Funding body
This thesis was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-42-4
Cover design by Claudia Flandoli
Printed and bound by Ipskamp Drukkers, Enschede

The research reported in this thesis was conducted at the Radboud University Centre for Language Studies, in Nijmegen, the Netherlands

Modelling multi-modal language learning:
from sentences to words

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 7 november 2022
om 16.30 uur precies

door
Danny Gijsbertus Martinus Merkx
geboren op 26 mei 1990
te 's-Hertogenbosch

Promotor:

Prof. dr. M.T.C. Ernestus

Copromotor:

Dr. S.L. Frank

Manuscriptcommissie:

Prof. dr. M.A. Larson (voorzitter)

Prof. dr. F. Moscoso del Prado Martín

Dr. A. Alishahi (Tilburg University)

Dr. D. Harwath (University of Texas at Austin)

Dr. O. Räsänen (Tampere University)

Contents

1 Introduction

When modelling the human cognitive capacity for language, one invariably has to consider how language is represented in the mind. Whether we study the processing and recognition of written language, sign language or speech, at some point the external linguistic stimuli become mental states that our brain is capable of operating on. Computational linguistic models try to formalise the operations our cognitive functions perform on linguistic stimuli, and as such force us to be explicit about the representations being operated on. For example in speech recognition models, the output is a 'recognised word', i.e., a mental state of the listener, and the modeller has to decide how to represent this output.

While nearly all cognitive models of language processing assume that we have mental representations of words, surprisingly few actually deal with where they come from. Prior knowledge of words is assumed, and, perhaps because many models are trained on (English) text, words are assumed to be 'the things between spaces'. This prior knowledge is not readily available however, when working with a language that has no orthography. During an internship at the Jelinek Memorial Workshop on Speech and Language Technology, I worked on a speech recognition system for unwritten languages. This posed a challenge, as many of the machine learning techniques for Natural Language Processing (NLP) I learned are dependent on text. How do you train such a system without relying on transcribed speech and what does speech recognition mean in a system that should not return a written transcript? We ended up implementing a system that instead finds a picture displaying what is described in the sentence, for instance, when hearing 'a dog running through a field', it should find a picture of a dog running through a field (Scharenborg et al., 2020).

While this speech recognition system is obviously limited to concrete descriptions of visual scenes, I wanted to continue investigating it and its potential as a cognitive model of language learning and speech recognition. While languages without orthography pose only a practical problem for an NLP application, they hint at a more fundamental question for cognitive modelling: can we model the representations that our mind operates on by depending solely on text? Infants initially have little understanding of what is being said around them, and yet, within a few years they are able to understand and speak their mother tongue. They learn by interaction with their environment and listening to other humans, a process that is well on its way before they sit down in a school bench to learn to read and write. In fact, many people learn their mother tongue without *ever* learning how to read and write. It seems clear to me that, because children obviously do not learn language through reading lots of text and many languages have no orthography at all, it is crucial that we look beyond written language and build our computational models around more natural linguistic input. Similarly in NLP, although cognitive plausibility is not an issue there, systems can benefit from more insight into how humans learn language. After all, human performance is still a golden standard in many machine learning tasks.

The speech recognition system I worked on learns to recognise speech by leveraging correlations between speech and the visual modality. Models that learn language by combining speech and vision have been around for quite a while (e.g., Roy and Pentland 1998). Recent advances in NLP and computer vision have made it possible to build such models on more than a toy lexicon, and sparked a new wave of interest in multi-modal learning. When the combination of linguistic and visual input is concerned, such a model is considered a Visual Grounding Model (VGM), that is, it grounds the representations of the linguistic units in visual information. While vision is not the only possible modality to include in multi-modal modals, it is the most prominent one and I will focus on a VGM in this dissertation.

My main goal is to investigate if a VGM can learn representations at the lexical and sentence level that capture cognitive aspects of meaning and predict human behavioural data. I aim to show that the solution to understanding the human language faculty will not come from increasingly large and complex text-based models, but from models that consider the wider range of the human sensory experience. In the next section, I discuss current approaches to representing linguistic units in computational models.

1.1 Linguistic representations

As said before, most computational models assume that the representations that model the mental states associated with the linguistic stimuli (which I will henceforth refer to as ‘linguistic representations’) are prior knowledge. The ‘mental lexicon’, containing all information about all the individual words a language user knows, including how to pronounce them, write them, what they mean and how to use them properly in a sentence (Emmorey and Fromkin, 1988) is prior knowledge to such models. Coming back to the speech recognition models, their lexicon is determines which words they are able to recognise in the first place, what they sound like and how they are written. Discovering what words there are in a language is no simple task however, as speech does not contain neat breaks or other clear clues as to where each word begins and ends.

It is compelling to compare the mental lexicon to a sort of dictionary; a collection of all known words, including their spelling, meaning and often an example of its use in a sentence. We could think of more elaborate dictionaries including pictures, sounds, synonyms etc. but somehow the comparison to a dictionary will always fall short. The mental lexicon is more than a collection of all the knowledge that you could possibly gather about words. An important question in psycholinguistic research is why certain words are faster to process than others. After all, in a dictionary, it should not take more time to look up one word or the other. This indicates that the structure of the mental lexicon is intricately linked to our daily *use* of these mental representations, influencing how words are activated and retrieved from memory when we read or hear them. Another fundamental difference with a dictionary is the fact that words in the mental lexicon are interconnected. Activation of one word can spread to related words, as emphasised for instance by semantic priming effects (e.g., Hutchison et al. 2013; D'Arcais et al. 1985; Schreuder et al. 1998).

However, we still know surprisingly little about what a human's mental representations look like, what information they contain, the ways in which they are connected, how they are learned, how multiple languages co-exist in the mental lexicon (or whether there is one for each language) and some even doubt the existence of the mental lexicon entirely (Elman, 2009). Computational models of the mental lexicon and the representations it contains are required to better understand it, and vice versa, computational models that rely on the mental lexicon can benefit from better linguistic representations. In the next section I discuss a widely used method to 'learn' linguistic representations from huge corpora using deep learning.

1.2 Word embeddings

The linguistic units (such as words and sentences) in computational linguistic models can be represented as embeddings. An embedding is a high-dimensional numerical vector that is not so much defined by the exact numbers in the vector as it is by its relation to other embeddings in the same embedding space. Distributional semantics models create embeddings that quantify word meaning based on the idea that a word's meaning depends on the contexts in which it appears. The same idea has also been applied to sentence meaning, that is, a sentence's meaning depends on the surrounding sentences. In such a learned embedding space, words or sentences with a similar meaning (i.e., appearing in similar contexts) should have a similar embedding. Another example is that word pairs with a similar relationship, such as 'king-queen' and 'man-woman', also have a similar difference vector in the embedding space. Embeddings are widely used in NLP systems, often to represent text inputs, and have been shown to improve for instance machine translation, named entity recognition and sentiment analysis to name a few (Qi et al., 2018; Sienčnik, 2015; Severyn and Moschitti, 2015).

The recent success of deep learning based distributional semantics in NLP has revived attention from the cognitive modelling community as well and semantic embeddings are now widely used as the linguistic input for various cognitive models, with research showing that they can account for response times in lexical decision tasks (Mandera et al., 2017; Rotaru et al., 2018; Petilli et al., 2021), decode brain data (Xu et al., 2016; Abnar et al., 2018), account for brain activity during text comprehension (Frank and Willems, 2017), and correlate with human judgements of word similarity (Kiela et al., 2018; Derby et al., 2018, 2020).

Distributional semantics models are in this sense used as computational models of where the linguistic representations come from and how they are learned. Aside from the fact that these word embeddings are only meant to quantify semantics and not for instance phonetic information, they are not cognitively plausible. Firstly, creating high-quality embeddings requires billions of word tokens. Obviously, humans are able to understand language after much less exposure, and furthermore, their language experience comes from much more than solely reading texts.

Secondly, these models treat word and sentence learning as entirely separate processes. Distributional semantics models are exposed to billions of word tokens to learn word meaning, without ever dealing with larger linguistic constructs. Distributional semantics models that learn sentence embeddings often start from pretrained word embeddings, implying that word meaning is learned first, before learning how they combine into a meaningful sentence (e.g., Conneau et al. 2017; Kiela et al. 2018).

Thirdly, any model that uses text, implicitly receives a lot of prior lexical information. Whereas speech does not contain neat breaks to indicate where words start and end, (English) text clearly demarcates the lexical items for which the model is supposed to learn representations. Even the knowledge that words exist at all can be considered prior knowledge. When learning language from speech and without prior knowledge, a model thus faces the difficult task of figuring out what sub-sentence units an utterance contains in the first place.

In the next section, I discuss the theories about human language learning that VGMs take inspiration from in order to create more cognitively plausible linguistic representations.

1.3 Human language learning

In order to create more cognitively plausible cognitive models, one unsurprisingly needs to look at how humans perform the cognitive function being modelled. As hinted at before, VGMs take inspiration from how children learn language. As children are able to learn language without any prior linguistic knowledge and without much explicit training, a plausible cognitive model should be able to do the same. While my VGM is not intended to be a computational model of child language acquisition, it focuses on the following two aspects of child language learning.

The first aspect is that language learning goes from utterances to words, and not the other way around. In usage-based theories of language there is, as the name suggests, a strong relationship between language use and the linguistic units involved. According to Tomasello (2009), a proponent of this theory, intention reading (i.e., the communicative intent behind language use) and pattern finding (i.e., identifying smaller linguistic units) are the key cognitive functions for learning language. The essential premise of this theory is that all language use has communicative intent, that is for instance, you want another person to do something or attend to something. Because the utterance is the smallest unit that conveys communicative intent, children would start with complete utterances as basic linguistic units when they learn language (Tomasello, 2000). Indeed, research shows that in young children, much of their language use is constrained to (parts of) utterances they have used before (Lieven et al., 2003) or comes from a small set of patterns like: 'Where is X' and 'Want more X' (Braine and Bowerman, 1976).

Later on, children's linguistic units become smaller, as they learn to identify slots in the linguistic patterns and learn which constituents of their linguistic units they can 'cut and paste' to create novel utterances (Pine and Lieven, 1993; Tomasello, 2000). Still, some relatively frequently used expressions such as 'how-are-you-doing' might become entrenched as a single linguistic unit even in adults. According to this view, the linguistic units in the mental lexicon can be 'the things between spaces' but can also refer to concepts that would require multiple 'words' to describe, or can even be complete sentences.

How exactly children learn language is not known, but from a usage-based view two things seem clear: firstly, learning about words and learning about sentences are not two separate, consecutive processes. Children's linguistic units become smaller (i.e., more like the traditional idea of words) by hearing sentences and finding patterns. Furthermore, in usage-based models, linguistic units are not necessarily the 'things between spaces', but multi-word expressions as well. Humans learn about words by finding patterns in sentences, they do not learn about sentences by learning how to combine words. Distributional semantics models consider word and sentence learning as separate processes and even in the wrong order. A model that first learns word representations through reading billions of word tokens, before learning how to use them in a sentence is not very plausible.

The second aspect of child language learning the VGM takes inspiration from is that humans have multiple streams of sensory information to learn language from, and text is not the one children start learning language with. According to embodied cognition theory, our conceptual knowledge is based on all our sensory experiences (Barsalou, 2008; Foglia and Wilson, 2013). For instance, hearing the word *coffee* brings back other sensory experiences associated with coffee, such as how it smells, looks and tastes. Embodied cognition theory thus assumes that all our sensory experiences contribute to our conceptual knowledge and processing, which should be reflected in human behaviour. If this is the case, we cannot model human language learning without considering a wider range of sensory experiences.

Of these sensory experiences, the most prominent one is the visual stream. It is theorised that infants learn to extract their words from speech by repeatedly hearing words while seeing the objects or actions these words refer to (Räsänen and Rasilo, 2015). For instance, parents might say 'the ball is on the table' and 'there's a ball on the floor' etc., while consistently pointing towards a ball. Children could learn what 'ball' means, because it is something both utterances and their associated visual scenes have in common. The combination of speech and visual context offers a possible mechanism for learning words when no prior linguistic information (e.g., segmentation) is given.

These two observations are essentially captured in the title of this dissertation: modelling multi-modal language learning, from sentences to words. As said before, the VGM is not intended to be a full computational account of child language acquisition. Rather, I hope to make the case that the VGM is a more cognitively plausible method for creating linguistic representations than distributional semantics and that in order to advance our cognitive models of language learning, we need to consider a wider range of sensory experiences, and treat sentence and word learning as a single end-to-end process.

1.4 Visually Grounded Models of language

Here I give a short overview of visually grounded models of language, but for a comprehensive and recent review, I refer to Chrupała (2022).

The potential of visual input for modelling the learning of linguistic units has long been recognised. One of the first VGMs of language learning is CELL, developed by Roy and Pentland (1998). Their model builds an 'audio-visual lexicon' by finding clusters in the visual input and looking for reoccurring segments in the acoustic signal. However, the model was limited to colours and shapes (utterances like 'this is a blue ball') and does not learn from more natural, less restricted input.

The VGM used in this dissertation is based on the recent wave of neural network based VGMs. Advances in machine learning have made it possible to train deep neural networks on large databases, and as a result, many such databases and models have emerged in the past ten years. In 2013, Hodosh, Young and Hockenmaier introduced Flickr8k (Hodosh et al., 2013), a database of images accompanied by written captions describing their contents, which was quickly followed by similar databases such as MSCOCO Captions (Chen et al., 2015). These datasets are now widely used for image-caption retrieval models (e.g., Karpathy and Fei-Fei 2015; Klein et al. 2015; Ma et al. 2015; Vendrov et al. 2016; Wehrmann et al. 2018; Dong et al. 2018) and caption generation (e.g., Karpathy and Fei-Fei 2015; Xu et al. 2015).

The move from written to spoken language was started by Harwath and Glass (2015), who collected spoken captions for the Flickr8k database and used it to train the first neural network based VGM based on speech. Initially, VGMs were intended as NLP systems: systems that retrieve relevant images for a given caption or vice versa, and even systems that could generate appropriate captions for an image. However, many studies have since begun considering VGMs as cognitive models and investigated the properties of their learned representations (e.g., Harwath et al. 2020a; Kiela et al. 2018; Chrupała et al. 2018; Hsu et al. 2020; Chrupała et al. 2020). Räsänen and Khorrami (2019) used a VGM to show that words can be discovered from recordings made by head-mounted cameras worn by infants during child-parent interaction. They showed that their model learned utterance representations from which several words (e.g., 'doggy', 'ball') could reliably be detected. This study was an important step towards showing that VGMs can acquire linguistic units from actual child-directed speech.

1.5 This dissertation

My first attempt at implementing a VGM was a DyNet (Neubig et al., 2017) implementation of the model by Harwath and Glass (2015), albeit with a different neural network structure. That implementation was made as part of my internship at the 2017 Jelinek Memorial Workshop on Speech and Language Technology (Scharenborg et al., 2018). At the time however, the system did not perform as it should, and we were unable to pinpoint the flaw in the implementation[1].

I decided to start the VGM implementation for this dissertation from scratch, working in Python and PyTorch, to get a working implementation that I could easily alter for the purposes of the research in this dissertation. Even though my main interest is in a model that learns language without requiring any text at all, I first built a text-based model, for the main reason that working with text is simply less complex than working with speech. The working text-based model was a stepping stone towards implementing the speech-based model. Another advantage of text is that it is more straightforward to test the sentence and word representations created by the model. Sentence- and word-embedding evaluation datasets are readily available for written data, as are alternative text-based sentence and word-embedding models to facilitate a comparative evaluation. By starting from a text-based model, I was able to implement and investigate both a text- and speech-based model, the results of which will be discussed in the remainder of this dissertation.

[1] Scharenborg et al. (2020), a follow up to that paper, includes a working implementation based on the work done for this dissertation in **Chapter 4**.

This dissertation is divided into three parts; the first part is an investigation of my text-based visual grounding model and in the second part, I present the speech-based model. Both start with an investigation of the sentence representations and conclude with an investigation of the model's word representations. The third part is a more in-depth investigation of the way deep learning models of cognition process language and how this relates to our knowledge of human language processing. Finally, in **Chapter 8** I present my overall conclusions and recommendations for future work.

1.5.1 Text model

In **Chapter 2**, which is based on Merkx and Frank (2019), I introduce the visual grounding model, which is my own Python implementation based on caption-image retrieval models such as those by Karpathy and Fei-Fei (2015), Wehrmann et al. (2018) and Dong et al. (2018). As caption-image retrieval models have hitherto been used primarily as NLP systems, the main goal of this chapter is to evaluate their viability as a cognitive models. Crucial for all subsequent research in this dissertation, the model has to be able to extract semantic knowledge without requiring any prior linguistic knowledge. Existing image-caption retrieval models as well as semantic sentence embedding models use pretrained word embeddings as input and the cognitive plausibility of such models is questionable. I investigate whether the VGM learns to capture aspects of semantic relatedness in its embedding space by using a large test battery for evaluating sentence embeddings called SentEval and comparing the model's sentence embeddings to a state-of-the-art text-based sentence embedding model.

In **Chapter 3** (Merkx et al., 2022), I build upon the results from the previous chapter and investigate whether the model can be used to create semantic word embeddings. As the model is trained on full sentences, it is not self-evident that it would also learn meaningful word representations, especially considering that previous implementations all used pretrained word embeddings as input. The goal of this chapter is to introduce a method to extract word embeddings from a model which in principal delivers sentence embeddings, and investigate whether these grounded embeddings capture cognitive aspects of word meaning that text-based approaches cannot. I test a central idea of embodied cognition theory, namely that all our sensory experiences contribute to our conceptual knowledge. If our visual experiences contribute to our conceptual knowledge, word embeddings incorporating visual features should be able to explain human behavioural data to a degree unattainable by purely text-based methods. To evaluate this claim, I perform two experiments on different types of human behavioural data, using well-known text-based word embedding methods as control variables to carefully separate the contribution of text-based knowledge from visual knowledge.

1.5.2 Speech model

The main goal of **Chapter 4** (Merkx et al., 2019) is to introduce the speech-based VGM. I build upon existing VGMs by implementing several improvements to their architecture and training. As with the text-based model, the speech-based VGM is trained on full sentences, and produces sentence embeddings. Inherent to the spoken input is the fact that the speech-based VGM receives no word boundary information, or even knows that words exist at all. I also perform a probing experiment to see if the model encodes the presence of words in its sentence embeddings. This is essential for the follow-up experiments and as the model is trained only on full utterances without any explicit clues about word boundaries, it is not self-evident that it encodes such information.

In **Chapter 5** (Merkx et al., 2021), I investigate whether the VGM learns to capture sentence semantics. This chapter can be seen as the *spoken* counterpart to **Chapter 2**. I collect a spoken equivalent to the pre-existing evaluation data used in the second chapter in order to test whether the speech-based VGM is able to learn sentence semantics, even though it is not explicitly trained to recognise speech, and in contrast to the text-based model from **Chapter 2**, does not even receive word boundary information. I also provide a critical note on the trend in computational linguistics and NLP to ‘improve’ models by creating larger neural networks trained on larger training corpora. I compare models trained on datasets that differ in composition but are similar in size, and aim to show that there are facets to corpus building other than its size to consider if we want to improve our models.

The speech-based VGM takes the leap from sentences to words in **Chapter 6** (Merkx et al., forthcoming). Whereas the fourth chapter introduced the first step in this direction, I here present the VGM as a model of human word recognition. The central question of this chapter is: does the model learn to recognise words, and does this generalise to isolated words? I also investigate whether the model’s word recognition performance is affected by known word competition effects. This chapter aims to be a closing piece for the speech-based part of this dissertation, showing that by combining multi-modal input, it can indeed learn language on both the sentence and word level without requiring any prior linguistic knowledge and without separating the processes of sentence and word learning.

1.5.3 Working memory in language processing

In **Chapter 7** (Merkx and Frank, 2021) I take a closer look at the way deep learning models process language. At the start of my project, Recurrent Neural Networks (RNN) were the main neural architecture used in cognitive linguistic models (including mine). The success of RNNs in explaining behavioural and neurophysiological data suggests that something akin to recurrent processing is involved in human sentence processing. Recently however, the Transformer architecture was introduced and proceeded to break record after record in NLP applications. The transformer differs drastically from the RNN in how it processes and 'remembers' previous inputs. The Transformer's success in processing language merits a closer look at this architecture as a possible processing mechanism in cognitive models as well. In this chapter I compare RNNs and Transformers as a model of human sentence processing. Even though at face-value the Transformer seems cognitively implausible, a re-evaluation of the RNN may be required if the Transformer outperforms the RNN as a cognitive model.

2 Learning semantic sentence representations from visually grounded language without lexical knowledge

Current approaches to learning semantic representations of sentences often use prior word-level knowledge. The current study aims to leverage visual information in order to capture sentence level semantics without the need for word embeddings. We use a multi-modal sentence encoder trained on a corpus of images with matching text captions to produce visually grounded sentence embeddings. Deep Neural Networks are trained to map the two modalities to a common embedding space such that for an image the corresponding caption can be retrieved and vice versa. We show that our model achieves results comparable to the current state-of-the-art on two popular image-caption retrieval benchmark data sets: MSCOCO and Flickr8k. We evaluate the semantic content of the resulting sentence embeddings using the data from the Semantic Textual Similarity benchmark task and show that the multi-modal embeddings correlate well with human semantic similarity judgements. The system achieves state-of-the-art results on several of these benchmarks, which shows that a system trained solely on multi-modal data, without assuming any word representations, is able to capture sentence level semantics. Importantly, this result shows that we do not need prior knowledge of lexical level semantics in order to model sentence level semantics. These findings demonstrate the importance of visual information in semantics.

2.1 Introduction

Distributional semantics, the idea that words that occur in similar contexts have similar meanings, has been around for quite a while (e.g., Rubenstein and Goodenough 1965; Deerwester et al. 1990). Rubenstein and Goodenough (1965)

This chapter is based on: Danny Merkx and Stefan L. Frank. Learning semantic sentence representations from visually grounded language without lexical knowledge. *Natural Language Engineering*, 25(4):451-466, 2019 already studied “how the proportion of words common to contexts containing word A and to contexts containing word B was related to the degree to which A and B were similar in meaning” (p.627). State-of-the-art word embedding methods such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) have shown meaningful clusters, correlations with human similarity judgements (De Deyne et al., 2017), and have become widely used features that boost performance in several natural language processing (NLP) tasks such as machine translation (Qi et al., 2018). With the success of word embeddings, researchers are looking for ways to capture the meaning of larger spans of text, such as sentences, paragraphs, and even entire documents. Much less is known about how to approach this problem and early solutions tried to adapt word embedding methods to larger spans of text, for example, Skip-Thought sentence embeddings (Kiros et al., 2015), FastSent (Hill et al., 2016), and Paragraph-Vector (Hill et al., 2016), which are related to the Skip-Gram word model by Mikolov et al. (2013a). Recently, there have also been successful sentence embedding models which are trained on a supervised task and then transferred to other tasks (e.g., Conneau et al. 2017; Yang et al. 2018; Kiela et al. 2018).

So far, existing sentence embedding methods often require (pretrained) word embeddings (Conneau et al., 2017; Kiela et al., 2018), large amounts of data (Hill et al., 2016), or both (de Boom et al., 2015; Yang et al., 2018). While word embeddings are successful at enhancing sentence embeddings, they are not very plausible as a model of human language learning. Firstly, a model using word embeddings makes the assumption that the words in its lexicon are the linguistic units bearing meaning. It is for instance not possible for the model to focus on only part of the morphology of such a predefined unit. Secondly, these models assume that the process of language acquisition begins with lexical level knowledge before learning how to process longer utterances. That is, the model already knows what a word is and in the case of pretrained word embeddings it receives considerable prior knowledge of lexical semantics. Both of these assumptions are questionable.

Tomasello (2000), a proponent of usage-based models of language, argues that children learn many relatively fixed expressions (e.g., ‘how-are-you-doing’) as single linguistic units. Furthermore, he argues that the linguistic units that children operate on early in language acquisition are entire utterances, before their language use becomes more adult-like. Indeed, research shows that in young children, much of their language use is constrained to (parts of) utterances they have used before (Lieven et al., 2003) or comes from a small set of patterns like: 'Where is X' and 'Want more X' (Braine and Bowerman, 1976). Children's linguistic units become smaller and more adult-like as they learn to identify slots in the linguistic patterns and learn which constituents of their linguistic units they can 'cut and paste' to create novel utterances (Pine and Lieven, 1993; Tomasello, 2000). Models that assume lexical items are the basic meaning bearing units and that language learning starts from lexical items towards understanding full sentences are thus not very plausible as models of language learning.

In the current study, we train a sentence encoder without prior knowledge of lexical semantics, that is, without using word embeddings. Instead of word embeddings, we use character level input in conjunction with visual features. The use of multi-modal data has proven successful on the level of word embeddings (see for instance Collell et al. 2017; Derby et al. 2018). For sentence semantics, the multi-modal task of image-caption retrieval, where given a caption the model must return the matching image and vice versa, has been proposed as a way of grounding sentence representations in vision (Harwath and Glass, 2015; Leidal et al., 2017). Recently Kiela et al. (2018) found that such models do indeed produce embeddings that are useful in tasks like natural language inference, sentiment analysis and subjectivity/objectivity classification.

Our model does not know a priori which constituents of the input are important. It may learn to extract features from spans of text both larger and smaller than words. Furthermore, we leverage the potential semantic information that can be gained from the visual features to create visually grounded sentence embeddings without the use of prior lexical level knowledge. We also probe the semantic content of the grounded sentence embeddings more directly than has so far been done, by evaluating on Semantic Textual Similarity, a well known benchmark test set consisting of sentence pairs with human-annotated semantic similarity ratings.

Our aim is to create a language model that learns semantic representations of sentences in a more cognitively plausible way, that is, not purely text-based and without prior lexical level knowledge. We evaluate our multi-modal sentence encoder on a large benchmark of human semantic similarity judgements in order to test if the similarity between the embeddings correlates with human judgements of semantic textual similarity. This is to the best of our knowledge the first evaluation of the sentence level semantics of a multi-modal encoder that does not make use of lexical information in the form of word embeddings. We find that the model produces sentence embeddings that account for human similarity judgements, with performance similar to competing models. Importantly, our model does so using visual information rather than prior knowledge such as word embeddings. We release the code of our preprocessing pipeline, models and evaluation on github as open source: `https://github.com/DannyMerkx/caption2image`.

2.2 Sentence embeddings

2.2.1 Text-only methods

Methods for creating sentence embeddings have thus far mostly been based solely on text data. Skip-Thought (Kiros et al., 2015), inspired by the idea behind word embeddings, assumes that sentences which occur in similar context have similar meaning. Skip-Thought encodes a sentence and tries to reconstruct the previous sentence and the next sentence from the resulting embedding. In a similar approach, Yang et al. (2018) try to match Reddit posts with their responses based on the assumption that posts with similar meanings will elicit similar responses.

InferSent, a recent model by Conneau et al. (2017), is one of the most successful models with regards to transfer learning and semantic content. Conneau et al. trained an RNN sentence encoder on the Stanford Natural Language Inference database (Bowman et al., 2015), a database with paired sentences annotated for entailment, neutral, or contradiction relationships. Conneau and Kiela (2018) released SentEval, a transfer learning evaluation toolbox for sentence embeddings, which includes a large number of human semantic similarity judgements. InferSent embeddings show a high correlation to several sets of semantic textual similarity judgements and perform well on various transfer tasks like sentiment analysis and subjectivity/objectivity detection.

2.2.2 Multi-modal methods

Image-caption retrieval is a multi-modal machine learning task involving challenges from both computer vision and language modelling. The task is to rank captions by relevance to a query image, or to rank images by relevance to a query caption, which is done by mapping the images and captions to a common embedding space and minimising the distance between the image and caption in this space.

Ma et al. (2015) used two Convolutional Neural Networks (CNN) to create image and sentence representations and another CNN followed by a Multilayer Perceptron (MLP) to derive a matching score between the images and captions. Klein et al. (2015) converted the captions to Fisher vectors (Jaakkola and Haussler, 1998) and used Canonical Correlations Analysis to map the caption and image representations to a common space. The model by Karpathy and Fei-Fei (2015) works at a different granularity: they encoded image regions selected by an object detection CNN and encoded each word in the sentence separately, thus ending up with multiple embeddings per caption and image. They then calculated the distances between all the embedded words and image regions.

Many image-caption retrieval models rely on pretrained neural networks and word embeddings. It is common practice to use a pretrained network such as VGG, Inception V2, or ResNet-152 to extract the visual features (e.g., Ma et al. 2015; Vendrov et al. 2016; Faghri et al. 2017; Wehrmann et al. 2018; Kiela et al. 2018). Furthermore, with the exception of the character-based model by Wehrmann et al. (2018), recent results are achieved by using pretrained Word2Vec or GloVe word embeddings to initialise the sentence encoder. The current state-of-the-art results are by Faghri et al. (2017), who fine-tuned a pretrained ResNet-152 and improved the sampling of mismatched image-caption pairs during training.

The approach of mapping the image-caption pairs to a common semantic embedding space is interesting because the produced embeddings could also be useful in other tasks, similar to how word embeddings can be useful in machine translation (Qi et al., 2018). Kiela et al. (2018) used a model similar to Dong et al. (2018), that is, a recurrent neural network caption encoder paired with a pretrained image recognition network which is trained to map the caption to the image features extracted by the image recognition network. Using SentEval, Kiela et al. (2018) showed that the resulting embeddings are useful in a wide variety of transfer tasks such as sentiment analysis in product and movie reviews, paraphrase detection and natural language inference. These results show that visually grounded sentence representations can be used for transfer learning, but do not directly probe the model's ability to learn sentence semantics.

The current study differs from previous research in three respects. Firstly, we train our model using character level input rather than word embeddings. Secondly, our model uses only the sentence representations that can be learned from the multi-modal training data. In contrast, Kiela et al. (2018) augmented their grounded representations by combining them with non-grounded (Skip-

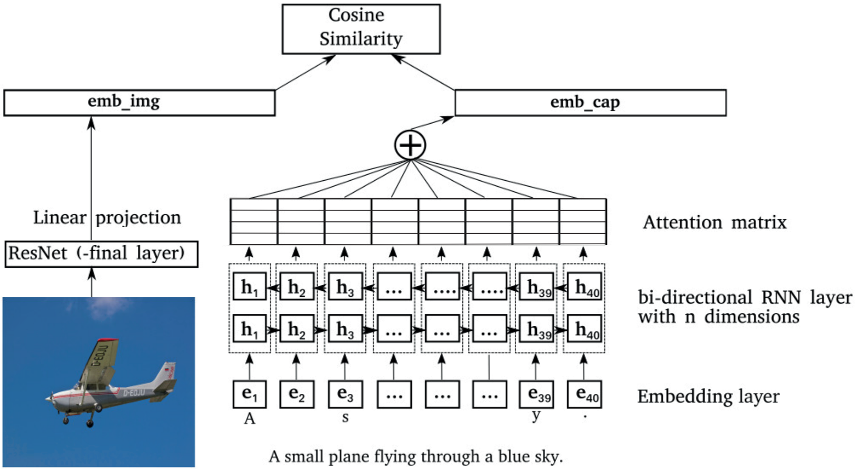
Cosine Similarity
emb_img
emb_cap
Linear projection
ResNet (-final layer)
Attention matrix
bi-directional RNN layer with n dimensions
Embedding layer
h_1 h_2 h_3 ... . h_{39} h_{40}
h_1 h_2 h_3 ... ... h_{39} h_{40}
e_1 e_2 e_3 ... ... e_{39} e_{40}
A s y .
A small plane flying through a blue sky.

Figure 2.1: Model architecture: the model consists of two branches with the image encoder on the left and the caption encoder on the right. The character embeddings are denoted by $\mathbf{e}_t$ and the RNN hidden states by $\mathbf{h}_t$. Each hidden state has n features which are concatenated for the forward and backward RNN into $2n$ dimensional hidden states. Then attention is applied which weighs the hidden states and then sums over the hidden states resulting in the caption embedding. At the top we calculate the cosine similarity between the image and caption embedding (**emb_img** and **emb_cap**).

Thought) representations. Finally, we probe the semantic content of our sentence representations more directly by evaluating the caption encoder on the Semantic Textual Similarity benchmark. This benchmark is included in the SentEval toolbox but has to the best of our knowledge not been used to evaluate visually grounded sentence representations.

2.3 Approach

In this section, we first describe our encoder architectures, where we combine several best practices and state-of-the-art methods in the field of deep learning. Next, we describe the training data and finally the semantic similarity tasks.

2.3.1 Encoder architectures

Image encoder

Our model maps images and corresponding captions to a joint embedding space, that is, the encoders are trained to make the embeddings of an image-caption pair lie close to each other in the embedding space. As such the model requires both an image encoder and a sentence encoder as illustrated in Figure 2.1.

The image features are extracted by a pretrained image recognition model trained on ImageNet (Deng et al., 2009). For this we used ResNet-152 (He et al., 2016), a residualised network with 152 layers from which we take the activations of the penultimate fully connected layer[1]. ResNet-152 has lower error rates on the ImageNet task than other networks previously used in the image captioning task such as VGG16, VGG19 and Inception V2.

For the image encoder we use a single layer linear projection on top of the pretrained image recognition model, and normalise the result to have unit L2 norm:

$$\mathbf{emb\_img} = \frac{\mathbf{img}A^T + \mathbf{b}}{||\mathbf{img}A^T + \mathbf{b}||_2}$$

where A and $\mathbf{b}$ are learned weights and bias terms, and **img** is the vector of ResNet image features.

Caption encoder

We built a caption encoder that trains on raw text, that is, character-level input. The sentence encoder starts with an embedding layer with embeddings $(\mathbf{e}_1, ..., \mathbf{e}_t)$ for the t characters in the input sentence. The embeddings are then fed into an RNN, followed by a self-attention layer and lastly normalised to have unit L2 norm:

$$\mathbf{emb\_cap} = \frac{\mathrm{Att}(\mathrm{RNN}(\mathbf{e}_1, ..., \mathbf{e}_t))}{||\,\mathrm{Att}(\mathrm{RNN}(\mathbf{e}_1, ..., \mathbf{e}_t))||_2}$$

where $\mathbf{e}_1, ..., \mathbf{e}_t$ indicates the caption represented as character embeddings and Att is the attention layer. The character embedding features are learned along with the rest of the network.

The RNN layer allows the network to capture long-range dependencies in the captions. Furthermore, by making the layer bidirectional we let the network

[1] The final layer of a pretrained visual network is a task-specific object classification layer while the penultimate layer contains generally useful image features. Madhyastha et al. (2018) document that the features of the penultimate layer yield better transfer learning results than the object classification layer.

process the captions from left to right and vice versa, allowing the model to capture dependencies in both directions. We then concatenate the results to create a single embedding. We test two types of RNN: the Long Short Term Memory unit (LSTM; Hochreiter and Schmidhuber 1997) and the Gated Recurrent Unit (GRU; see Greff et al. 2017 and Chung et al. 2014 for detailed descriptions of these RNNs). The GRU is a recurrent layer that is widely used in sequence modelling (e.g., Zhu et al. 2015; Patel et al. 2016; Conneau et al. 2017). The GRU requires fewer parameters than the LSTM while achieving comparable results or even outperforming LSTMs in many cases (Chung et al., 2014). On the other hand, Conneau et al. (2017) found that an LSTM not only performed better than a GRU on their training task, but also generalised better to other tasks including semantic similarity. We test both architectures as it is not clear which is better suited for the image-captioning task.

The self-attention layer computes a weighted sum over all the hidden RNN states:

$$\mathbf{a}_t = \text{softmax}(V \tanh(W\mathbf{h}_t + \mathbf{b}_w) + \mathbf{b}_v)$$
$$\text{Att}(\mathbf{h}_1, ..., \mathbf{h}_t) = \sum_t \mathbf{a}_t \circ \mathbf{h}_t$$

where $\mathbf{a}_t$ is the attention vector for hidden state $\mathbf{h}_t$ and W, V, $\mathbf{b}_w$, and $\mathbf{b}_v$ indicate the weights and biases. The applied attention is then the sum over the Hadamard product between all hidden states $(\mathbf{h}_1, ..., \mathbf{h}_t)$ and their attention vector.

While attention is part of many state-of-the-art NLP systems, Conneau et al. (2017) found that attention caused their model to overfit on their training task, giving worse results on transfer tasks. As a simpler alternative to attention, we also test max pooling, where we take for each feature the maximum value over the hidden states.

Both encoders are jointly trained to embed the images and captions such that the cosine similarity between image and caption pairs is larger (by a certain margin) than the similarity between mismatching pairs, minimising the so-called hinge loss. The network is trained on a minibatch B of correct image-caption pairs (cap, img) where all other image-caption pairs in the minibatch serve to create counterexamples (cap, img') and (cap', img). We calculate the cosine similarity $\cos(x, y)$ between each embedded image-caption pair and subtract the similarity of the mismatched pairs from the matching pairs such that the loss is only zero when the matching pair is more similar by a margin α. The hinge loss L as a function of the network parameters θ is given by:

$$L(\theta) = \sum_{(cap,img),(cap',img')\in B} \Big(\max(0, \cos(cap, img') - \cos(cap, img) + \alpha) + \\ \max(0, \cos(img, cap') - \cos(img, cap) + \alpha) \Big)$$

where $(cap, img) \neq (cap', img')$.

2.3.2 Training data

The multi-modal embedding approach requires paired captions and images for which we use two popular image-caption retrieval benchmark datasets: Flickr8k (Hodosh et al., 2013) and MSCOCO (Chen et al., 2015).

Flickr8k is a corpus of 8,000 images taken from the online photo sharing application Flickr.com. Each image has five captions created using Amazon Mechanical Turk (AMT) where workers were asked to "write sentences that describe the depicted scenes, situations, events and entities (people, animals, other objects)" (Hodosh et al., 2013, p. 860). We used the data split provided by Karpathy and Fei-Fei (2015), with 6,000 images for training and a development and test set of 1,000 images each.

To extract the image features, all images are resized such that the smallest side is 256 pixels while keeping the aspect ratio intact. We take ten 224 × 224 crops of the image: one from each corner, one from the middle and the same five crops for the mirrored image. We use ResNet-152 pretrained on ImageNet to extract visual features from these ten crops and then average the features of the ten crops into a single vector with 2,048 features. The character input is provided to the networks as is, including all punctuation and capitals.

Microsoft Common Objects in Context (MSCOCO) is a large dataset of 123,287 images with five captions per image. The captions were gathered using AMT, with workers being asked to describe the important parts of the scene. Like Vendrov et al. (2016), we use 113,287 images for training and 5,000 for development and testing each. The image and text features are extracted from the data following the same procedure used for Flickr8k. The only difference is that the captions are provided in a tokenised format and we create the character level input by concatenating the tokens with single spaces and adding a full stop to the end of each caption.

2.3.3 Training procedure

The image-caption retrieval performance on the development set is used to tune the hyperparameters for each network. We found a margin $\alpha = 0.2$ for the loss function to work best on both the GRUs and LSTMs. Although performance was relatively stable in the range $0.15 \leq \alpha \leq 0.25$, it quickly degraded outside this range. The networks were trained with a single layer bidirectional RNN and we tested hidden layer sizes $n \in \{512, 1024, 2048\}$. The number of hidden units determines the embedding size, which is $2n$ (due to the RNN being bidirectional). The attention layer has 128 hidden units. The image encoder has $2n$ dimensions to match the size of the sentence embeddings. We use 20-dimensional character embeddings and found that varying the size of these embeddings has very little effect on performance.

The networks are trained using Adam (Kingma and Ba, 2015) with a cyclic learning rate schedule based on Smith (2017). The learning rate schedule varies the learning rate lr smoothly between a minimum and maximum bound (lr_{min} and lr_{max}) over the course of four epochs as given by:

$$lr = 0.5(lr_{\text{max}} - lr_{\text{min}})(1 + \cos(\pi(1 + 0.5 step \times mb))) + lr_{\text{min}}$$

where $step$ indicates the step size, that is, the number of minibatches for a full cycle of the learning rate, and mb is the number of minibatches processed so far. We set the step size such that the learning rate cycle is four epochs. The cyclic learning rate has two advantages. Firstly, fine-tuning the learning rate can be a very time consuming process. Smith (2017) found that the cyclic learning rate works within reasonable upper and lower bounds which are easy to find: simply set the upper and lower bound by selecting the highest and lowest learning rates for which the loss value decreases. Secondly, the learning rate schedule causes the network to visit several local minima during training, allowing us to use snapshot ensembling (Huang et al., 2017). By saving the network parameters at each local minimum, we can ensemble the caption embeddings of multiple networks at no extra cost.

We train the networks for 32 epochs and take a snapshot for ensembling at every fourth epoch. For ensembling we use the two snapshots with the highest performance on the development data. We found that for Flickr8k an upper bound on the learning rate of 10^{-3} and a lower bound of 10^{-6} worked well and for MSCOCO we had to adjust the upper bound to 10^{-4}.

2.3.4 Semantic evaluation

For the semantic evaluation we use the SentEval toolbox introduced by Conneau and Kiela (2018). This toolbox is meant to test sentence embeddings on a diverse set of transfer tasks, from sentiment analysis and paraphrase detection to entailment prediction. For semantic textual similarity analysis, SentEval includes the Semantic Textual Similarity and Sentences Involving Compositional Knowledge datasets which we briefly review here. After training our multi-modal encoder network, we simply discard the image encoder, and the caption encoder is used to encode the test sentences in SentEval.

Semantic Textual Similarity (STS) is a shared task hosted at the SemEval workshop. SentEval covers the STS datasets from 2012 to 2016. The datasets consist of paired sentences from various sources labelled by humans with a similarity score between zero ('the two sentences are completely dissimilar') and five ('the two sentences are completely equivalent, as they mean the same thing') for a total of five annotations per sentence pair (Agirre et al. 2015, p. 254, see also for a full description of the annotator instructions). The evaluation performed on the STS 2012 to 2016 tasks measures the correlation between the cosine similarity of the sentence embeddings and the human similarity judgements.

The STS Benchmark set (STS-B) consists of 8,628 sentence pairs selected from all STS tasks (Cer et al., 2017). STS-B consists of a training, development and test set (5,749, 1,500 and 1,379 sentence pairs respectively). For the STS-B task, the SentEval toolbox trains a classifier which tries to predict the similarity scores using the sentence embeddings resulting from our model. Table 2.1 gives an overview of the datasets. For full descriptions of each dataset see Agirre et al. (2012, 2013, 2014, 2015, 2016).

Sentences Involving Compositional Knowledge (SICK) is a database created for a shared task at SemEval-2014 with the purpose of testing compositional distributional semantics models (Bentivogli et al., 2016). The dataset consists of 10,000 sentence pairs which were generated using sentences taken from Flickr8k and the STS 2012 MSRvid data set. The sentences were altered to display linguistic phenomena that the shared task was meant to evaluate, such as negation. This resulted in sentences like 'there is no biker jumping in the air' and 'two angels are making snow on the lying children' (altered from 'two children are lying in the snow and are making snow angels', Bentivogli et al. 2016, p. 6) which do not occur in the Flickr8k training data.

For the semantic evaluation of our sentence embeddings we used the SICK Relatedness (SICK-R) annotations. For the SICK-R task, annotators were asked

Table 2.1: Description of the various STS tasks and their subtasks. Some subtasks appear in multiple STS tasks, but consist of different sentence pairs drawn from the same source. The image description datasets are drawn from the PASCAL VOC-2008 dataset (Everingham et al., 2008) and so do not overlap with Flickr8k or MSCOCO.

Task	Subtask	#Pairs	Source
STS 2012	MSRpar	750	newswire
	MSRvid	750	videos
	SMTeuroparl	459	glosses
	OnWN	750	WMT eval.
	SMTnews	399	WMT eval.
STS 2013	FNWN	189	newswire
	HDL	750	glosses
	OnWN	561	glosses
STS 2014	Deft-forum	450	forum posts
	Deft-news	300	news summary
	HDL	750	newswire headlines
	Images	750	image descriptions
	OnWN	750	glosses
	Tweet-news	750	tweet-news pairs
STS 2015	Answers forum	375	Q&A forum answers
	Answers students	750	student answers
	Belief	375	committed belief
	HDL	750	newswire headlines
	Images	750	image descriptions
STS 2016	Answer-Answer	254	Q&A forum answers
	HDL	249	newswire headlines
	Plagiarism	230	short-answer plagiarism
	Postediting	244	MT postedits
	Question-Question	209	Q&A forum questions
Total		12,544	

to rate the relatedness of sentence pairs on a 5-point scale for a total of ten annotations per sentence pair. Unlike for STS, there were no specific descriptions attached to the scale; participants were only instructed using examples of related and unrelated sentence pairs. Similar to STS-B, a classifier is trained on top of the embeddings, using 45 percent of the data as training set, 5 percent as development set and 50 percent as test set.

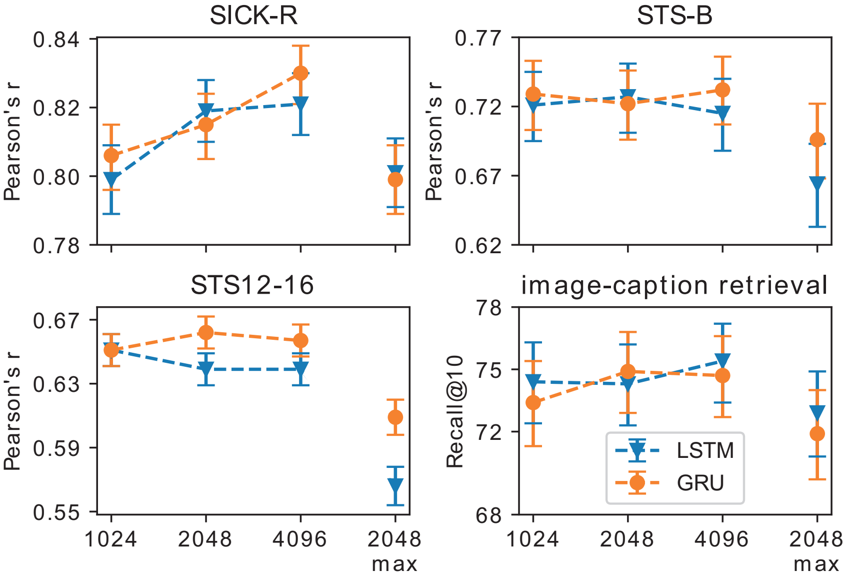

Figure 2.2: Model performance on the semantic (SICK-R, STS-B, and STS12-16) and training task (image-caption retrieval) measures including the 95 percent confidence interval. Training task performance is measured in recall@10. The semantic performance measure is Pearson's r. The horizontal axis shows the embedding size with "max" indicating the max pooling model.

2.4 Results and discussion

2.4.1 Model selection

We perform model selection after training on only the Flickr8k database. Due to the considerably larger size of MSCOCO it is more efficient to train and test our models on Flickr8k, and train on MSCOCO using only the best setup found on Flick8k.

To select the DNN architecture with the best performance we compare our architectures on image-caption retrieval performance and on their ability to capture semantic content. The image-caption retrieval performance is measured by Recall@10: the percentage of images (or captions) for which the correct caption (or image) was in the top ten retrieved items. For the purpose of model selection we use the average of the bidirectional (caption to image and image to caption) retrieval results on the development set. For the semantic evaluation we use correlation coefficients (Pearson's r) between embedding distances and human similarity judgements from STS-B and SICK-R. We also aggregate the Pearson's r scores for the STS 2012 through 2016 tasks.

Figure 2.2 shows the results for our models trained on Flickr8k. There is no clear winner in terms of performance: the GRU 2048 (referring to the embedding size) performs best on STS, GRU 4096 on SICK-R and STS-B, and LSTM 4096 on the training task. Although there are differences between the GRU and the LSTM, they are only statistically significant for STS12-16. Furthermore, the max pooling models are outperformed by their attention-based counterparts. We only tested the max pooling with an embedding size of 2048. Due to the clear drop in both training and semantic task performance we did not run any further experiments.

As our main goal is the evaluation of semantic content, we continue with the GRUs as they perform significantly better on STS12-16. There is no clear winner between the GRU 2048 and GRU 4096 as the performance differences on all measures are relatively small. The 4096 model performs significantly better on SICK-R but the 2048 model performs slightly better on STS12-16. As STS12-16 is the main interest in our evaluation we pick the GRU 2048 as our best performing Flickr8k model and train a GRU 2048 model on MSCOCO. We will from now on refer to this model as char-GRU, shorthand for character-based GRU.

2.4.2 Image-caption retrieval

We compare our char-GRU model with the current state-of-the-art in image-caption retrieval on both Flickr8k and MSCOCO. Table 2.2 shows the bidirectional retrieval results on both Flickr8k and MSCOCO. For MSCOCO we report both the results on the full test set (5000 items) and average results on a five-fold test set of 1000 items to be able to compare our results to previous work. Our models perform comparable to the state-of-the-art on both image to caption and caption to image retrieval on all metrics for Flick8k. The MSCOCO model by Faghri et al. (2017), which fine-tuned the ResNet-152 network during training, is the only model that significantly outperforms our own across the board.

All systems except the one by Wehrmann et al. (2018) and our own made use of word embeddings. Wehrmann et al. (2018) report that their CNN model trained on Flickr8k could only achieve such high recall scores when fine-tuning a model that was pretrained on MSCOCO, which they hypothesised is due to the small number of training examples in Flickr8k. Using our char-GRU model we outperform their convolutional approach without any pretraining on MSCOCO,

Table 2.2: Image-Caption retrieval results on the Flickr8k and MSCOCO test sets. R@N is the percentage of items for which the correct image or caption was retrieved in the top N (higher is better). Med r is the median rank of the correct image or caption (lower is better). We also report the 95 percent confidence interval for the R@N scores. For MSCOCO we report the results on the full test set (5,000 items) and the average results on five folds of 1,000 image-caption pairs.

Flickr8k	Caption to Image				Image to Caption			
	R@1	R@5	R@10	med r	R@1	R@5	R@10	med r
Klein et al. (2015)	21.2±1.1	50.0±1.4	64.8±1.3	5.0	31.0±2.9	59.3±3.0	73.7±2.7	4.0
Wehrmann et al. (2018)	26.9±1.2	-	69.6±1.3	4.0	32.4±2.9	-	73.6±2.7	3.0
Dong et al. (2018)	-	-	-	-	36.3±3.0	66.4±2.9	78.2±2.6	-
char-GRU	27.5±1.2	58.2±1.4	70.5±1.3	4.0	38.5±3.0	68.9±2.9	79.3±2.5	2.0
MSCOCO	1k results							
Vendrov et al. (2016)	37.9±0.6	-	85.9±0.4	2.0	46.7±1.4	-	88.9±0.9	2.0
Faghri et al. (2017)	52.0±0.6	84.3±0.5	92.0±0.3	1.0	64.6±1.3	90.0±0.8	95.7±0.6	1.0
Wehrmann et al. (2018)	40.4±0.6	-	88.6±0.4	2.0	49.5±1.4	-	91.3±0.8	1.6
char-GRU	41.4±0.6	76.8±0.5	88.0±0.4	2.0	51.2±1.4	83.5±1.0	92.1±0.7	1.2
MSCOCO	5k results							
Vendrov et al. (2016)	18.0±0.5	-	57.6±0.6	7.0	23.3±1.2	-	65.0±1.3	5.0
Faghri et al. (2017)	30.3±0.6	59.4±0.6	72.4±0.6	4.0	41.3±1.4	71.1±1.3	81.2±1.1	2.0
Kiela et al. (2018)	17.1±0.5	43.0±0.6	57.3±0.6	8.0	27.1±1.2	55.6±1.4	70.0±1.3	4.0
char-GRU	20.2±0.5	46.9±0.6	60.9±0.6	6.0	25.7±1.2	54.3±1.4	68.8±1.3	4.0

indicating that Flickr8k has enough training examples for a recurrent architecture to take advantage of.

2.4.3 Semantic evaluation

We now look at the semantic properties of the sentence embeddings in more detail and compare our models with previous work. Figure 2.3 displays Pearson's *r* scores on all the subtasks of the STS tasks for our char-GRU model, InferSent (Conneau et al., 2017), and a Bag Of Words (BOW) baseline using the average over a sentence's GloVe vectors.

Comparing Flickr8k with MSCOCO

First of all, our Flickr8k model significantly outperforms the MSCOCO model on 6 out of 26 tasks, while the MSCOCO model only outperforms the Flickr8k model on MSRvid, Images (STS 2014) and SICK-R. It seems that the larger amount of image-caption data in MSCOCO allows the model to become better at what it was already good at, that is, video and image descriptions. On the other hand, specialising in image and video descriptions seems to decrease the models' generalisation to other tasks indicating that it is overfitting. That being said, the Flickr8k model performs quite well, beating the InferSent and BOW models on some tasks and performing comparably on most of the other tasks even though

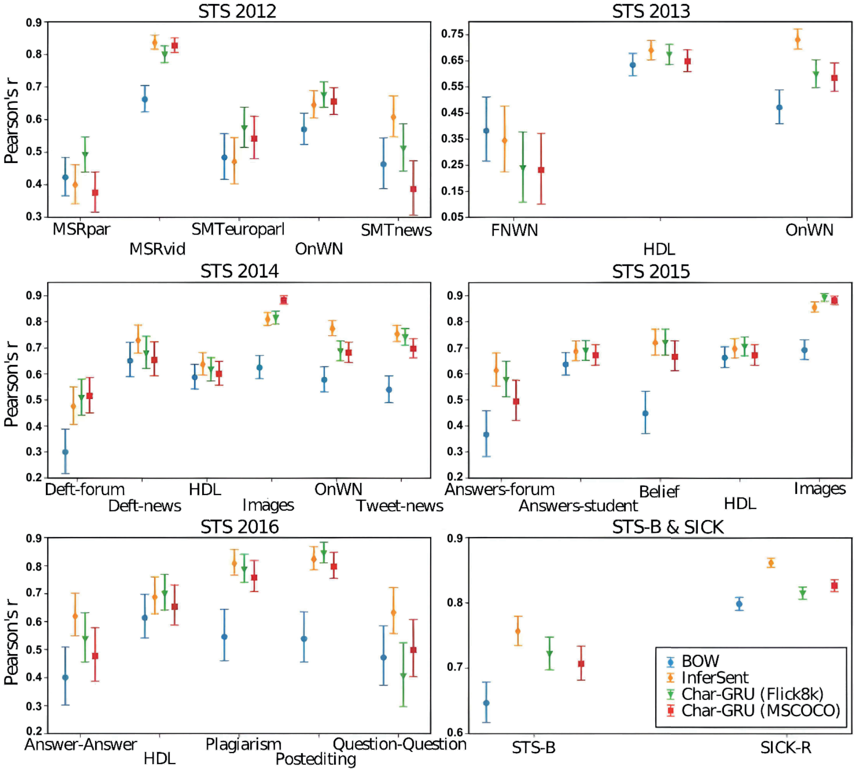

Figure 2.3: Semantic evaluation task results: Pearson correlation coefficients with their 95 percent confidence interval for the various subtasks (see Table 2.1). BOW is a bag of words approach using GloVe embeddings and InferSent is the model reported by Conneau et al. (2017). A supplement with a table of the results shown here is included in the github repository.

the Flickr8k database is only about five percent of the size of MSCOCO and about one percent of what InferSent is trained on.

Comparing with BOW baseline

It is important to note that models using GloVe vectors receive a considerable amount of prior lexical semantic knowledge. GloVe vectors are trained on an 840-billion-word corpus with a vocabulary of over 2.2 million words and InferSent gets all of this extracted semantic knowledge for free. If the model encounters a word in the transfer tasks that it has never seen during training, it still has knowledge of the word's semantic relatedness to other words through that word's GloVe vector.

This makes the BOW model a useful baseline model. It uses the prior word knowledge that InferSent uses (GloVe vectors) but it is not trained to create sentence embeddings. While InferSent is a significant improvement over the BOW model on most tasks (22 out of 26), it does not improve on the BOW model on 4 out of 26 tasks. Figure 2.3 shows that the BOW model performs close to the three trained models on many tasks. InferSent and the BOW model have the same input, but InferSent is trained on large amounts of data in order to extract information from this input. This then makes it reasonable to assume that a large part of InferSent's performance is due to the word level semantic information available in the GloVe vectors.

Our char-GRU model does not have such information available but instead benefits from being grounded in vision. By learning language from the ground up from multi-modal data, our model learns to capture sentence semantics with a performance comparable to models which receive prior knowledge of lexical semantics. Even though the system's only language input consists of image captions, Figure 2.3 shows that our model generalises well to a wide variety of domains. The Flickr8k model significantly outperforms the BOW baseline on 20 out of 26 tasks.

Comparing with InferSent

Next, we compare InferSent with our Flickr8k char-GRU in more detail. Our model performs on par with InferSent on 16 out of 26 tasks. It is not surprising that our char-GRU model performs well on the Images sets, with a significant improvement over InferSent on Images (STS 2015). Our char-GRU also outperforms InferSent significantly by quite a margin on SMTeuroparl (transcriptions from European Parliament sessions) and MSRpar (a news set scraped from the internet), both very different from each other and different from image captions. Table 2.3 contains examples of these datasets to highlight what we will discuss next.

On closer inspection, SMTeuroparl contains sentence pairs with high word overlap and relatively high similarity scores given by the human annotators. Even though word embedding based models should be just as capable of exploiting high word overlap as our char-GRU model, perhaps they are more prone to make mistakes if the two sentences differ by a very rare word such as 'pontificate' in the example. The embedding for such a rare word could be very skewed

Table 2.3: Example sentence pairs with their human-annotated similarity score taken from STS tasks.

Dataset	Similarity	Example pair
SMTeuroparl	3.5	We often pontificate here about being the representatives of the citizens of Europe. We are proud often here to represent the citizens of Europe.
MSRpar	4.0	South Asia follows, with 1.1 millions youths infected — 62 percent of them female. Of the 1.1 million infected in South Asia, 62 % are female.
FNWN	0.4	This frame contains words that describe an item's static position on a scale with respect to some property variable. Lacking in specific resources, qualities or substances.
Question-Question	4.0	How do I make a height adjustable desk? How can I build a wall mounted adjustable height desk?

towards an unrepresentative context when learning the embeddings. The MSRpar dataset contains many proper nouns for which no embedding might exist and it is common practice to then remove the word from the input. In contrast, our character-based method does not remove such proper nouns and thereby benefits from morphological similarity between the two sentences, even though the proper noun has never been seen before. Indeed, our model seems to work reasonably well on the other news databases as well, achieving state-of-the-art performance equal to InferSent on all HDL (news headlines) sets.

InferSent significantly outperforms our Flickr8k trained char-GRU model on 7 out of 26 tasks. Especially noticeable is our model's performance on the Question-Question (forum question) dataset and on FNWN (WordNet definitions), the only task where our model is outperformed significantly by the BOW model. FNWN contains definition-like sentences, often with structures that one does not find in an image description. In the example in Table 2.3, for instance, the first sentence of the pair is very lengthy and contains parentheses and abbreviations, while the second sentence is very short and lacks a subject. Concerning the question database, our model has never seen a question during training. Questions have a different syntactic structure than what our model has seen during training. Furthermore, most image descriptions tend to start with the word 'A' (e.g., 'A man scales a rock in the forest.'), whereas questions tend to start with 'What', 'Should' and 'How', for example.

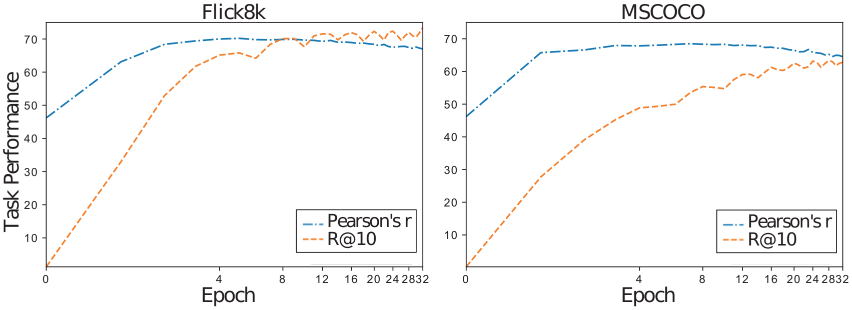

Figure 2.4: The training task performance (R@10) and the semantic task performance (Pearson's $r \times 100$) as they develop over training, with the number of epochs on a logarithmic scale. For MSCOCO (right) we show the training task performance on the 5,000 item test set.

Trade-off between training task and transfer task performance

We further investigate how prone our model is to overspecialising on image descriptions. Figure 2.4 shows how the bidirectional image-caption retrieval performance and the semantic task performance (SICK-R and STS12-16 combined) develop during training.

Epoch zero is the performance of an untrained model, and it is clear that both measures increase substantially during the first few epochs. Most improvement in both training task and semantic task performance happens in the first four epochs. After that the training task performance still increases by 12.8 and 28.5 percent for Flickr8k and MSCOCO, respectively. On the other hand, semantic task performance peaks around epoch four and then slowly decreases by 4.6 and 5.8 percent towards the last epoch for Flickr8k and MSCOCO, respectively. So even though our model is capable of learning how to extract semantic information from image-caption pairs, it is prone to overspecialising on the training task. The performance drop on the semantic task is only small, but trade-offs between the performance on different tasks pose a challenge to the search for universal sentence embeddings.

2.5 Conclusion

We investigated whether sentence semantics can be captured in sentence embeddings without using (prior) lexical knowledge. We did this using a multi-modal encoder which grounds language in vision using image-caption pairs. Harwath and Glass (2015) have claimed that this method produces a multi-modal semantic embedding space and, indeed, we found that the distances between resulting sentence embeddings correlate well with human semantic similarity judgements, in some cases more so than models based on word embeddings. Importantly, this shows that we do not need to use word embeddings, which has hitherto been the standard in sentence embedding methods. The addition of visual information during training allows our model to capture semantic information from character-level language input. The model generalises well to linguistic domains such as European Parliament transcriptions, which are very different from the image descriptions it was trained on, but our model also has difficulty with some of the subtasks. For instance, our model scored significantly lower than InferSent on the SICK and forum question databases suggesting that our grounding approach alone is not enough to learn semantics for all linguistic domains. This could be because some visual information is hardly ever explicitly written down (few people will write down obvious facts like 'bananas are yellow'), while more abstract concepts will not appear in images or their descriptions (e.g., the words 'intent' and 'attempted' from our test sentences in Table 2.3 are hard to capture in an image). Future work could combine the visual grounding approach with text-only methods in order to learn from more diverse data. In such a multi-task learning setting, our grounded sentence encoder could be fine-tuned on for instance natural language inference data, combining our approach with that of InferSent (Conneau et al., 2017).

In future work, we plan to work on spoken utterances. Unlike text, speech is not neatly segmented into lexical units, posing a challenge to conventional word embedding methods. However, the results presented here show that it is possible to learn sentence semantics without such prior lexical semantic knowledge and segmentation into lexical units. So far, studies of sentence meaning have mostly focused on written language, even though we learn to listen and speak long before we learn how to read and write. Learning representations of sentence meaning directly from speech therefore seems more intuitive than separately learning word and sentence representations from written sources. Furthermore, most languages have no orthography and only exist in spoken form. Capturing semantics directly from the speech signal provides a way to model sentence semantics for these languages. While there is previous work on spoken caption-image retrieval (e.g., Harwath et al. 2016; Chrupała et al. 2017) we have barely scratched the surface of transfer learning using spoken input.

3 | Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge

Distributional semantic models capture word-level meaning that is useful in many natural language processing tasks and have even been shown to capture cognitive aspects of word meaning. The majority of these models are purely text based, even though the human sensory experience is much richer. In this paper we create visually grounded word embeddings by combining English text and images and compare them to popular text-based methods, to see if visual information allows our model to better capture cognitive aspects of word meaning. Our analysis shows that visually grounded embedding similarities are more predictive of the human reaction times in a large priming experiment than the purely text-based embeddings. The visually grounded embeddings also correlate well with human word similarity ratings. Importantly, in both experiments we show that the grounded embeddings account for a unique portion of explained variance, even when we include text-based embeddings trained on huge corpora. This shows that visual grounding allows our model to capture information that cannot be extracted using text as the only source of information.

3.1 Introduction

Distributional semantic models create word representations that quantify word meaning based on the idea that a word's meaning depends on the contexts in which the word appears. Such representations (also called embeddings) are widely used as the linguistic input for computational linguistic models, with research showing that they can account for response times in lexical decision tasks (Mandera et al., 2017; Rotaru et al., 2018; Petilli et al., 2021), decode brain data

This chapter is based on: Danny Merkx, Stefan L. Frank and Mirjam Ernestus. Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1-11. ACL, 2022

(Xu et al., 2016; Abnar et al., 2018), account for brain activity during text comprehension (Frank and Willems, 2017), and correlate with human judgements of word similarity (Kiela et al., 2018; Derby et al., 2018, 2020).

While such embeddings have proven useful, they are not cognitively plausible as creating high quality embeddings requires billions of word tokens. For instance, the GloVe embeddings developed by Pennington et al. (2014) are trained on 840 billion words. It would take a human 80 years of constant reading at about 330 words per second to digest that much information. Obviously, humans are able to understand language after much less exposure, and furthermore, their sensory experience is much richer than solely reading texts.

Embodied cognition theory poses that our conceptual knowledge is based on the entirety of our sensory experience (Barsalou, 2008; Foglia and Wilson, 2013). For instance, reading the word *dog* elicits sensory experiences we have with dogs, such as their sound and how they look. Embodied cognition theory thus assumes that all our sensory experiences contribute to our conceptual knowledge and processing, which should be reflected in human behaviour. Early priming studies have indeed found that visual similarities can elicit priming effects (D'Arcais et al., 1985; Schreuder et al., 1998).

If visual features are part of our conceptual knowledge, word embeddings incorporating visual features should be able to explain human behavioural data to a degree unattainable by purely text-based methods (that is, if we assume visual sensory experiences can never be fully captured by textual descriptions). That is why recent research has taken an interest in multi-modal word embeddings, combining text with a second source of information, resulting in visually grounded embeddings (VGEs) in the case of visual information.

3.1.1 Related work

Using image tags as a source of visual context, Bruni et al. (2013) create visual distributional semantic embeddings and use dimensionality reduction to map visual and text-based embeddings to the common VGE space. Derby et al. (2018) combine text-based embeddings with the network activations of an object recognition model and show that these visual features improve the embeddings' performance in downstream tasks. Petilli et al. (2021) use visual embeddings created by an object recognition network, and show that the embedding similarities are predictive of priming effects over and above text-based similarities.

The studies described above involve separately trained word and visual embeddings. An end-to-end approach to combine visual and linguistic information is through a deep neural network based caption-to-image retrieval (C2I) models (e.g., Karpathy and Fei-Fei 2015; Kamper et al. 2017a). While these models are trained to encode images and corresponding written or spoken captions in a common embedding space such that relevant captions can be retrieved given an image and vice versa, the resulting embeddings have been shown to capture sentence-level semantics (e.g. Chrupała et al. 2017 and **Chapters 2** and **5**). Kiela et al. (2018) showed that pretrained embeddings correlated better with human intuition about word meaning after being fine-tuned as learnable parameters in their C2I model.

3.1.2 Current study

In this study we investigate whether VGEs created by a C2I model explain human behavioural data. Our research question is: can VGEs capture aspects of word meaning that (current) text-based approaches cannot? To answer this question we investigate novel end-to-end trained VGEs and test them on two types of human behavioural data thought to rely on conceptual/semantic knowledge. Secondly, we take care to separate the contribution of the image modality from that of the linguistic information to see whether visual grounding captures word properties that cannot be learned by purely text-based methods. We do this by comparing our VGEs to three well-known text-based methods.

Throughout our experiments we will use two versions of the text-based methods: custom trained on the same data as our VGEs and pretrained on large corpora. From a cognitive modelling perspective, the former of these is more interesting. While the use of large corpora may not be problematic for natural language processing applications where performance comes first, we aim to create cognitively plausible embeddings, that is, from a realistic amount of linguistic exposure. However, the inclusion of pretrained embeddings serves to answer our main research question.

Semantic similarity judgements

In our first experiment we test whether the VGEs correlate better with a measure of human intuition about word meaning than text-based embeddings. A well-known method to capture human intuition about word meaning is simply by asking subjects how similar two words are in meaning. To evaluate word embeddings, one can then see if embedding similarities for those word pairs correlate with the human judgements (e.g., Bruni et al., 2013; Baroni et al., 2014; Speer and Chin, 2016; Kiela et al., 2018; Derby et al., 2020).

While the study by Kiela et al. (2018) performed a similar investigation on pretrained word embeddings fine-tuned through their C2I model, they did not take into account the fact that text might also contain visual knowledge. It is not unreasonable to assume that some visual knowledge can be gained from a large corpus of sentences solely describing visual scenes. We account for this visual knowledge from text by incorporating word embeddings trained on the image descriptions in order to investigate the contribution of the *image* modality included in the VGEs.

Collecting word similarity ratings typically involves showing participants two words and asking them to rate how similar or related their meanings are, or picking the most related out of several pairs. Semantic relatedness refers to the strength of the association between two word meanings. For instance, 'dog' and 'leash' have a strong relationship but are not similar in meaning. Semantic similarity refers to two words sharing semantic properties, for instance 'dogs' and 'cats' which are both animals that people keep as pets (Hill et al., 2015).

Semantic priming

In the second experiment, we test whether our VGEs are predictive of semantic priming effects from a large priming experiment (Hutchison et al., 2013). Semantic priming effects occur when activation of a semantically related prime word facilitates the processing of the target word, resulting in shorter reaction times. If all our sensory experiences contribute to word meaning, we would expect visual perceptual properties of the prime-target pair to influence the response times.

Petilli et al. (2021) performed a similar experiment using visual embeddings derived from activation features from an object recognition network and text-based word embeddings. Their results show that after accounting for the text-based similarity, the visual embedding similarities contribute to explaining the human reaction times only for lexical decision trials with a short stimulus onset asynchrony (SOA), and not for the naming task or long SOA trials. They attribute this to: 1) the lexical decision task being more sensitive to semantic effects than the naming task (Lucas, 2000), and 2) visual information being activated in early linguistic processing and rapidly decaying (Pecher et al., 1984; Schreuder et al., 1998). We will further test these interactions in our own experiment.

3.2 Methods

In our experiments, we compare the VGEs from our own model with three well known text-based distributional semantic models: FastText (Bojanowski et al., 2017), Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). For the purpose of this study, we take two approaches: 1) we train our own text-based distributional models to allow for a fair comparison to the VGEs, and 2) we use the pretrained models to investigate whether our VGEs capture semantic information that even models trained on large text corpora do not. We released the code for this project on github: `https://github.com/DannyMerkx/speech2image/tree/CMCL2022`.

3.2.1 Training data

MSCOCO is a database intended for training image recognition, segmentation and captioning models (Chen et al., 2015). It has 123,287 images and 605,495 written English captions, that is, five captions paired to each image. Captions were collected by asking annotators to describe what they saw in the picture. Five thousand images (25,000 captions) are reserved as a development set.

The captions are provided in tokenised format. In order to use them in our models we only de-capitalised all words and removed the punctuation at the end of each sentence. This results in a total of 6,184,656 word tokens and 28,415 unique word types, to which we add start- and end-of-sentence tokens for training our visually grounded model.

The images are preprocessed by resizing the images such that the shortest side is 256 pixels, while keeping the original aspect ratio. We take ten 224 by 224 crops of the image: one from each corner, one from the middle and the same five crops for the mirrored image. We use ResNet-152 (He et al., 2016) pretrained on ImageNet to extract visual features from these ten crops and then average the features of the ten crops into a single vector with 2,048 features. These features are extracted by removing ResNet's classification layer and taking the activations of the penultimate layer.

3.2.2 Models

Visually grounded model

Our visually grounded model is based on the implementation presented in **Chapter 2**, and we refer to that chapter for the details. Here we will provide a brief overview of the model, any differences with the original model and the parameter settings tested in this study.

The VGE model maps images and their corresponding captions to a common embedding space. It is trained to make the embeddings for matching images and captions as similar as possible, and those for mismatched images and captions dissimilar. The model consists of two parts; an image embedder and a caption embedder. The image embedder is a single-layer linear projection on top of the image features extracted with ResNet-152. We train only the linear projection and do not further fine-tune ResNet.

The caption embedder consists of a word embedding layer followed by a two-layer bi-directional recurrent Long Short Term Memory (LSTM) layer and finally a self-attention layer. The embedding layer has 300 dimensions and is used to represent the input words as learnable embeddings. The purpose of the LSTM is to create a contextualised hidden state for each time-step (input word). Its first layer has 1028 hidden units, while its second layer acts as a bottleneck with 300 hidden units. Finally, the purpose of the attention layer is to weigh each time-step in order to create a single fixed-length embedding for the entire caption. The attention layer has 128 hidden units.

The image embedder has 2 × 300 dimensions so that the output matches the size of the caption embeddings. Both image and caption embedding are L2 normalised and we take their distance as the loss signal for the batch hinge loss function (see section 2.3.1). The networks are trained for 32 epochs using Adam with a cyclic learning rate schedule based on Smith (2017), which varies the learning rate smoothly between 10^{-3} and 10^{-6}.

The obvious way to extract word embeddings from the trained model would be to use the trained weights of the embedding layer. Unlike for instance in GloVe, where each word's embedding is based on its full co-occurrence distribution, these embeddings are not trained specifically to capture word context or meaning and they are not necessarily the best word embeddings. Our initial tests showed that they indeed performed very poorly as semantic embeddings when trained from a random initialisation [1]. Rather than taking the input embeddings we create our own embeddings from the hidden representations of the model.

We create our VGEs from the hidden activations of the bottleneck LSTM layer. We use the trained caption encoder to encode all training sentences in MSCOCO.

[1] Kiela et al. (2018) were able to use the input embeddings because they were initialised using pretrained embeddings.

However, we remove the attention layer that creates the sentence embedding and we retain the individual activations of the LSTM at each time step. As the word representations in this layer can be used to create semantic sentence embeddings that capture human intuition about sentence meaning (as shown for instance in **Chapters 2** and **5**), we expect these representations to better capture word meaning than the input embeddings.

The embedding for each word is then created by summing and normalising its LSTM layer activations from all its occurrences in the dataset. As opposed to the model in **Chapter 2**, where we used a single recurrent layer and found no further benefit of additional layers in terms of sentence embedding quality, we found that the quality of our VGEs improves when we use a two-layer LSTM, with the second layer acting as a bottleneck from which we derive the embeddings.

Text-based models

The text-based distributional models are trained on the MSCOCO captions. We train Word2Vec and FastText using the *Gensim* package (Řehůřek and Sojka, 2010). We train GloVe using the code that Pennington et al. (2014) made publicly available[2].

Word2Vec and FastText were trained as the Skip-gram variant with embedding size 300, a context window of 10 and 10 negative samples. GloVe was trained with embedding size 300 and a context window of 10. All resulting word embeddings are then L2 normalised.

In addition, we use the following pretrained vectors (all 300 dimensional): Word2Vec trained on 100 billion tokens of the Google News corpus (Mikolov et al., 2013b), FastText trained on 600 billion tokens of Common Crawl (Mikolov et al., 2018) and GloVe trained on 840 billion tokens of Common Crawl (Pennington et al., 2014).

3.2.3 Evaluation data

Semantic similarity judgements

We include both semantic relatedness and similarity datasets in our analysis. It has been argued that subjects' intuitive understanding of similarity is not necessarily in line with the 'scientific' notions of similarity and relatedness explained in the introduction (Hill et al., 2015). Thus, if subjects are not clearly instructed on

[2] `https://nlp.stanford.edu/projects/glove/`

Table 3.1: Description of the word similarity/relatedness evaluation datasets. #available is the number of word pairs included in the evaluation. Type indicates whether the dataset captures similarity or relatedness. NA indicates subjects were not specifically instructed on the difference.

Dataset	#word-pairs	#available	type
WordSim353	353	240	NA
WordSim-S	203	147	Similarity
WordSim-R	252	166	Relatedness
SimLex999	999	793	Similarity
-SimLex999 Q1	249	141	Similarity
-SimLex999 Q4	250	249	Similarity
MEN	3000	2889	Relatedness
RareWords	2034	204	NA

these notions of similarity or relatedness, we consider the nature of the dataset undefined.

The WordSim353 dataset by Finkelstein et al. (2002) contains 353 word pairs annotated with similarity ratings. While the name suggests it is a similarity rating dataset, more recent studies consider it a hybrid dataset, as subjects were not specifically instructed to judge relatedness or similarity. In a later study by Agirre et al. (2009), the WordSim353 data was split into similar and related pairs by annotating the word pairs. WordSim-S (similar) contains word pairs annotated as being synonyms, antonyms, identical, or hyponym-hyperonym. WordSim-R (related) contains word pairs annotated as being meronym-holonym, and pairs with none of the above relationships but with a similarity score greater than 5 (out of 10). Both sets contain all unrelated words (words not annotated with any of the above relationships and a similarity lower than 5).

SimLex999 was created with the caveats of the original WordSim353 in mind in order to create a dataset of 999 word pairs annotated for similarity rather than relatedness (Hill et al., 2015). SimLex999 furthermore contains concreteness ratings for the word pairs. Hill et al. (2015) divided the dataset into concreteness quartiles based on the sum of the concreteness ratings for each pair. Using these quartiles we also look at the 25% most concrete word pairs versus the 25% most abstract pairs in the dataset, of course expecting our grounded model to perform best on the concrete words.

MEN contains 3000 word pairs annotated for semantic relatedness (Bruni et al., 2013). Ratings were collected by showing subjects two word pairs and asking them to select the most related one. MEN was specifically collected to test multi-modal models, by selecting only words that have a visual referent that appeared in a large image database.

The RareWords dataset contains 2034 word pairs, where at least one word of each pair has a low frequency in Wikipedia (Luong et al., 2013). Modelling low-frequency words is a challenge for many models of distributional semantics.

Not all of the words in these databases are available in our training data and thus some will not have a word embedding. Table 3.1 contains an overview of the datasets described here and the number of word pairs that could be entered in our evaluations.

Semantic priming

The Semantic Priming Project (SPP) dataset (Hutchison et al., 2013) contains lexical decision times and naming times from a large priming experiment. The database is large for its kind, with 1,661 target words (and 1,661 non-words for the lexical decision task), each paired with a strong and weak prime and two unrelated primes. Furthermore, each prime-target pair was presented with a short (200ms) and a long (1200ms) SOA. Every combination of prime-target and SOA received responses from 32 subjects.

This gives us 26,576 (1661 target words × 4 priming conditions × 2 SOAs × 2 tasks) trials (disregarding the non-word word trials). We preprocessed the data by removing target words that mistakenly had more or fewer than the required four primes, trials with erroneous responses and missing data. We also lowered any capitals in the prime and target words, averaged the response times over the 32 subjects, and removed any prime-target pair that did not occur in our training data, resulting in 18,326 datapoints.

3.2.4 Analysis

Semantic similarity judgements

To test whether the word embedding models capture human intuitions on word similarity, we use the models to calculate embedding cosine similarities for each word pair and correlate them with the human annotations. From the correlations r we derive R^2 values, that is, the percentage of variance in the human similarity judgements that is explained by the model similarity scores. This allows us to evaluate our custom trained word embeddings to see which method best extracts word-level semantics from the MSCOCO dataset.

Next, we also compute semi-partial correlations between the human annotations and our VGE model using each of the text-based models as a control. Simply put, the semi-partial correlation between the VGE similarities and human annotations removes the effect of the control (i.e., text-based similarities) from the VGE similarities. Semi-partial R^2 gives us the percentage of variance that is uniquely explained by the VGE similarities. Given that all models are trained on the same textual data, with only the VGEs having access to the visual modality, this allows us to see whether visual grounding captures information that the text-based methods do not.

Finally we also test the semi-partial correlations using the pretrained embeddings as a control. For each pretrained model we also add in its custom MSCOCO-trained equivalent as a control, to take into account the information that text-based models can extract from the MSCOCO captions.

Semantic priming

Using linear regression models, we analyse how well embedding similarities predict human (log-transformed) reaction times in the SPP data using the Statsmodels package in Python (Seabold and Perktold, 2010). We code SOA and Task as factor variables. The reaction times are not on the same scale due to differences in the required response for the lexical decision and naming tasks so we standardise the log-transformed reaction time data separately for each combination of SOA and Task. This removes the main effects of SOA and Task but we include them in the regression as we are interested in their interactions with the similarity measures.

We fit a baseline regression including the target length (number of characters), Task and SOA as regressors. We furthermore include several regressors based on SUBTLEX-US (Brysbaert and New, 2009): log-transformed word-frequency counts, contextual diversity (the number of SUBTLEX-US documents a word appears in) and the orthographic neighbourhood density (the number of SUBTLEX-US words that are one character edit away) for the target words.

Next, for each of our embedding models, we include the prime-target embedding similarities as a regressor to the baseline model. We also add two two-way interactions to test the claims made in Petilli et al. (2021): 1) the interaction between the embedding similarities and Task to test the difference between lexical decision and naming in terms of sensitivity to semantic effects and 2) the interaction between the embedding similarities and SOA to test their claim about the time-frame in which visual information plays a role. These regression models allow us to compare the word embedding models to each other and to the baseline using the Akaike Information Criterion (AIC), where a lower AIC indicates a better model fit.

We also test if our VGEs can explain variance in the human reaction times that the text-based methods do not. We do this by refitting the regression models for each of the text-based similarity measures and adding the VGE similarity measures and their interactions with Task and SOA as extra regressors. For each of these regressions we then calculate the log-likelihood ratio (LLR) with the corresponding regression without the VGEs, indicating the decrease in model deviance due to adding the VGE similarity measures. Higher LLRs indicate a larger contribution of the VGEs to explaining variance in the human response times beyond what the text-based embedding similarities explain. Because the LLR follows a χ^2 distribution, we can test whether including the VGEs significantly improves the regression model.

We apply a similar approach to the pretrained text-based embeddings, but we also want to account for the information that text-based embedding models can extract from the MSCOCO captions. We do this by fitting a regression model as in the previous step except that we include both the pretrained and MSCOCO trained embeddings and their interactions with SOA and Task. We then follow the same procedure as described above by adding the VGE similarities and calculate LLRs to see if adding VGEs improves the regression fit.

3.3 Results

3.3.1 Semantic similarity judgements

Figure 3.1 shows the R^2 (explained variance) based on the Pearson correlation coefficients between the human similarity annotations and the embedding similarities. On top of the text-based R^2 values, we display the semi-partial R^2 of the VGEs using the text-based model as control. As total explained variance equals the semi-partial R^2 plus R^2 of the control(s), this clearly visualises both the total amount of explained variance and the amount of *extra* variance that is uniquely explained by the VGEs. All Pearson correlations were positive, as expected, except for two non-significant semi-partial correlations which are therefore not included in the figure.

For the MSCOCO models (left panel) we see that while GloVe has the worst performance on each dataset, there is no single best model. Furthermore, while

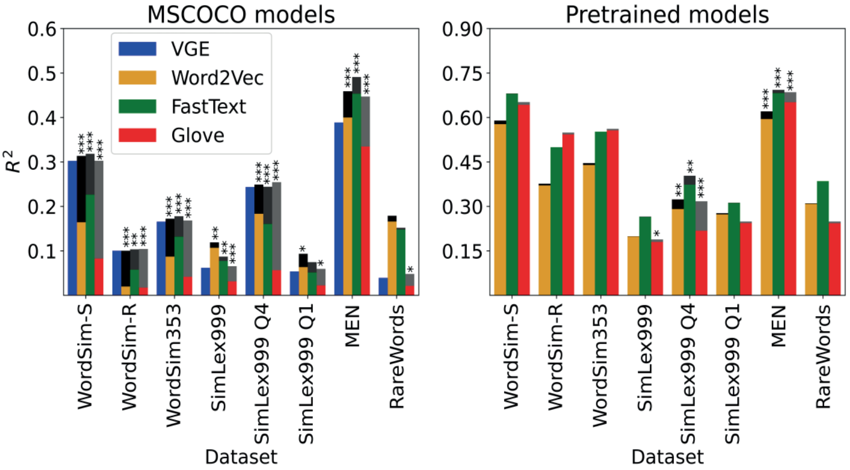

Figure 3.1: The coloured bars indicate the R^2 scores of the four word embedding models. The grey-scale bars on top of the R^2 scores of the text-based models indicate the semi-partial R^2 scores and their significance ($*p < .05, **p < .01, ***p < .001$, corrected using the Benjamini and Hochberg (1995) procedure with a false discovery rate of 0.05) of the VGEs after controlling for the variance explained by that text-based model. Left panel: models trained on MSCOCO. Right panel: pretrained text-based models.

the VGEs are outperformed by FastText and Word2Vec on SimLex999, we see that the VGEs perform best on the most concrete words (Q4) in SimLex999. A bit surprising then, is that the VGEs are outperformed by FastText and Word2Vec on MEN, which contains solely picturable nouns.

Looking at the semi-partial R^2, that is, the extra variance explained by the VGEs after controlling for one of the other embedding models, we see that for nearly every dataset and every model, the VGEs explain a significant portion of variance that is not explained by the text-based models. This is not very surprising on WordSim, where the VGEs were the best performing embeddings by quite a margin. However, we also see that even though the VGEs are outperformed by FastText and Word2Vec on MEN, they still explain a large extra portion of variance even though the R^2 for these models was already quite high.

Lastly, the pretrained models (right panel) outperform the MSCOCO models. This was expected, as the used training data is several orders of magnitude larger than MSCOCO. However, the semi-partial correlations still show that the VGEs explain a significant portion of extra variance on SimLex999 Q4 and MEN.

Table 3.2: AIC comparison of regression models (lower is better). Δ indicates the difference in AIC compared to the VGE model or the Baseline model. β indicates the coefficient of the embedding similarity main effect (lower is better) and its significance.

Model	AIC	ΔVGE	ΔBaseline	β
VGE	46997.55	—	−211.04	−.67***
FastText	47101.90	104.35	−106.86	−.54***
GloVe	47163.70	166.15	−44.88	−.20**
Word2Vec	47184.45	186.90	−24.13	−.22**
Baseline	47208.58	211.03	—	—

Table 3.3: LLRs between regression models with the indicated text-based similarity measures and the same model with the VGE similarities as extra regressors. β VGE are the regression coefficients for the VGE similarities in each model. Higher LLRs indicate a larger improvement in model quality due to adding the VGEs.

	MSCOCO		+ Pretrained	
	LLR	β VGE	LLR	β VGE
Word2Vec	193.72***	−.77***	69.72***	−.49***
FastText	111.46***	−.63***	47.32***	−.42***
GloVe	168.34***	−.72***	49.80***	−.36***

3.3.2 Semantic priming

The ΔAIC scores in Table 3.2 show that all word embedding models trained on MSCOCO improve the regression fit above the baseline. The embedding similarity effects were all negative, that is, a higher similarity correctly predicts a lower reaction time. We furthermore see that the VGE-derived similarity measures result in the best model fit by quite a margin, as evidenced by the AIC scores and effect size.

We also find significant interactions between Task and the embedding similarities for the VGE ($\beta = 0.201, P = 0.009$) and FastText regression models ($\beta = 0.197, P = 0.027$), meaning that the effect of embedding similarity is stronger for the lexical decision task. We find no significant interactions between the embedding similarities and SOA.

Table 3.3 shows the LLRs between the regression models including the (pretrained) text-based and our VGE word similarity measures and the corresponding model including only the text-based measures. We see that our VGEs significantly improve the regression fit for every type of text-based method, even when we include both the pretrained and MSCOCO text-based measures. The coefficients of the VGE effects in these models are all negative, meaning a higher VGE similarity predicts a lower reaction time.

In the regression models including the VGEs and the MSCOCO text-based embeddings we found significant interactions between the VGE similarities and Task in the regression models that also include Word2Vec ($\beta = 0.239, P = 0.007$) or GloVe ($\beta = 0.234, P = 0.01$) and no other interactions with Task or SOA.

Lastly, in the regression models including the VGEs and both pretrained and MSCOCO text-based embeddings, we find significant interactions with Task for Word2Vec ($\beta = 0.312, P < 0.001$), FastText ($\beta = 0.297, P = 0.001$) and GloVe ($\beta = 0.443, P < 0.001$) vectors, and none for the VGEs.

3.4 Discussion

We created Visually Grounded Embeddings using a caption-image retrieval model in order to test if these embeddings can capture information about word meaning that text-based approaches cannot. Importantly, by testing our VGEs on human behavioural measures typically thought to rely on conceptual/semantic knowledge, we test a central idea of embodied cognition theory, namely that our visual experiences contribute to our conceptual knowledge.

3.4.1 Semantic similarity judgements

Our first experiment showed that, when trained on the same corpus, our VGEs are on par with text-based methods. While there is no clear overall best method, the VGEs perform well on WordSim and, as might be expected, on the datasets with concrete picturable nouns. Even though the text-based methods outperform the VGEs on one of these (MEN), the VGEs still explain a significant amount of extra variance over and above what is explained by the text-based methods. This indicates that the text-based embeddings and VGEs capture non-overlapping conceptual knowledge, which we attribute to the visual grounding of the VGEs, given that the training materials were otherwise equal.

The only database where the VGEs performed notably worse than the text-based methods was RareWords. This is perhaps because during training, the VGEs are grounded in the image corresponding to the text input, even if not all words in the sentence are visible in the picture. As the words in RareWords are generally not picturable nouns, any visual information incorporated into the word-embedding is unlikely to be helpful, or, as evidenced by the results, counterproductive.

We furthermore found that our VGEs explain additional variance in the human similarity ratings even after accounting for both the MSCOCO text-based models and pretrained models trained on massive text corpora. The fact that the VGEs explain a significant amount of extra variance even after the text-based models have seen billions of tokens of text, suggests that some aspects of word meaning cannot be captured solely from text and that visual similarity plays a role in human intuition about word meaning.

3.4.2 Semantic priming

In our second experiment, the VGEs outperformed the text-based methods on explaining human reaction times from the Semantic Priming Project. Even after we account for both the MSCOCO text-based models and pretrained models in our regression, the VGEs still explain a significant amount of variance in the reaction times.

In previous work, Petilli et al. (2021) only found a significant contribution of visual information in the short SOA lexical decision task. We found no further proof for their hypothesis that visual information is activated in early linguistic processing and thereafter rapidly decays. Rather, we find that our VGEs improve the model quality for both short and long SOA trials.

We did find a significant positive interaction with Task, meaning that the word embeddings explain less variance in the naming task than in the lexical decision task. This interaction was not specific to the VGEs but also occurred in the models including FastText and for all the pretrained embeddings. As claimed in Petilli et al. (2021) and Lucas (2000) this suggests that naming tasks are in general less sensitive to semantic effects.

3.5 Conclusion

We set out to test an end-to-end approach to combining visual and textual input in a single embedding, trained on a cognitively plausible amount of data. The results from our two experiments suggest that VGEs capture aspects of word meaning that text-based approaches cannot. Even though we include word embeddings trained on corpora several orders of magnitude greater than any human's exposure to language, our VGEs still explain a unique portion of variance in both human behavioural measures.

While our results indicate that visual grounding can provide complementary information for certain words, it may not play a role in our conceptual knowledge of rare, abstract words, as shown by our results on the RareWords corpus. Similar to Petilli et al. (2021) this then does not support the strongest formulations of embodied cognition theory which suggest total equivalence between conceptual and sensorimotor processing (Glenberg, 2015).

Of course, one could always claim that it is just current word-embedding models that do not fully capture word meaning yet. However, given that VGEs trained on a relatively small amount of visual data can complement text-based embeddings, we do not think even larger text-corpora or more complex embedding models can ever fully capture human semantic knowledge. The human experience is rich and varied, and our computational models can never fully capture human word knowledge while ignoring visual aspects of this experience.

4 Language learning using speech to image retrieval

Humans learn language by interaction with their environment and listening to other humans. It should also be possible for computational models to learn language directly from speech but so far most approaches require text. We improve on existing neural network approaches to create visually grounded embeddings for spoken utterances. Using a combination of a multi-layer GRU, importance sampling, cyclic learning rates, ensembling and vectorial self-attention our results show a remarkable increase in image-caption retrieval performance over previous work. Furthermore, we investigate which layers in the model learn to recognise words in the input. We find that deeper network layers are better at encoding word presence, although the final layer has slightly lower performance. This shows that our visually grounded sentence encoder learns to recognise words from the input even though it is not explicitly trained for word recognition.

4.1 Introduction

Most computational models of natural language processing (NLP) are based on written language; machine translation, sentence meaning representation and language modelling to name a few (e.g., Wang et al. 2018; Kiros et al. 2015). Even if the task inherently involves speech, such as in automatic speech recognition, models require large amounts of transcribed speech (Wang and Wang, 2016). Yet, humans are capable of learning language from raw sensory input, and furthermore children learn to communicate long before they are able to read. In fact, many languages have no orthography at all and there are also languages of which the writing system is not widely used by its speakers. Text-based models cannot be used for these languages and applications like search engines and automated translators cannot serve these populations.

This chapter is based on: Danny Merkx, Stefan L. Frank and Mirjam Ernestus. Language learning using speech to image retrieval. In *Interspeech 2019 - 20th Annual Conference of the International Speech Communication Association*, pages 1841-1846, 2019

There has been increasing interest in learning language from more natural input, such as directly from the speech signal, or multi-modal input (e.g., speech and vision). This has several advantages such as removing the need for expensive annotation of speech, being applicable to low resource languages and being more plausible as a model of human language learning.

An important challenge in learning language from spoken input is the fact that the input is not presented in neatly segmented tokens. An auditory signal does not contain neat breaks in between words like the spaces in text. Furthermore, no two realisations of the same spoken word are ever exactly the same. As such, spoken input cannot be represented by conventional word embeddings (e.g., Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014)). These text-based embeddings are trained to encode word-level semantic knowledge and have become a mainstay in work on sentence representations (e.g., Conneau et al. 2017; Kiela et al. 2018). When we want to learn language directly from speech, we will have to do so in a more end-to-end fashion, without prior lexical level knowledge in terms of both form and semantics.

In **Chapter 2** we used image-caption retrieval, where given a written caption the model must return the matching image and vice versa. We trained deep neural networks (DNNs) to create sentence embeddings without the use of prior knowledge of lexical semantics (see Kiela et al. 2018; Karpathy and Fei-Fei 2015; Faghri et al. 2017 for other studies on this task). The visually grounded sentence embeddings that arose capture semantic information about the sentence as measured by the Semantic Textual Similarity task (see Agirre et al. 2016), performing comparably to text-only methods that require word embeddings.

In the current study we present an image-caption retrieval model that extends our previous work to spoken input. Harwath and colleagues adapted text-based caption-image retrieval (e.g., Karpathy and Fei-Fei 2015) and showed that it is possible to perform speech-image retrieval using convolutional neural networks on spectral features (Harwath and Glass, 2015; Harwath et al., 2016). Our work is most closely related to the models presented by Harwath and Glass (2015), Harwath et al. (2016, 2020b) and Chrupała et al. (2017). In the current study we improve upon these previous approaches to visual grounding of speech and present state-of-the-art image-caption retrieval results.

The work by Harwath and Glass (2015), Harwath et al. (2016, 2020b), Chrupała et al. (2017) and the results presented here are a step towards more cognitively plausible models of language learning as it is more natural to learn language without prior assumptions about the lexical level. For instance, research indicates that the adult lexicon contains many relatively fixed multi-word expressions (e.g., 'how-are-you-doing') (Tomasello, 2000). Furthermore, early during language acquisition the lexicon consists of entire utterances before a child's language use becomes more adult-like (Tomasello, 2000; Braine and Bowerman, 1976; Pine and Lieven, 1993; Lieven et al., 2003). Image to spoken-caption retrieval models do not know a priori which constituents of the input are important and have no prior knowledge of lexical level semantics. We probe the resulting model to investigate whether it learns to recognise lexical units in the input without being explicitly trained to do so.

We test two types of acoustic features; Mel Frequency Cepstral Coefficients (MFCCs) and Multilingual Bottleneck (MBN) features. MFCCs are features that can be computed for any speech signal without needing any other data, while the MBN features are 'learned' features that result from training a network on top of MFCCs in order to recognise phoneme states. While MBN features have been shown to be useful in several speech recognition tasks (e.g., Nguyen et al. 2014; Fer et al. 2017), learned audio features face the same issue as word embeddings, as humans learn to extract useful features from the audio signal as a result of learning to understand language and not as a separate process. However, the MBN features can still be useful where system performance is more important than cognitive plausibility, for instance in a low resource setting. Furthermore, these features could provide a clue as to what performance would be possible if we had more sophisticated models or more data to improve the feature extraction from the MFCCs in an end-to-end fashion.

In summary, we improve on previous spoken-caption to image retrieval models and investigate whether it learns to recognise words in the speech signal. We show that our model achieves state-of-the-art results on the Flickr8k database, outperforming previous models by a large margin using both MFCCs and MBN features. We find that our model learns to recognise words in the input signal and show that the deeper layers are better at encoding this information. Recognition performance drops a little in the last two layers as the network abstracts away from the detection of specific words in the input and learns to map the utterances to the joint embedding space. We released the code for this project on github: `https://github.com/DannyMerkx/speech2image/tree/Interspeech19`.

4.2 Image to spoken-caption retrieval

4.2.1 Materials

Our model is trained on the Flickr8k database (Hodosh et al., 2013). Flickr8k contains 8,000 images taken from online photo sharing application Flickr.com, for which five English captions per image are available. Annotators were asked to 'write sentences that describe the depicted scenes, situations, events and entities (people, animals, other objects)'. Spoken captions for Flickr8k were collected by Harwath and Glass (2015) by having Amazon Mechanical Turk workers pronounce the original written captions. We used the data split provided by Karpathy and Fei-Fei (2015), with 6,000 images for training and a development and test set both of 1,000 images.

4.2.2 Image and acoustic features

To extract image features, all images are resized such that the smallest side is 256 pixels while keeping the aspect ratio intact. We take ten 224 by 224 crops of the image: one from each corner, one from the middle and the same five crops for the mirrored image. We use ResNet-152 (He et al., 2016) pretrained on ImageNet to extract visual features from these ten crops and then average the features of the ten crops into a single vector with 2,048 features.

We test two types of acoustic features; Mel Frequency Cepstral Coefficients (MFCCs) and Multilingual Bottleneck (MBN) features. The MFCCs were created using 40 Mel-spaced filterbanks. We use 12 MFCCs and the log energy feature and add the first and second derivatives resulting in 39-dimensional feature vectors. We compute the MFCCs using 25 ms analysis windows with a 5 ms shift.

The MBN features are created using a pretrained DNN made available by Fer et al. (2017). In short, the network is trained on multilingual speech data (11 languages, no English) to classify phoneme states. The MBN features consist of the outputs of intermediate network layers where the network is compressed from 1500 features to 30 features (see Fer et al. (2017) for the full details of the network and training).

4.2.3 Model architecture

Our multi-modal encoder maps images and their corresponding captions to a common embedding space. The idea is to make matching images and captions

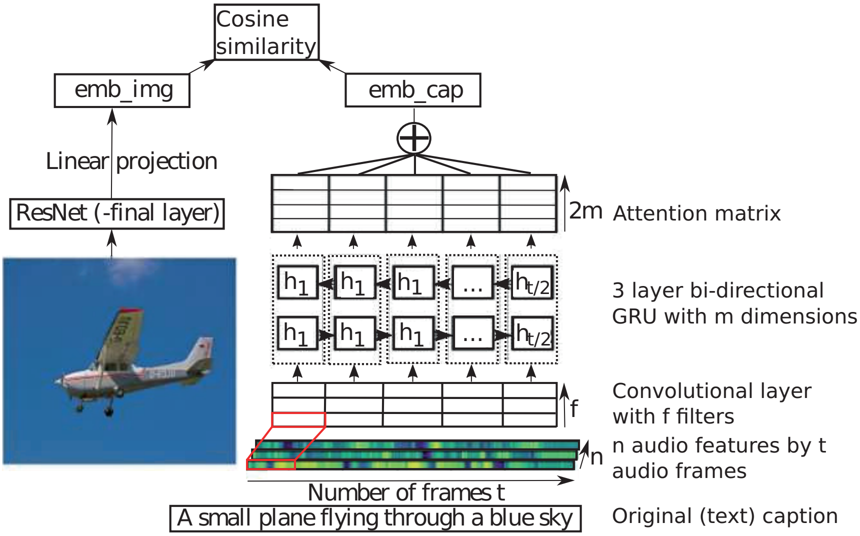
Cosine similarity
emb_img
emb_cap
Linear projection
ResNet (-final layer)
2m
Attention matrix
h_1
$h_{t/2}$
...
3 layer bi-directional GRU with m dimensions
f
Convolutional layer with f filters
n
n audio features by t audio frames
Number of frames t
A small plane flying through a blue sky
Original (text) caption

Figure 4.1: Model architecture: the model consists of two branches with the image encoder on the left and the caption encoder on the right. The audio features consist of n features by t frames and the GRU has $\mathbf{h}_{t/2}$ hidden states. Each GRU hidden state has m features which are concatenated for the forward and backward GRU into $2m$ dimensional hidden states. Vectorial attention is applied which weighs and sums the hidden states resulting in the caption embedding. At the top we calculate the cosine similarity between the image and caption embedding (**emb_img** and **emb_cap**).

lie close together and mismatched images and captions lie far apart in the embedding space. Our model consists of two parts; an image encoder and a sentence encoder as depicted in Figure 4.1. The approach is based on our own text-based model described in **Chapter 2** and on the speech-based models presented by Harwath et al. (2016) and Chrupała et al. (2017) and we refer to those studies for more details. Here, we focus on the differences with previous work.

For the image encoder we use a single-layer linear projection on top of the pretrained image recognition model, and normalise the result to have unit L2 norm. The image encoder has 2048 input units and 2048 output units.

Our caption encoder consists of three main components. First we apply a 1-dimensional convolutional layer to the acoustic input features. The convolution has a stride of size 2, kernel size 6 and 64 output channels. This is the only layer where the model differs from the text-based model, which features a character embedding layer instead of a convolutional layer. The resulting features are then fed into a bi-directional Gated Recurrent Unit (GRU) followed by a self-attention layer and is lastly normalised to have unit L2 norm.

We use a 3-layer bi-directional GRU which allows the network to capture long-range dependencies in the acoustic signal (see Chung et al. 2014 for a more detailed description of the GRU). Furthermore, by making the layer bi-directional we let the network process the output of the convolutional layer from left to right and vice versa, allowing the model to capture dependencies in both directions. We use a GRU with 1024 units, and concatenate the bidirectional representations resulting in hidden states of size 2048. Finally, the self-attention layer computes a weighted sum over all the hidden GRU states:

$$\mathbf{a}_t = \text{softmax}(V \tanh(W\mathbf{h}_t + \mathbf{b}_w) + \mathbf{b}_v) \tag{4.1}$$

$$\text{Att}(\mathbf{h}_1, ..., \mathbf{h}_t) = \sum_t \mathbf{a}_t \circ \mathbf{h}_t \tag{4.2}$$

where $\mathbf{a}_t$ is the attention vector for hidden state $\mathbf{h}_t$ and W, V, $\mathbf{b}_w$, and $\mathbf{b}_v$ indicate the weights and biases. The applied attention is then the sum over the Hadamard product between all hidden states $(\mathbf{h}_1, ..., \mathbf{h}_t)$ and their attention vector. We use 128 units for W and 2048 units for V.

4.2.4 Training

Following the approach in **Chapter 2**, the model is trained to embed the images and captions such that the cosine similarity between image and caption pairs is larger (by a certain margin) than the similarity between mismatching pairs. This so called hinge loss L as a function of the network parameters θ is given by:

$$\begin{aligned} L(\theta) = \sum_{(c,i),(c',i') \in B} \Big(& \max(0, \cos(c, i') - \cos(c, i) + \alpha) + \\ & \max(0, \cos(i, c') - \cos(i, c) + \alpha) \Big) \end{aligned} \tag{4.3}$$

where $(c, i) \neq (c', i')$. B is a minibatch of correct caption-image pairs (c, i), where the other caption-image pairs in the batch serve to create mismatched pairs (c, i') and (c', i). We take the cosine similarity $\cos(x, y)$ and subtract the similarity of the mismatched pairs from the matching pairs such that the loss is only zero when the matching pair is more similar than the mismatched pairs by a margin α. We use importance sampling to select the mismatched pairs; rather than using all the other samples in the mini-batch as mismatched pairs (as done in

Chapter 2 and Chrupała et al. 2017), we calculate the loss using only the hardest examples (i.e. mismatched pairs with high cosine similarity). While Faghri et al. (2017) used only the single hardest example in the batch for text-captions, we found that this did not work for the spoken captions. Instead we found that using the hardest 25 percent worked well.

The networks are trained using Adam (Kingma and Ba, 2015) with a cyclic learning rate schedule based on the work by Smith (2017). The learning rate schedule varies the learning rate smoothly between a minimum and maximum bound which were set to 10^{-6} and 2×10^{-4} respectively. The learning rate schedule causes the network to visit several local minima during training, allowing us to use snapshot ensembling (Huang et al., 2017). By saving the network parameters at each local minimum, we can ensemble the embeddings of multiple networks at no extra cost. We use a margin $\alpha = 0.2$ for the loss function. We train the networks for 32 epochs and take a snapshot for ensembling at every fourth epoch. For ensembling we use the two snapshots with the highest performance on the development data and simply sum their embeddings.

The main differences with the approaches described by Harwath et al. (2016) and Chrupała et al. (2017) are the use of multi-layered GRUs, importance sampling, the cyclic learning rate, snapshot ensembling and the use of vectorial rather than scalar attention.

4.3 Word presence detection

While our model is not explicitly trained to recognise words or segment the speech signal, previous work has shown that such information can be extracted by visual grounding models (Chrupała et al., 2017; Kamper et al., 2017a). Chrupała et al. (2017) use a binary decision task: given a word and a sentence embedding, decide if the word occurs in the sentence. Our approach is similar to the spoken-bag-of-words prediction task described by Kamper et al. (2017a). Given a sentence embedding created by our model, a classifier has to decide which of the words in its vocabulary occur in the sentence.

Based on the original written captions, our database contains 7,374 unique words with a combined occurrence frequency of 324,480. From these we select words that occur between 50 and a 1,000 times and are over 3 characters long so that there are enough examples in the data that the model might actually learn to recognise them, and to filter out punctuation, spelling mistakes, numerals and most function words. This leaves 460 unique words, mostly verbs and nouns, with a combined occurrence frequency of 87,020 in our data. We construct a vector for each sentence in Flickr8k indicating which of these words is present. We do not encode multiple occurrences of the same word in one sentence.

The words described above are used as targets for a neural network classifier consisting of a single feed forward layer with 460 units. This layer simply takes an embedding vector as input and maps it to the 460 target words. We then apply the standard logistic function and calculate the Binary Cross Entropy loss to train the network.

We train five word detection networks for both the MFCC and the MBN-based caption encoders, in order to see how word presence is encoded in the different neural network layers. We train networks for the final output layer, the three intermediate layers of the GRU and the acoustic features. For the final layer we simply use the output embedding as input to the word detection network. We apply some post-processing to the acoustic features and the intermediate layer outputs to ensure that our word detection inputs are all of the same size. As the intermediate GRU layers produce 2048 features for each time step in the signal, we use average-pooling along the temporal dimension to create a single input vector and normalise the result to have unit L2 norm. The acoustic features consist of 30 (MBN) or 39 (MFCC) features for each time step, so we apply the convolutional layer followed by an untrained GRU layer to the input features, use average-pooling and normalise the result to have unit L2 norm.

The word detection networks are trained for 32 epochs using Adam (Kingma and Ba, 2015) with a constant learning rate of 0.001. We use the same data split that was used for training the multi-modal encoder, so that we test word presence detection on data that was not seen by either the encoder or the decoder.

4.4 Results

Table 4.1 shows the performance of our models on the image-caption retrieval task. The caption embeddings are ranked by cosine distance to the image and vice versa where R@N is the percentage of test items for which the correct image or caption was in the top N results. We compare our models to Harwath and Glass (2015) and Chrupała et al. (2017), and include our own character-based model from **Chapter 2** for comparison. Harwath and Glass (2015) used a convolutional approach, whereas Chrupała et al. (2017) used recurrent highway networks with scalar attention. The character-based model is similar to the model we use here and was trained on the original Flickr8k text captions (see

Table 4.1: Image-Caption retrieval results on the Flickr8k test set. R@N is the percentage of items for which the correct image or caption was retrieved in the top N (higher is better). Med r is the median rank of the correct image or caption (lower is better). We also report the 95 percent confidence interval for the R@N scores.

Model	Caption to Image			
	R@1	R@5	R@10	med r
Harwath and Glass (2015)	-	-	17.9±1.1	-
Chrupała et al. (2017)	5.5±0.6	16.3±1.0	25.3±1.2	48
MFCC-GRU	8.4±0.8	25.7±1.2	37.6±1.3	21
MBN-GRU	12.7±0.9	34.9±1.3	48.5±1.4	11
Chapter 2 Char-GRU	27.5±1.2	58.2±1.4	70.5±1.3	4
Model	Image to Caption			
	R@1	R@5	R@10	med r
Harwath and Glass (2015)	-	-	24.3±2.7	-
MFCC-GRU	12.2±2.0	31.9±2.9	45.2±3.1	13
MBN-GRU	16.0±2.5	42.8±3.1	56.1±3.0	8
Chapter 2 Char-GRU	38.5±3.0	68.9±2.9	79.3±2.5	2

Table 4.2: Area under the curve of the receiver operating characteristic for both models.

Model	AUC				
	input	layer 1	layer 2	layer 3	attention
MBN	.57	.80	.86	.85	.82
MFCC	.54	.68	.80	.75	.75

Chapter 2 for a full description). Both our MFCC and MBN-based model significantly outperform previous spoken caption-to-image methods on the Flickr8k dataset. The largest improvement is the MBN model which outperforms the results reported in Chrupała et al. (2017) by as much as 23.2 percentage points on R@10. The MFCC model also improves on previous results but scores significantly lower than the MBN model across the board, improving as much as 12.3 percentage points over previous work. There is a large performance gap between the text-caption to image retrieval results and the spoken-caption to image results, showing there is still a lot of room for improvement.

The results of the word presence detection task are shown in Figure 4.2 and Table 4.2. Figure 4.2 shows the F1 score for all the classifiers at 20 equally spaced detection thresholds (i.e. a word is classified as 'present' if the word detection output is above this threshold). Table 4.2 displays the area under the curve for the receiver operating characteristic. Even though the MBN model outperforms the MFCC model for all layers we see the same pattern emerging from both the

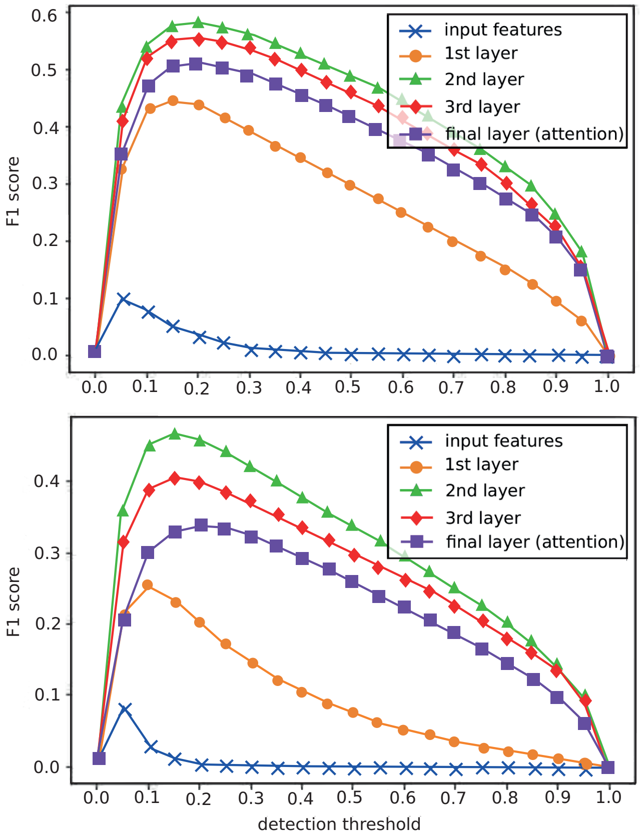

Figure 4.2: Plots of the F1 scores for the word presence classifiers at 20 equally spaced activation thresholds. The top figure shows the classifiers trained on the MBN model, and the bottom figure the MFCC model.

F1 score and the AUC. The performance on the feature level is not much better than random. Predicting ‘not present’ for every word would be the best random guess as this is a heavy majority class in this task. Inspection of the predictions shows that the classifier is indeed heavily biased towards the majority class for the input features. Then we see the performance increasing for the first layer and peaking at the second layer. The performance then drops slightly for the third layer and the attention layer.

4.5 Discussion and conclusion

We trained an image-caption retrieval model on spoken input and investigated whether it learns to recognise linguistic units in the input. As improvements over previous work we used a 3-layer GRU and employed importance sampling, cyclic learning rates, ensembling and vectorial self-attention. Our results on both MBN and MFCC features are significantly higher than the previous state-of-the-art. The largest improvement comes from using the learned MBN features but our approach also improves results for MFCCs, which are the same features as were used by Chrupała et al. (2017). The learned MBN features provide better performance whereas the MFCCs are more cognitively plausible input features.

The probing task shows that the model learns to recognise these words in the input. The system is not explicitly optimised to do so, but our results show that the lower layers learn to recognise this form related information from the input. After layer 2, the performance starts to decrease slightly which might indicate that these layers learn a more task-specific representation and it is to be expected that the final attention layer specialises in mapping from audio features to the multi-modal embedding space.

In conclusion, we presented what are, to the best of our knowledge, the best results on spoken-caption to image retrieval. Our results improve significantly over previous approaches for both untrained and trained audio features. In a probing task, we show that the model learns to recognise words in the input speech signal.

We are currently collecting the Semantic Textual Similarity (STS) database in spoken format and the next step will be to investigate whether the model presented here also learns to capture sentence level semantic information and understand language in a deeper sense than recognising word presence. The work presented by Chrupała et al. (2017) has made the first efforts in this regard and we aim to extend this to a larger database with sentences from multiple domains. Furthermore, we want to investigate the linguistic units that our model learns to recognise. In the current study, we only investigated whether the model learns to recognise words, but the potential benefit of our model is that it might learn multi-word statements or might even learn to look at sub-lexical level information. Harwath et al. (2020b) and Drexler and Glass (2017) have recently shown that the speech-to-image retrieval approach can be used to detect word boundaries and even discover sub-word units. Our interest is in investigating how these word and sub-word units develop over training and through the network layers.

5 Semantic sentence similarity: size does not always matter

This study addresses the question whether visually grounded speech recognition (VGS) models learn to capture sentence semantics without access to any prior linguistic knowledge. We produce synthetic and natural spoken versions of a well known semantic textual similarity database and show that our VGS model produces embeddings that correlate well with human semantic similarity judgements. Our results show that a model trained on a small image-caption database outperforms two models trained on much larger databases, indicating that database size is not all that matters. We also investigate the importance of having multiple captions per image and find that this is indeed helpful even if the total number of images is lower, suggesting that paraphrasing is a valuable learning signal. While the general trend in the field is to create ever larger datasets to train models on, our findings indicate other characteristics of the database can be just as important.

5.1 Introduction

The idea that words that occur in similar contexts have similar meaning has been investigated for decades (e.g., Rubenstein and Goodenough 1965; Deerwester et al. 1990). Advances in deep learning and computational power have made it possible to create models that learn useful and meaningful representations of larger spans of text such as sentences, paragraphs and even complete documents (Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017; Yang et al., 2018; Kiela et al., 2018; Devlin et al., 2019). A caveat of such models is the need to be trained on enormous amounts of text and the current trend is to use ever larger training corpora to create better models. Whereas BERT (Devlin et al., 2019) is trained on 2.5 billion tokens of text the more recent GPT-3 (Brown et al., 2020) is trained on nearly 500 billion tokens. It is obvious that humans are able to

This chapter is based on: Danny Merkx, Stefan L. Frank and Mirjam Ernestus. Semantic sentence similarity: size does not always matter. In *Interspeech 2021 - 22nd Annual Conference of the International Speech Communication Association*, pages 4393-4397, 2021 understand and use language after much less exposure; one would need to read 200 words per second, 24 hours a day for 80 years to digest as much information as GPT-3. People are able to hear and speak long before they are able to read, and many people never learn to read at all. Moreover, writing is a relatively recent invention, which only arose after spoken language.

Visually Grounded Speech (VGS) models aim to learn language without using written text data or prior information on the linguistic units in the speech signal. Instead, these models combine speech signals with visual information to guide learning; a VGS model learns to create representations for an image and its corresponding spoken caption that are similar to each other in the embedding space. Such models have been shown to learn to extract meaningful linguistic units from speech without explicitly being told what these units are, as shown in word recognition experiments (e.g., **Chapter 4** and Havard et al. 2019) and semantic keyword spotting (Kamper and Roth, 2018). Recent research has shown that VGS models with quantisation layers learn to extract phonetic and word-like units that are useful in zero-shot learning and speech synthesis (Harwath et al., 2020a; Hsu et al., 2020).

As with text-based models, there is a trend in VGS models to use ever larger training corpora. CELL, one of the earliest VGS models, used a database of around 8,000 utterances (Roy and Pentland, 1998). Harwath and colleagues introduced the first 'modern' neural network based approach which was trained on the Flickr8k Audio Caption corpus, a corpus with 40,000 utterances (Harwath and Glass, 2015). This corpus was quickly followed up by Places Audio Captions (400,000 utterances) (Harwath et al., 2020b) and, most recently, by SpokenCOCO (600,000 utterances) (Hsu et al., 2020).

However, previous work on visual grounding using written captions has shown that larger databases do not always result in better models. In **Chapter 2**, we compared models trained on the written captions of Flickr8k (Hodosh et al., 2013) and MSCOCO (Chen et al., 2015). We showed that, although the much larger MSCOCO (600k sentences) achieved better performance on the training task, the model trained on the smaller Flickr database performed better at transfer tasks; the resulting embeddings correlated better with human semantic relatedness ratings. As the MSCOCO model only performed better on visually descriptive sentences, these results suggest that there is a trade-off between getting better at processing image descriptions and creating generally useful sentence representations.

There is another interesting difference between the VGS training corpora besides their size. While both Flickr8k Audio and SpokenCOCO have five captions per image, Places Audio has only one. Consequently, even though SpokenCOCO has more captions than Places, Places has 400,000 images while SpokenCOCO has only 120,000. The more fundamental difference is how models trained on Places and Flickr8k handle paraphrases. In a VGS, captions with similar images (i.e., likely paraphrases) should have similar representations. So, in a way, a VGS can be said to implicitly learn that paraphrases share one meaning. However, the paraphrasing in SpokenCOCO and Flickr8k is more explicit than in Places because there are always five captions per image, and these should ideally have the same representation in the embedding space.

Our first research question is: do VGS models learn to capture sentence semantics? So far, testing of the usability of the sentence representations created by VGS models has been limited, and recent research has focused more on whether useful sub-sentence units can be extracted (e.g., Havard et al. 2019; Harwath et al. 2020a; Hsu et al. 2020). To answer this question we will investigate whether the representations learned by a VGS are predictive of semantic sentence similarity as judged by humans. In order to test this, we create spoken versions of the Semantic Textual Similarity (STS) database. STS consists of sentence pairs that were annotated by humans for semantic similarity. We look at the correlation between the human similarity ratings and the similarity of sentence representations created by our VGS model.

We compare models trained on the three spoken image caption databases; Flickr8k Audio Captions, Places Audio Captions and SpokenCOCO. It is tempting to simply move on to the bigger corpus once one becomes available without investigating whether this actually constitutes an improvement. Using more data will likely lead to an increase in training task performance, but comparisons between corpora based on metrics other than training task performance are scarce. We investigate which model creates sentence representations that best capture sentence semantics, the only difference between these models being the database they were trained on. Our test material (STS) contains sentences from a wide range of domains, so a model needs to be able to generalise well to perform well on this task.

We will also investigate the importance of paraphrasing in corpora having multiple captions for each image. Our second research question is: is it beneficial for VGS models to have multiple captions per image? We answer this question by training models on subsets of SpokenCOCO where we fix the total number of captions, but vary the number of captions per image and consequently the number of images in the training data.

5.2 Methods

5.2.1 Semantic similarity data

For the semantic evaluation we use the Semantic Textual Similarity (STS) data. STS is a shared task hosted at the SemEval workshop. These datasets contain paired sentences from various sources labelled by humans with a similarity score between zero ('the two sentences are completely dissimilar') and five ('the two sentences are completely equivalent, as they mean the same thing') averaged over five annotators per sentence pair (see Agirre et al. 2015 for a full description of the annotator instructions).

We use the STS 2012 to 2016 tasks, which are included in the SentEval toolbox for testing textual sentence representations (Conneau and Kiela, 2018), allowing for a comparison between speech-based models and previous work using SentEval. Table 5.1 gives an overview of the STS tasks by year, and the sources from which the sentences were taken. We had the sentences produced by speech production software (synthetic speech) and by humans. All synthetic and natural utterances are made publicly available in .wav format as the SpokenSTS database, which can be found at: `https://doi.org/10.17026/dans-z48-3ev6`.

Synthetic speech

The synthetic speech was created with Google's Wavenet using three male and three female voices with a US accent. All utterances were produced using all six voices for a total of 75,264 utterance pairs. We applied as little preprocessing to the STS text as possible. To identify the necessary preprocessing steps, we sampled 10% of the STS sentence pairs to convert to synthetic speech without any preprocessing. This sample was used to identify text characteristics that were troublesome to Wavenet and to apply the necessary preprocessing steps in order to correct these where possible. For example, Wavenet pronounces the quotation marks (saying /quote/) if there is a space between the period and a quotation mark at the end of a sentence. Wavenet also pronounces certain non-capitalised abbreviations as if they were words rather than spelling them out (e.g., 'usa' is pronounced /usa/ instead of /u/, /s/, /a/). A full overview

Table 5.1: Description of the STS subtasks by year. Some subtasks appear in multiple years, but consist of different sentence pairs drawn from the same source. The image description datasets are drawn from the PASCAL VOC-2008 dataset (Everingham et al., 2008) and do not overlap with the training material of our models.

Task	Subtask	#Pairs	Source
STS 2012	MSRpar	750	newswire
	MSRvid	750	videos
	SMTeuroparl	459	glosses
	OnWN	750	WMT eval.
	SMTnews	399	WMT eval.
STS 2013	FNWN	189	newswire
	HDL	750	glosses
	OnWN	561	glosses
STS 2014	Deft-forum	450	forum posts
	Deft-news	300	news summary
	HDL	750	newswire headlines
	Images	750	image descriptions
	OnWN	750	glosses
	Tweet-news	750	tweet-news pairs
STS 2015	Answers forum	375	Q&A forum answers
	Answers students	750	student answers
	Belief	375	committed belief
	HDL	750	newswire headlines
	Images	750	image descriptions
STS 2016	Answer-Answer	254	Q&A forum answers
	HDL	249	newswire headlines
	Plagiarism	230	short-answer plagiarism
	Postediting	244	MT postedits
	Question-Question	209	Q&A forum questions
Total		12,544	

of all preprocessing applied, our code and our data can be found at `https://github.com/DannyMerkx/speech2image/tree/Interspeech21`.

Natural speech

We selected a random sample of 5% of the STS sentence pairs (638 pairs) evenly distributed across the different STS subsets. These sentences were recorded by four native speakers of English (two male, two female) with a North American accent. Recordings were made in a sound-attenuated booth using Audacity in sessions of one and a half hour including breaks. Speakers read the sentences out loud from a script. They were instructed to pronounce the sentences as they found most appropriate (e.g., saying ‘an apple’ even though the original STS sentence might be misspelled as ‘a apple’) and to pronounce large numbers according to their preference either in full or digit by digit. Speakers were paid 10 euros per hour in gift certificates.

After recording was done, the audio was processed by an annotator. Utterances were automatically detected and labelled in Audacity, checked by the annotator for deviations from the script and where possible these deviations were corrected. For instance, when speakers made a mistake, they were allowed to continue from a natural break like a comma and so the annotator combined the correct parts from multiple attempts. If speakers misspoke and corrected themselves mid-utterance without re-recording (part of) the utterance, the mistake was removed. Furthermore, silences longer than 500ms were shortened.

5.2.2 Visually Grounded Speech model

The VGS architecture used in this study is our own implementation presented in **Chapter 4** and we refer to that paper for more details. Here, we present a description of the model and the differences with **Chapter 4**.

Our VGS model maps images and their corresponding captions to a common embedding space. It is trained to make matching images and captions lie close together, and mismatched images and captions lie far apart, in the embedding space. The model consists of two parts; an image encoder and a caption encoder. The image encoder is a single-layer linear projection on top of ResNet-152 (He et al., 2016), a pretrained image recognition network, with the classification layer removed. We train only the linear projection and do not further fine-tune ResNet.

The caption encoder consists of a 1-dimensional convolutional layer followed by a bi-directional recurrent layer and finally a self-attention layer. The only difference with **Chapter 4** is the use of a four-layer LSTM instead of a three-layer GRU. Audio features consist of 13 Cepstral mean-variance normalised MFCCs and their first and second order derivatives calculated for 25ms frames with 10ms frame-shift.

5.2.3 Training material

We train separate models on each of the three training corpora. Flickr8k (Hodosh et al., 2013) has 8,000 images and 40,000 written captions, five per image. We use the spoken versions of these captions collected using Amazon Mechanical

Turk (AMT) by Harwath and Glass (2015). The data split is provided by Karpathy and Fei-Fei (2015), with 6,000 images for training and a development and test set both of 1,000 images.

Places has 400,000 images drawn from the Places205 corpus (Zhou et al., 2014) for which a single audio description per image was collected by Harwath et al. (2020b) using AMT. Whereas Flickr8k Audio consisted of written captions which were then read out loud by workers, here, workers were tasked with describing the Places images as no written captions existed. We use the most recent official split[1] with 400,000 images for training and a development and test set of 1,000 images.

MSCOCO has 123,287 images and 605,495 written captions (Chen et al., 2015), for which Hsu et al. (2020) collected spoken versions using AMT which they released as SpokenCOCO. Five thousand images are reserved as a development set and no official test set is provided. In order to keep results comparable between models we use 1,000 images from the development set for development and reserve 1,000 images as a test set.

5.2.4 Experiments

All models reported in this study are trained for 32 epochs. The models are trained using a cyclical learning rate which smoothly varies the learning rate between 2×10^{-4} and 2×10^{-6} over the course of four epochs. After a model is trained, we select the epoch with the lowest development set error for further testing.

To answer our first research question, we use the trained caption encoders to encode the SpokenSTS sentences. We calculate the cosine similarity between each pair of encoded sentences and then calculate the Pearson correlation coefficient between the embedding similarity scores and the human similarity judgements.

To answer our second research question, we train five more models on subsets of SpokenCOCO where we vary the number of images in the training set and the number of captions per image. As a lower bound on the amount of data we take the size of Flickr8k; 6,000 images and 30,000 captions, five per image. We then increase the amount of visual information (i.e., number of images) while keeping the total number of captions fixed at 30,000; 7,500 images with four captions per image, 10,000 images with three captions per image, 15,000 images with

[1] Available at: https://groups.csail.mit.edu/sls/downloads/placesaudio/downloads.cgi

Table 5.2: Image-Caption retrieval results of each database's respective test set. R@N is the percentage of items for which the correct image or caption was retrieved in the top N (higher is better). Med r is the median rank of the correct image or caption (lower is better).

Model	Caption to Image			
	R@1	R@5	R@10	med r
Flickr8k Audio	12.7	35.1	48.4	12
Places Audio	30.6	62.6	73.8	3
SpokenCOCO	30.6	64.1	79.8	3
	Image to Caption			
Flickr8k Audio	20.3	44.8	58.8	7
Places Audio	29.5	62.0	74.3	3
SpokenCOCO	39.2	75.3	86.4	2

two captions per image and finally 30,000 images with one caption per image, similar to the Places database. If paraphrasing is helpful to the model, we expect model performance to decrease with a decreasing number of captions per image, even though the total number of captions remains the same. While we obviously cannot make sure that the models are trained on the same data, the data in the model with five captions per image is a subset of the data for the model with four captions per images and so on, so that the training data for each model overlaps as much as possible given the experimental setup. All code used in this study is available at `https://github.com/DannyMerkx/speech2image/tree/Interspeech21`

5.3 Results

In Table 5.2 we compare the image-caption retrieval performance of the three models trained on different datasets (Flickr8k Audio, Places Audio and SpokenCOCO). This indicates how well the models perform on the training task. In order to retrieve images using a caption or captions using an image, the caption embeddings are ranked by their similarity to the image embeddings, and vice versa. It is clear that training task performance increases with database size.

The results of the sentence semantics evaluation are shown in Figure 5.1. We show Pearson correlation coefficients between the human similarity judgements and the embedding similarities generated by the trained models. As each sentence is pronounced by six voices we calculate the embedding similarity for each pair of voices and average over the resulting 36 pairs. In general, we see that both the Flickr8k and the SpokenCOCO model tend to outperform the

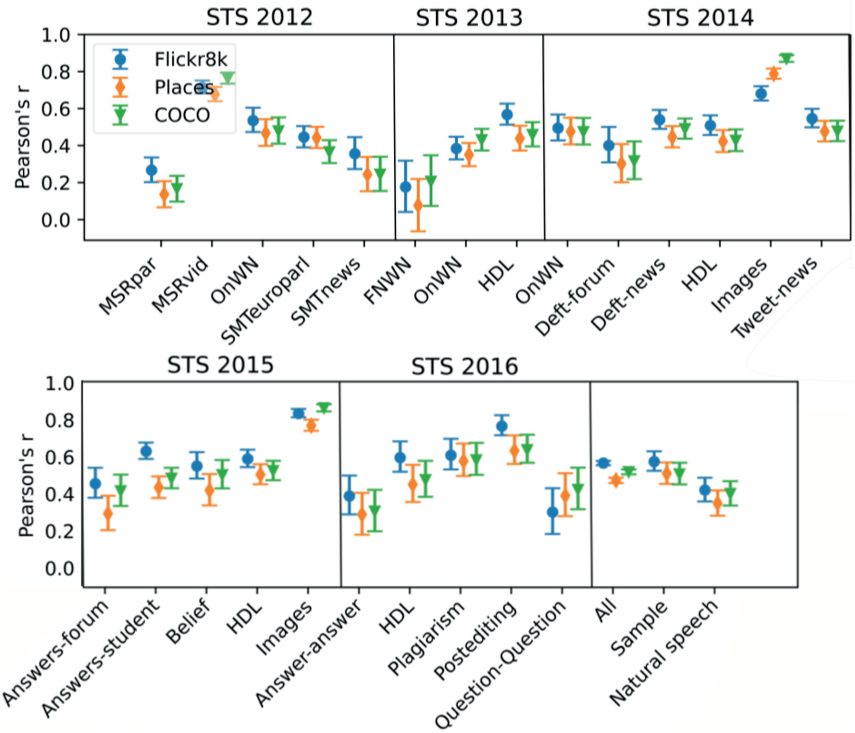

Figure 5.1: Semantic evaluation task results: Pearson correlation coefficients with their 95 percent confidence interval for the various subtasks using the synthetic SpokenSTS(see Table 5.1). The bottom rightmost section shows the average over all STS subsets (All), the results on the natural speech recordings (Natural speech) and the results on the synthetic version of the natural speech sample (Sample).

Places model, and that the Flickr8k model tends to outperform the SpokenCOCO model. This is confirmed by the significant differences in Pearson's *r* calculated on the complete SpokenSTS database (indicated as All).

Lastly, it is clear that all models perform worse on natural speech. In Figure 5.1, Sample indicates the subset of synthetic speech representing the same sample of STS sentences that was used for the natural speech. Model performance on this subset is similar to the performance on the entirety of SpokenSTS indicating that the sample is representative of the entire corpus.

The results of the paraphrasing experiment are shown in Figure 5.2. The results show a trend where models trained on more captions per image (i.e., more paraphrases) perform better, even though the total number of captions is the same across models and the models with more paraphrases receive less visual

Table 5.3: AIC comparison of regression models (lower is better). ΔAIC indicates the difference in AIC compared to the best model, LL indicates the model's log likelihood

No. captions	AIC	ΔAIC	LL
5	127974.5	0.00	−63984.23
3	127985.3	10.81	−63989.64
4	128116.3	141.80	−64055.13
2	128218.3	243.80	−64106.13
1	128269.7	295.26	−64131.86

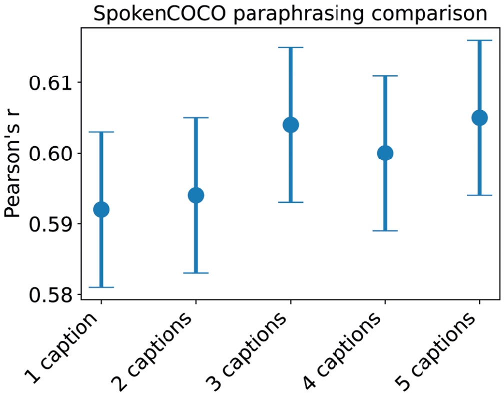

Figure 5.2: Comparison of the five models trained on subsets of SpokenCOCO with differing numbers of captions per image. We show Pearson correlation coefficients over the entire synthetic SpokenSTS with 95 percent confidence intervals.

information. To further investigate this trend we performed five separate regression analyses with the human similarity judgements as dependent variable and each of the five models' similarity ratings as regressors. Embedding similarities were averaged over the 36 voice pairs. Table 5.3 shows a comparison of the Akaike Information Criteria (AIC) of these regression models. These results show the same trend as Figure 5.2 and clearly indicate that the similarity ratings generated by models with more captions per image provide a better fit to the human similarity ratings.

5.4 Discussion and conclusion

We collected synthetic and natural speech for a large corpus of human sentence similarity judgements in order to investigate whether VGS models learn to capture sentence semantics. Furthermore, we investigated the merits of database size and the availability of paraphrases in the training data.

The results show that similarity scores generated by our VGS models correlate quite well human similarity judgements overall. This shows that a model tasked with mapping images to captions and vice versa learns to capture sentence semantics. However there are also some subsets of STS (MSRpar, FNWN) on which the model performs quite poorly, and unsurprisingly all models clearly perform best on subtasks consisting of visual descriptions. Furthermore, we found that even though the models trained on Places and SpokenCOCO outperform the Flickr8k model in terms of training task performance, the Flickr8k model performs better on the SpokenSTS task. This confirms our previous results on text-based grounding models in **Chapter 2** which compared models trained on the written versions of Flickr8k and MSCOCO. As in **Chapter 2** we see that SpokenCOCO outperforms Flickr8k mainly on the subtasks containing visual descriptions. The models that were trained on a smaller subsets of SpokenCOCO for the paraphrasing experiment performed better than the model trained on the entire database. This indicates that training on more data might cause the model to overspecialise; it performs better on sentences which are similar to the training data, but becomes worse at generalising to sentences from other domains.

Next, we investigated whether the presence of paraphrases in the data (i.e., multiple captions per image) is beneficial to the model. By training models on subsets of SpokenCOCO where we fixed the total number of captions but varied the number of captions per image, we found that having more captions per image increases model performance, even though these models consequently are trained on less visual information. This also explains why the Places model performs worst out of the three, even though the amount of data is in the same ballpark as SpokenCOCO (it has fewer captions but more images). An interesting question for future research is whether this trend continues beyond five captions per image. Collecting more captions for existing databases, rather than collecting more image captions pairs, could be an important consideration for future data collection efforts.

In conclusion, we show VGS models are capable of capturing sentence semantics. Importantly, our results show that database size is not all that matters when it comes to training VGS models. Even though it is enticing to collect ever larger databases to increase training task performance, this does not always translate to better transfer learning results. Our Flickr8k model outperforms our SpokenCOCO model even though it has 20 times less data. Furthermore, other characteristics of a database might be even more important than its size; in the case of VGS this is the presence of multiple captions per image.

6 | Modelling human word learning and recognition using visually grounded speech

Many computational models of speech recognition assume that the set of target words is already given. This implies that these models learn to recognise speech in a biologically unrealistic manner, i.e., with prior lexical knowledge and explicit supervision. In contrast, visually grounded speech models learn to recognise speech without prior lexical knowledge by exploiting statistical dependencies between spoken and visual input. While it has previously been shown that visually grounded speech models learn to recognise the presence of words in the input, we explicitly investigate such a model as a model of human speech recognition. We investigate the time-course of noun and verb recognition as simulated by the model using a gating paradigm to test whether its recognition is affected by well-known word-competition effects in human speech processing. We furthermore investigate whether vector quantisation, a technique for discrete representation learning, aids the model in the discovery and recognition of words. Our experiments show that the model is able to recognise nouns in isolation and even learns to properly differentiate between plural and singular nouns. We also find that recognition is influenced by word competition from the word-initial cohort and neighbourhood density, mirroring word competition effects in human speech comprehension. Lastly, we find no evidence that vector quantisation is helpful in discovering and recognising words, though our gating experiment does show that the LSTM-VQ model is able to recognise the target words earlier.

6.1 Introduction

Infants initially have little understanding of what is being said around them, and yet at approximately nine months old are able to produce their first words. When they start producing their first multi-word utterances around 18 months, they

This chapter is based on: Danny Merkx, Sebastiaan Scholten, Stefan L. Frank, Mirjam Ernestus and Odette Scharenborg. Modelling human word learning and recognition using visually grounded speech. *Cognitive Computation*, forthcoming.

can already produce about 45 words and comprehend many more (Benedict, 1979; Snyder et al., 1981). One of the challenges infants face is that speech does not contain neat breaks between words, which would allow them to segment the utterance into words. To complicate things further, words might be embedded in longer words (e.g., *ham* in *hamster*) and furthermore, no two realisations of the same spoken word are ever the same due to speaker differences, accents, co-articulation and speaking rate, etc. (Eisner and McQueen, 2018). In this study, we investigate whether a computational model of speech recognition inspired by infant learning processes can learn to recognise words without prior linguistic knowledge.

Cognitive science has long tried to explain our capacity for speech comprehension through computational models (see Weber and Scharenborg 2012 for an overview). Models such as Trace (Elman and McClelland, 1988), Cohort (Marslen-Wilson, 1987), Shortlist (Norris, 1994), Shortlist B (Norris and McQueen, 2008) and FineTracker (Scharenborg, 2010) attempt to explain how variable and continuous acoustic signals are mapped onto a discrete and limited-size mental lexicon. These models all assume that the speech signal is first mapped to a set of pre-lexical units (e.g., phones, articulatory features) and then to a set of lexical units (words). The exact set of units is predetermined by the model developer, avoiding the issue of learning what these units are in the first place. Even the recently introduced DIANA model (ten Bosch et al., 2015), which does away with fixed pre-lexical units, uses a set of predetermined lexical units.

While all these models have proven successful at explaining behavioural data from listening experiments, they all require prior lexical knowledge in the form of a fully specified set of (pre-)lexical units. In contrast, infants learn words without prior lexical knowledge (or, arguably, any other linguistic knowledge) as well as without explicit supervision. A viable computational model should simulate word learning in a similar manner.

We take inspiration from the way infants learn language in order to model human word learning and recognition in a more cognitively plausible and ‘human like’ manner. While learning language, children are exposed to a wide range of sensory experiences beyond purely linguistic input. On the other hand, current computational models of word learning and recognition are often limited to linguistic input. Using a multi-modal model, we aim to show that it is possible to learn to recognise words without prior lexical knowledge and explicit supervision if the model is exposed to sensory experiences beyond speech. While there are many sensory experiences that could contribute to language learning, we focus on the most prominent of the human senses: vision. The model that we investigate in the current work exploits visual context in order to learn to recognise words in speech without supervision or prior lexical knowledge.

6.1.1 Visually Grounded Speech

Humans have access to multiple streams of sensory information besides the speech signal, perhaps most prominently the visual stream. It has been suggested that infants learn to extract words from speech by repeatedly hearing words while seeing the associated objects or actions (Räsänen and Rasilo, 2015), and indeed speech is often used to refer to and describe the world around us. For instance, parents might say ‘the ball is on the table’ and ‘there’s a ball on the floor’ etc., while consistently pointing towards a ball.

Visually Grounded Speech (VGS) models are speech recognition models inspired by this learning process. The idea behind VGS models (e.g. De Deyne et al. 2021; Harwath et al. 2020a; Kamper et al. 2019) is to make use of co-occurrences between the visual and auditory streams. For instance, from the sentences ‘a dog playing with a stick’ and ‘a dog running through a field’ along with images of these scenes, a model could learn to link the auditory signal for ‘dog’ to the visual representation of a dog because they are common to both image-sentence pairs. This allows the model to discover words, that is, to learn which utterance constituents are meaningful linguistic units. While there is a wide variety of VGS models, they all share the common concept of combining visual and auditory information in a common multi-modal representational space in which the similarity between matching image-sentence pairs is maximised while the similarity between mismatched pairs is minimised.

The potential of visual input for modelling the learning of linguistic units has long been recognised. In 1998, Roy and Pentland introduced their model of early word learning (Roy and Pentland, 1998). While many models at the time (and even today) relied on phonetic transcripts or written words, they implemented a model that learns solely from co-occurrences between the visual and auditory inputs. This model builds an ‘audio-visual lexicon’ by finding clusters in the visual input and looking for reoccurring units in the acoustic signal. It performs many tasks that are still the focus of research today: unsupervised discovery of linguistic units, retrieval of relevant images, and generation of relevant utterances. However, the model was limited to colours and shapes (utterances such as ‘this is a blue ball’) and has not been shown to learn from more natural, less restricted input.

The tasks performed by Roy and Pentland's model involve challenges for both computer vision and natural language processing. Advances in both fields have led to renewed interest in multi-modal learning, and with it increased the need for multi-modal datasets. In 2013, Hodosh, Young and Hockenmaier introduced Flickr8k (Hodosh et al., 2013), a database of images accompanied by written captions describing their contents, which was quickly followed by similar databases such as MSCOCO Captions (Chen et al., 2015). These datasets are now widely used for image-caption retrieval models (e.g., Karpathy and Fei-Fei 2015; Klein et al. 2015; Ma et al. 2015; Vendrov et al. 2016; Wehrmann et al. 2018; Dong et al. 2018 and **Chapter 2**) and caption generation (e.g., Karpathy and Fei-Fei 2015; Xu et al. 2015).

Harwath and Glass collected spoken captions for the Flickr8k database and used it to train the first neural network based VGS model (Harwath and Glass, 2015). Since then, there have been many improvements to the model architecture (Harwath et al. 2016; Chrupała et al. 2017; Havard et al. 2020; Harwath et al. 2020b; Scharenborg et al. 2020; Kamper and Roth 2018 and **Chapter 4**), as well as new applications of VGS models such as semantic keyword spotting (Kamper et al. 2017b,a, 2019), image generation (Wang et al., 2021), recovering of masked speech (Srinivasan et al., 2020), and even the combination of speech and video (Palaskar et al., 2018).

Many studies have since investigated the properties of the representations learned by such VGS models (e.g., Harwath et al. 2020a; Chrupała et al. 2018; Hsu et al. 2020; Chrupała et al. 2020 and **Chapter 5**). Perhaps the most prominent question is whether words are encoded in these utterance embeddings even though VGS models are not explicitly trained to encode words and are only exposed to complete sentences. The VGS model presented by Harwath et al. (2020b) showed that representations of a speech unit and a visual patch are often most similar when the visual patch contains the speech unit's visual referent. Chrupała et al. (2017) and our results in **Chapter 4**, show that VGS models encode the presence of individual words that can reliably be detected in the resulting sentence representation.

Räsänen and Khorrami (2019) made a VGS model that was able to discover words from even more naturalistic input than image captions: recordings from head-mounted cameras worn by infants during child-parent interaction. The authors showed that their model was able to learn utterance representations in which several words (e.g., 'doggy', 'ball') could reliably be detected. Even though their model used visual labels indicating the objects the infants were paying attention to rather than the actual video input, this study is an important step towards showing that VGS models can acquire linguistic units from actual child-directed speech.

While the presence of individual words is encoded in the representations of a VGS model, the model does not explicitly yield any segmentation or discrete linguistic units. A technique which allows for the unsupervised acquisition of such discrete units is Vector Quantisation (VQ). VQ layers were recently popularised by van den Oord et al. (2017), who showed that these layers could efficiently learn a discrete latent representational space. Harwath et al. (2020a) have recently applied these layers in a VGS model, and showed that their model learned to encode phones and words in its VQ layers.

Havard and colleagues went beyond simply detecting the presence of words in sentence representations: they presented isolated nouns to a VGS model trained on whole utterances, and showed that the model was able to retrieve images of the nouns' visual referents (Havard et al., 2019). This shows that their model does not merely encode the presence of these nouns in the sentence representations, but actually 'recognises' individual words and learns to map them onto their visual referents. So, regarding the example mentioned above, the model learned to link the auditory signal for 'dog' to the visual representation of a dog.

However, the model by Havard et al. (2019) was trained on synthetic speech. Word recognition in natural speech is known to be more challenging, as shown for instance by a large performance gap between VGS models trained on synthetic and real speech (Chrupała et al., 2017). Dealing with the variability of speech is an important aspect of human speech recognition. If VGS models are to be plausible as computational models of speech recognition, it is important that these models implicitly learn to extract words from natural speech.

6.1.2 Current study

The goal of this study is to investigate whether a VGS model discovers and recognises words from natural, as opposed to synthetic, speech. We furthermore go beyond earlier work because we investigate the model's cognitive plausibility by testing whether its word recognition performance is affected by word competition known to take place during human speech comprehension. We aim to answer the following questions:

1. Does a VGS model trained on natural speech learn to recognise words, and does this generalise to isolated words?

2. Is the model's word recognition process affected by word competition?
3. Does the model learn the difference between singular and plural nouns?
4. Does the introduction of VQ layers for learning discrete linguistic units aid word recognition?

Our **first** experiment is a continuation of our previous work (Scholten et al., 2021) and the work by Havard et al. (2019). Following Havard et al. (2019), we present isolated target words to the VGS model and measure its word recognition performance by looking at the proportion of retrieved images containing the target word's visual referent. If the model is indeed able to recognise a word in isolation, it should be able to retrieve images depicting the word's visual referent, indicating that the model has learned a representation of the word from the multi-modal input. Whereas previous work focused on the recognition of nouns, we also include verbs as our target words.

For this experiment, we collect new speech data, consisting of words pronounced in isolation. On the one hand, such data can be thought of as 'cleaner' than words extracted from sentences (as in Scholten et al. 2021) due to the absence of co-articulation. On the other hand, the model was trained on words in their sentence context, co-articulation included, and might have learned to rely on this contextual information too heavily to also recognise words in isolation. Thus, to answer our first research question, we investigate whether our VGS model learns to recognise words independently of their context. Furthermore, we investigate whether linguistic and acoustic factors affect the model's recognition performance similarly to human performance. For instance, we know that faster speaking negatively impacts human word recognition (e.g. Koch and Janse 2016).

In our **second** experiment we investigate the time course of word recognition in our VGS model. This allows us to test whether the model's word recognition performance is affected by word competition as is known to take place during human speech comprehension. For this experiment, we look at two measures of word competition: word-initial cohort size and neighbourhood density. In the Cohort model of human speech recognition (Marslen-Wilson, 1987), the incoming speech signal is mapped onto phone representations. These activated phone representations activate every word in which they appear. As more speech information becomes available, activation reduces for words that no longer match the input. The word that best matches the speech input is recognised. The number of activated or competing words is called the word-initial cohort size and plays a role in human speech processing: the larger the cohort size (i.e., the more competitors there are), the longer it takes to recognise a word (Norris et al., 1995). Words with a denser neighbourhood of similar sounding words are also harder to recognise as they compete with more words (Luce and Pisoni, 1998).

We also use our model to test the interaction between neighbourhood density and word frequency. Several studies have investigated this interaction, with inconclusive results. In a gating study, Metsala (1997) found an interaction where recognition was facilitated by a dense neighbourhood for low-frequency words and by a sparse neighbourhood for high-frequency words. Goh et al. (2009) found that response latencies in word recognition were shorter for words with sparser neighbourhoods. They furthermore found a higher recognition accuracy for sparse-neighbourhood high-frequency words as opposed to the other conditions (i.e., sparse-low, dense-high, dense-low). This means that, unlike Metsala, they found no facilitatory effect of neighbourhood density for low-frequency words. Others found no interaction between lexical frequency and neighbourhood density at all (Rispens et al., 2015; Garlock et al., 2001).

For this experiment, we use a gating paradigm, a well known technique borrowed from human speech processing research (e.g., Cotton and Grosjean 1984; Smith 2017). In the gating experiment, a word is presented to the VGS model in speech segments of increasing duration, that is, with an increasing number of phones, and the model is asked to retrieve an image of the correct visual referent on the basis of the speech signal available so far. We then analyse the effects of word competition and several control factors on word recognition performance.

In our **third** experiment we investigate whether our VGS model learns to differentiate between singular and plural instances of nouns. By the same principle of co-occurrences between the visual and auditory streams that allows the model to discover and recognise nouns, it may also be able to differentiate between their singular and plural forms. We test this by presenting both forms of all nouns to the model, and analysing whether the retrieved images contain single or multiple visual referents of that noun.

Our **fourth** question investigates VQ, a technique that was recently first applied to VGS models by Harwath et al. (2020a). Their model acquired discrete linguistic units, including words. However, it is still an unanswered question whether such VQ-induced word units also aid the recognition of words in isolation. If they do, the addition of VQ layers should improve word recognition results of our VGS model. Havard et al. (2020) improved retrieval performance of their VGS model by providing explicit word boundary information, thereby showing that knowledge of the linguistic units is indeed beneficial to the model. Rather than explicitly providing word-boundary information, VQ layers allow units to emerge in an end-to-end fashion. Because prior knowledge of word boundaries is not cognitively plausible, VQ layers are a more suitable approach for our cognitive model. To investigate if the introduction of VQ layers indeed aids word recognition, all our experiments compare the baseline VGS model to a VGS model with added VQ layers.

To foreshadow our results, we find that (1) our VGS model does learn to recognise words in isolation but performance is much higher on nouns than on verbs; (2) word recognition in the model is affected by competition similarly to humans; (3) the model can distinguish between singular and plural nouns to a limited extent; and (4) the use of VQ layers does not improve the model's recognition performance.

6.2 Methods

6.2.1 Visually Grounded Speech model

Model architecture

Our VGS model consists of two deep neural networks as depicted in Figure 6.1; one to encode the images and one to encode the audio captions. The model is trained to embed both input streams in a common embedding space; its training goal is to minimise the cosine distance between image-caption pairs while maximising the distance between mismatched pairs. We do not fine-tune the hyper-parameters of the model but use the best parameters found in **Chapter 4** – this is because it is not our current goal to improve the training task score but to perform experiments in order to learn more about the unsupervised discovery and recognition of words in a VGS model.

It is common practice to use a pretrained image recognition network for the image branch of a VGS model (e.g., Kamper et al. 2017a; Chrupała et al. 2017; Harwath et al. 2020a). We use the ResNet-152 network (He et al., 2016), which is a pretrained convolutional network that was trained on ImageNet (Deng et al., 2009), to extract image features. This is done by taking the activations of ResNet-152's penultimate fully connected layer by removing the final object-classification layer. Our image branch then is a single linear layer of size 2048 applied to these image features. Finally, we normalise the results to have unit L2 norm. The goal of the linear projection is to map the image features to the same

2048-dimensional embedding space as the audio representations. The image embedding **i** is given by:

$$\mathbf{i} = \frac{\mathbf{img}A^T + \mathbf{b}}{||\mathbf{img}A^T + \mathbf{b}||_2}, \tag{6.1}$$

where A and **b** are learned weight and bias terms, and **img** is the vector of ResNet-152 image features.

The audio branch consists of a 1-d convolutional neural network of size 6, stride 2 and 64 output channels, which sub-samples the signal along the temporal dimension. The resulting features are fed into a 4-layer bi-directional Long Short Term Memory (LSTM) with 1024 units.[1] The 1024 bi-directional units are concatenated to create a 2048 feature vector. The self-attention layer computes a weighted sum over all the hidden LSTM states:

$$\mathbf{a}_t = \text{softmax}(V \tanh(W\mathbf{h}_t + \mathbf{b}_w) + \mathbf{b}_v), \tag{6.2}$$

where $\mathbf{a}_t$ is the attention vector for hidden state $\mathbf{h}_t$, and W, V, $\mathbf{b}_w$, and $\mathbf{b}_v$ indicate the weights and biases. The learnable weights and biases are implemented as fully connected linear layers with output sizes 128 and 2048, respectively. The applied attention is then the sum over the Hadamard product between all hidden states $(\mathbf{h}_1, ..., \mathbf{h}_t)$ and their attention vector:

$$\text{Att}(\mathbf{h}_1, ..., \mathbf{h}_t) = \sum_t \mathbf{a}_t \circ \mathbf{h}_t. \tag{6.3}$$

The resulting embeddings are normalised to have unit L2 norm. The caption embedding **c** is thus given by:

$$\mathbf{c} = \frac{\text{Att}(\text{LSTM}(\text{CNN}(\mathbf{a}_1, ..., \mathbf{a}_t)))}{||\,\text{Att}(\text{LSTM}(\text{CNN}(\mathbf{a}_1, ..., \mathbf{a}_t)))||_2}, \tag{6.4}$$

where $\mathbf{a}_1, ..., \mathbf{a}_t$ indicates the caption represented as t frames of MFCC vectors and Att, LSTM and CNN are the attention layer, stacked LSTM layers, and convolutional layer, respectively.

Next, we also implement a VGS model with added VQ layers (van den Oord et al., 2017). We will refer to our regular model and the model with VQ layers as LSTM and LSTM-VQ models, respectively. Our implementation most closely follows Harwath et al. (2020a), who were the first to apply these layers in a VGS model, and showed that their model learned discrete linguistic units. VQ

[1] In **Chapter 4** we used a 3-layer Gated Recurrent Unit, but it has since then become practically feasible to train larger models on our hardware.

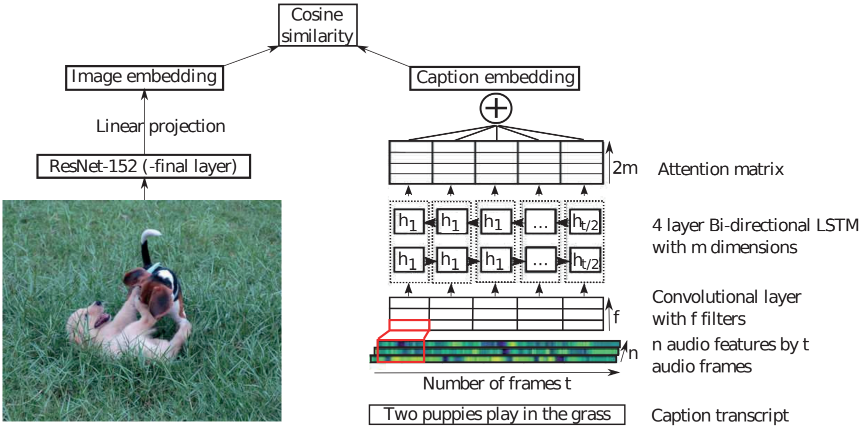
Cosine similarity
Image embedding
Caption embedding
Linear projection
ResNet-152 (-final layer)
2m
Attention matrix
h_1
$h_{t/2}$
...
4 layer Bi-directional LSTM with m dimensions
Convolutional layer with f filters
f
n
n audio features by t audio frames
Number of frames t
Two puppies play in the grass
Caption transcript

Figure 6.1: Model architecture: the model consists of two branches with the image encoder depicted on the left and the caption encoder on the right. The audio features consist of 13 MFCC with 1st and 2nd order derivatives by t frames. Each LSTM hidden state $\mathbf{h}_t$ has 1024 features which are concatenated for the forward and backward LSTM into 2048-dimensional hidden states. Vectorial attention weighs and sums the hidden states resulting in the caption embedding. The linear projection in the image branch maps the image features to the same 2048-dimensional space as the caption embedding. We calculate the cosine similarity between the image and caption embedding.

layers consist of a 'codebook' which is a set of n-dimensional embeddings. A VQ layer discretises incoming input by mapping it to the closest embedding in the codebook and passing this embedding to the next layer:

$$VQ(\mathbf{x}) = \mathbf{e}_k, \text{ where } k = \text{argmin}_j ||\mathbf{x} - \mathbf{e}_j||_2, \tag{6.5}$$

where $\mathbf{x}$ is the VQ layer input and $\mathbf{e}_j$ are the codebook embeddings.

For the LSTM-VQ model we insert VQ layers in the LSTM stack after the first and after the second LSTM layer, with 128 and 2048 codes, respectively. We use two layers because, as shown by Harwath et al. (2020a), this made a hierarchy of linguistic units emerge: the first layer best captured phonetic identity while in the second layer, several codes emerged that were sensitive to specific words.

We use our own PyTorch implementation of the models and the VQ layer described here, adapted from our previous work presented in **Chapters 2** and **4**, which is in turn most closely related to, and based on, the VGS models presented by Harwath et al. (2016) and Chrupała et al. (2017). Our implementation and data can be found on `https://github.com/DannyMerkx/speech2image/tree/CogComp2022`.

Training data

We train the model on Flickr8k (Hodosh et al., 2013), a well-known dataset of 8,000 images from the online photo sharing platform Flickr.com, with five written English captions per image. Annotators were asked to 'write sentences that describe the depicted scenes, situations, events and entities (people, animals, other objects)' (Hodosh et al., 2013). We use the spoken captions Harwath and Glass (2015) collected by having Amazon Mechanical Turk (AMT) workers pronounce the original written captions. We use the data split provided by Karpathy and Fei-Fei (2015), with 6,000 images for training and a development and test set of 1,000 images each.

Image features are extracted by resizing all images while maintaining the aspect ratio such that the smallest side is 256 pixels. Ten crops of 224 by 224 pixels are taken, one from each of the corners, one from the middle and similarly for the mirrored image. We use ResNet-152 (He et al., 2016) to extract visual features from these ten crops and then average the features of the ten crops into a single vector with 2,048 features.

The audio input consists of Mel Frequency Cepstral Coefficients (MFCCs). We compute the MFCCs using 25 ms analysis windows with a 10 ms shift. The MFCCs were created using 40 Mel-spaced filterbanks. We use 12 MFCCs and the log energy feature, and add the first and second derivatives resulting in 39-dimensional feature vectors. Lastly, we apply per-utterance cepstral mean and variance normalisation.

Training

The model is trained to embed the images and captions such that the cosine similarity between image and caption embeddings is larger for matching pairs than the similarity between mismatching pairs. The batch hinge loss L as a function of the network parameters θ is given by:

$$\begin{split} L(\theta) = \sum_{(\mathbf{c},\mathbf{i}),(\mathbf{c}',\mathbf{i}')\in B} \Big(& \max(0, \cos(\mathbf{c},\mathbf{i}') - \cos(\mathbf{c},\mathbf{i}) + \alpha) + \\ & \max(0, \cos(\mathbf{i},\mathbf{c}') - \cos(\mathbf{i},\mathbf{c}) + \alpha) \Big), \end{split} \tag{6.6}$$

where $(\mathbf{c}, \mathbf{i}) \neq (\mathbf{c}', \mathbf{i}')$, B is a minibatch of matching caption-image pairs $(\mathbf{c}, \mathbf{i})$, and the other caption-image pairs $(\mathbf{c}', \mathbf{i}')$ in the batch serve to create mismatching pairs: $(\mathbf{c}, \mathbf{i}')$ and $(\mathbf{c}', \mathbf{i})$. We take the cosine similarity and subtract the similarity of the mismatching pairs from the matching pairs such that the loss is only zero when the matching pair is more similar than the mismatching pairs by a margin α, which was set to 0.2.

Training-task performance is evaluated by caption-to-image and image-to-caption retrieval score Recall@N on the 1000-image test set. For these retrieval tasks, the caption embeddings are ranked by cosine distance to the image and vice versa, and Recall@N is the percentage of test items for which the correct image or caption was in the top N results. Furthermore, we evaluate the median rank of the correct image or caption.

Because the VQ operation is indifferentiable, a trick called *straight through estimation* is required to pass a learning signal to layers before the VQ layer (Bengio et al., 2013). Put simply, as there is no gradient for the VQ operation, the gradients for the VQ output are copied and used as an approximation of the gradients for the VQ input.

The VQ layer learns to make the codebook codes more similar to their inputs and vice versa. The first is accomplished by an exponential moving average. When a code is activated, it gets multiplied by a decay factor γ and summed with $(1-\gamma)\mathbf{x}$, where $\mathbf{x}$ is the input that activated the code. Making the inputs more similar to the codes is accomplished by a separate VQ loss, which is the mean squared error between each input and its closest code.

The networks are trained using Adam (Kingma and Ba, 2015) with a cyclic learning rate schedule based on the work by Smith (2017). The learning rate schedule varies the learning rate smoothly between a minimum of 10^{-6} and maximum of 2×10^{-4}.

We train the regular LSTM-based network for 16 epochs. Following Harwath et al. (2020a), we *warm start* the LSTM-VQ model by taking the trained LSTM network, inserting the VQ layers and training for another 16 epochs. While, unlike Harwath et al. (2020a), we did not encounter a large performance loss for *cold started* networks, we did find that a cold started VQ network frequently suffered from codebook collapse (van Niekerk et al., 2020). This is an issue where suddenly all VQ inputs are mapped to only a few (often even just one) codes and from which the model never recovers.

We trained 20 VGS models of each type (with and without VQ) using different seeds for the pseudo-random number generator, to average over random effects of weight initialisation and training data presentation order.

6.2.2 Data collection

Target words

Word learning by VGS models exploits the fact that words in the speech signal tend to co-occur with visual referents in the corresponding images. We can therefore expect that any words the system learns to recognise will be words with visual referents in the images. Hence, we limit our analysis to the recognition of nouns and verbs. We only look at high-frequency words that the model has had ample opportunity to learn to recognise.

We selected the 50 nouns and 50 verbs with the most frequent lemma in the Flickr8k database, excluding some words like 'air' and 'stand' as their referents appear in nearly every picture and, consequently, whether the words are recognised cannot be established. Other examples of rejected words are verbs such as 'try' for which it is not possible to set objective standards for the visual referent. The selected words are shown in Table 6.1.

To test word recognition performance, we present the selected target verbs and nouns in isolation. Two North American native speakers of English (one male, one female), not present in the Flickr8k database, were asked to read the target words out loud from paper. The words were recorded in isolation by asking the speakers to leave at least a second of silence in between words. To keep conditions close to those of the Flickr8k spoken captions (and other captioning databases collected through AMT), the speakers recorded the words at home using their own hardware. They were asked to find a quiet setting and record the words in a single session. They received a $20 gift card for their participation.

The nouns were presented in both their singular and plural form (where applicable)[2]. All verbs were recorded in root form, third person singular form, and progressive participle form. We did not record past tense forms as these are rarely, if ever, used in the image descriptions.

The speech data were recorded in stereo at 44.1kHz in Audacity. We downsampled the utterances to 16kHz and converted them to mono to match the

[2] 'Shorts' and 'sunglasses' are syntactically plural, but we group them under the singular nouns as their use in the data is most often in reference to a single object.

Table 6.1: Selected target nouns and verbs in order of occurrence in the training set transcripts. A * indicates nouns for which only the singular or plural form was recorded, + indicates words that were not included in the analysis because there were not enough images depicting their visual referent in the test set.

Nouns		Verbs	
dog	man	play	run
boy	girl	jump	sit
woman	water*	hold	walk
shirt	ball	ride	climb
grass*	beach	smile	pose
snow*	group	catch	carry
street	rock	leap	perform
camera	bike	fly	dance
mountain	hat	swim	eat
pool	player	pull	hang
jacket	ocean	chase	slide
basketball	sand*	splash	point
car	building	kick	throw
soccer*	swing	fight	swing
football	sunglasses*	lie	lay
shorts*	park	laugh	ski
dress	table	surf	drive
hand	tree	fall	follow
lake	hill	race	roll
toy	baby	hit	reach
tennis*+	river	wade	lean
wave	snowboarder	push	bite
bench	game	spray	paddle
surfer	stick	light+	bend
team	skateboard	cross	raise

conditions of the Flickr8k captions, after which we applied the same MFCC processing pipeline used for the Flickr8k training data.

Image annotations

We test whether the VGS model learned to recognise the recorded target words by presenting them to the model and checking whether the retrieved images contain the words' visual referents. The problem with this approach, however, is that Flickr8k contains no ground truth image annotations for such a test. The captions can serve as an indication: if annotators mention an action or object in the caption we can be reasonably sure it is visible in the picture. In contrast, it is definitely not the case that if an object or action is not mentioned, it is not in the picture. Hence, using captions as ground truth would lead to an underestimation of model performance.

We created a ground truth labelling for the visual referents of our target words by manually annotating the 1000 images in the Flickr8k test set for visual presence of each target word. For the nouns, we also indicate whether the visual referent occurred only once or multiple times in the images, allowing us to test whether the model learns to differentiate between plural and singular nouns.

There were two annotators, one covering the nouns and one the verbs. To check the quality of the annotations, the first author annotated a sample of 5% of the images. The inter annotator agreement based on this sample was $\kappa = 0.70$ for verbs and $\kappa = 0.76$ for nouns.

6.2.3 Word recognition

We take the retrieval of images containing a target word's visual referent as indicative of successful word recognition. As this is a retrieval task where multiple correct images can be found per word, we use precision@10 (P@10) to measure word recognition performance, following Havard et al. (2019). That is, for each target word embedding we calculate the cosine similarity to all test image embeddings and retrieve the ten most similar images. P@10 is then the percentage of those images that contains the visual referent according to our annotations. We excluded two target words from this analysis as there were fewer than ten test images containing their visual referent. Although we annotated whether an image contains a single or multiple visual referents, unless stated otherwise, multiple visual referents were counted as correct for a singular noun and vice versa for the purpose of calculating P@10.

We also compute P@10 scores for two baseline models. Our *random* baseline is simply the averaged score over five randomly initialised and untrained VGS models. This results in a random selection of images but since some words' visual referents occur in dozens to hundreds of test images, the recognition scores are far from zero. Our *naive* baseline is the recognition score of a model that always retrieves the ten images with the highest number of visual referents (i.e., always the same ten images, selected separately for the nouns and verbs). Note that this baseline is not realistic and requires knowledge of the contents of the test set (namely the number of visual referents per image). Still, it is useful to compare our model performance to a model that has only a single response regardless of the input.

We then examine the influence of linguistic and acoustic factors on the model's word recognition performance as measured by P@10, using a Generalised Linear Mixed Model (GLMM) with beta-binomial distribution[3] and canonical logit link function. We used the *glmmTMB* package in R (Brooks et al., 2017).

The GLMM examines the effects of signal duration (i.e., number of speech frames), speaking rate (number of phones per second), number of vowels, number of consonants, morphology (singular or plural)[4] and VQ (LSTM or LSTM-VQ model), with the VGS model's word recognition performance (P@10) as the outcome variable. As control variables, we furthermore include the (log-transformed) counts of the target word and its lemma in the training set as we expect better recognition for words that are seen more often during training. The correlation between lemma count and word count is .48, so they are expected to explain unique portions of variance. We also include speaker-ID to account for differences in recognition performance between the two speakers. Number of vowels and consonants are centered, all other non-categorical variables are standardised. VQ (LSTM $= -1$, LSTM-VQ $= 1$), morphology (plural $= -1$, singular $= 1$) and speaker ID ($\#1 = -1, \#2 = 1$) were sum coded.

The GLMM includes by-lemma and by-model (each of the 20 random initialisations) random intercepts. We first included all fixed effects that vary within lemma or model-ID as by-lemma or by-model random slopes but this model was unable to converge. As a maximal model is thus not possible, we reduced the model until it converged: we tried a zero-correlation-parameter GLMM, which also did not converge. Next, we split the GLMM into one with only the by-lemma and one with only the by-model random slopes (uncorrelated). The by-model GLMM resulted in a singular fit for the speaker ID, morphology, and VQ random slopes. After removing these by-model slopes, the combined GLMM, with all remaining uncorrelated by-lemma and by-model slopes, converged. None of the removed random slopes could be added back into the combined GLMM without causing convergence issues. The final GLMM formula is:

```
   p@10 ∼ speaking rate + duration + lemma count + word count +
#vowels + #consonants + VQ + speaker id + morphology + (1 +
speaking rate + duration + word count + #vowels + #consonants + VQ +
speaker id + morphology || lemma) + (1 + speaking rate + duration +
```

[3] Our P@10 data, which is discrete and has a floor of 0 and a ceiling of 10, is not suited for standard linear modelling. Our response variable is best described as a series of Bernoulli trails with successes and failures in terms of correct and incorrect retrieval.

[4] As seen in Section 6.3.1, word recognition results on the verbs were overall a lot worse than for the nouns so we decided not to continue our analysis on the verbs.

`lemma count + word count + #vowels + #consonants || model id)`, where the double pipe symbol (||) means that correlations between random slopes are not estimated.

6.2.4 Word competition

We perform a gating experiment to investigate word competition in our models. We present the models with the target words in segments of increasing length, using one gate per phone. Simply put, if the target word is 'dog' with the phones /d-ɔ-g/, we evaluate performance after the model has processed /d/, /d-ɔ/, and finally the whole word /d-ɔ-g/. Performance is measured in P@10 as described in 6.2.3.

For the gating experiment we need to know when each phone starts and ends. We use the Kaldi toolkit to make a forced alignment of our target words and their phonetic transcripts (Povey et al., 2011), taken from the CMU Pronouncing Dictionary available at `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

We define the word-initial cohort of a target word at a certain gate to be the set of words in the Flickr8k dataset that share the target's word-initial phone sequence up to the gate. That is, the number of words in the word-initial cohort equals the number of words that cannot be distinguished from the target given the sequence so far, and thus the number of words competing for recognition.

We define neighbourhood density as the number of words in Flickr8k that differ by exactly one phone from the target word (Vitevitch and Luce, 2016). These words are expected to compete for recognition and so affect word recognition. Research shows that words with a dense neighbourhood are harder to recognise than those with a sparse neighbourhood (Luce and Pisoni, 1998).

For both the word-initial cohort and the neighbourhood density, we use phonetic transcripts from the CMU pronouncing dictionary, which contains the transcripts for a total of 6431 words in the Flickr8k captions.

We use a GLMM to test whether the neighbourhood density and word-initial cohort size affect word recognition in our model. Furthermore, we are interested in three interaction effects: as previously discussed, we test the interactions between neighbourhood density and the word and lemma counts. The third interaction is between VQ and the number of phones processed so far (gate number). The VGS model with VQ layers is forced to map its inputs to discrete units even as early as the first gate. As the second VQ layer has been shown to learn discrete word-like representations (Harwath et al., 2020a), we might expect that words are recognised earlier, as would be indicated by a smaller effect of gate number for the LSTM-VQ model.

The GLMM's fixed effects are the neighbourhood density, gate number, the size of the word-initial cohort, VQ, morphology, the number of vowels and the number of consonants. Again we also add the occurrence frequencies of the target word and its lemma in the training set and speaker-ID to account for expected effects of training data frequency and speaker differences. The number of vowels, number of consonants and gate number are centered, all other non-categorical variables are standardised.

The GLMM has by-lemma and by-model random intercepts. We started with maximal by-lemma and by-model random slopes but had to reduce the complexity due to convergence issues, using the same procedure as described before. However, after removing all random slopes that yielded singular fits in the GLMM with only by-model random effects, the combined model (with by-model and by-lemma random effects) still failed to converge. We proceeded to use the variance estimates of the separate GLMMs to remove the smallest variance components until the combined GLMM converged. This led to the removal of all by-model random slopes and the by-lemma slopes for number of vowels and word count. The final GLMM formula for analysis of the gating experiment is:

```
   p@10 ∼ (lemma count + word count) * density + VQ * gate +
initial cohort size + speaker id + morphology + #vowels +
#consonants + (1 + density + VQ + gate + initial cohort size +
speaker id + morphology + #consonants || lemma) + (1 | model id)
```

6.3 Results

All results presented here are averaged over the 20 random initialisations of the VGS model. We first evaluate how well the models perform on the training task and compare their performance to other VGS models. The scores in Table 6.2 show the result for the speech caption-to-image and image-to-caption retrieval tasks. This indicates how well the model learned to embed the speech and images in the common embedding space. As expected, the VQ layers are beneficial to the VGS model's training task performance (Harwath et al., 2020a).

Table 6.2: Image-caption retrieval results on the Flickr8k test set. R@N is the percentage of items for which the correct image or caption was retrieved in the top N (higher is better) with 95% confidence interval. Med r is the median rank of the correct image or caption (lower is better). We compare our VGS models to previously published results on Flickr8k. '-' means the score is not reported in the cited work.

Model	Caption to Image			
	R@1	R@5	R@10	med r
Harwath and Glass (2015)	-	-	17.9±1.1	-
Chrupała et al. (2017)	5.5±0.6	16.3±1.0	25.3±1.2	48
Chapter 4	8.4±0.8	25.7±1.2	37.6±1.3	21
Wang et al. (2021)	10.1±0.8	28.8±1.3	40.7±1.4	-
LSTM	12.5±0.2	33.8±0.3	46.8±0.3	12
LSTM-VQ	12.9±0.2	34.5±0.3	47.3±0.3	12
Model	Image to Caption			
	R@1	R@5	R@10	med r
Harwath and Glass (2015)	-	-	24.3±2.7	-
Chapter 4	12.2±2.0	31.9±2.9	45.2±3.1	13
Wang et al. (2021)	13.7±2.1	36.1±3.0	49.3±3.1	-
LSTM	18.5±0.5	42.4±0.7	55.8±0.7	8
LSTM-VQ	19.6±0.6	45.4±0.7	58.1±0.7	7

6.3.1 Word recognition

In the first experiment, we presented isolated words to the model. Table 6.3 shows the average P@10 scores. The singular nouns are recognised best with P@10 scores of .519 and .529 for the LSTM and LSTM-VQ model, respectively. This means that, on average, more than five out of the ten retrieved images contain the correct visual referent. For the plural nouns the average performance is .479 and .449 for the LSTM and LSTM-VQ model, respectively. However, seven target nouns have no plural form, so the scores for plural and singular nouns are not directly comparable. Therefore, we also calculate singular noun performance only on those words that also have a plural form. The results show that singular and plural forms are recognised equally well by the LSTM model. However, the LSTM-VQ model recognises plural target words slightly less accurately than singular words.

The histograms in Figure 6.2 show the distribution of the P@10 scores by word type (noun or verb), morphology and whether the VGS model included VQ layers. This highlights that the recognition of the verbs is overall much worse than for the nouns: many verbs have a P@10 of zero, meaning they are not recognised at all. For the nouns on the other hand, only two words are not

Table 6.3: Word recognition results for each noun and verb type for the trained models, the random model, and the naive baseline. In parentheses are the recognition scores when only evaluating the subset of target words that also have plural forms.

			Baseline	
Morphology	LSTM	LSTM-VQ	Random	Naive
singular noun	.519(.479)	.529(.485)	.137	.278
plural noun	.479	.449	.140	.267
root verb	.185	.193	.082	.188
third-person verb	.176	.164	.078	.188
participle verb	.246	.260	.083	.188

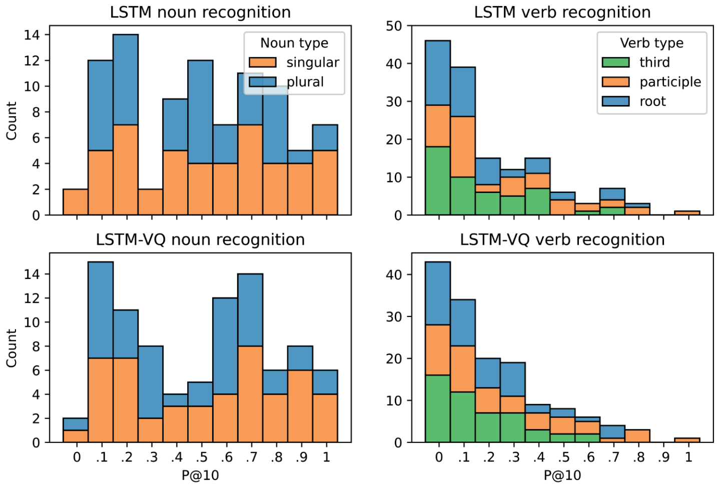

Figure 6.2: Histograms of the word recognition experiment results for each word type.

recognised at all. While both LSTM models outperform the random baseline on verb recognition, only on the participles is performance better than the naive baseline's, with scores over .7 on some words. As the recognition performance for the verbs is obviously a lot worse than for nouns, we continue our analysis on the nouns only.

Havard et al. (2019) reported a median P@10 of 0.8 on 80 nouns (from the synthetic speech database MSCOCO), while our models achieve median P@10 scores of 0.6 and 0.5 on singular and plural nouns, respectively. Even though the models recognise most nouns and even their plural forms (with only two words

Table 6.4: Estimated model effects for the word recognition GLMM and the results of Type III Wald χ^2 tests. Plural, LSTM and speaker 1 are the reference levels for Morphology, VQ and Speaker id respectively.

Effect	Estimate	Std. error	χ^2	p
Intercept	−0.26	0.70	1.20	0.27
Speaking rate	−2.03	0.91	4.98	**0.03**
Duration	−0.88	0.60	2.14	0.14
Lemma count	1.98	0.70	7.97	**0.005**
Word count	0.33	0.40	0.69	0.41
#Vowels	1.33	1.35	0.98	0.32
#Consonants	2.06	0.81	6.46	**0.01**
VQ	0.02	0.04	0.34	0.56
Speaker id	−0.37	0.25	2.13	0.14
Morphology	−0.28	0.44	0.42	0.52

per model not being recognised at all), this indicates a large drop in recognition performance going from the synthetic speech dataset used by Havard et al. (2019) to our natural speech. Note, however, that as Havard et al. used the most frequent nouns for their dataset (MSCOCO), the target words do not fully overlap with ours.

The results of the GLMM for the word recognition experiment are summarised in Table 6.4. Speaking rate and number of consonants have a significant effect on the VGS model's word recognition performance. The positive coefficient of the number of consonants indicates that words with more consonants are on average recognised better. The negative coefficient for speaking rate indicates that words are harder to recognise if they are spoken faster. Unsurprisingly, lemma count also has a significant effect on word recognition: lemmas that were seen more often during training are recognised better. The results further confirm that plural and singular nouns are recognised equally well and that there is no difference in recognition performance between the two speakers.

While overall these results show no difference in word recognition performance between the LSTM-VQ and the LSTM models, it is notable that only LSTM-VQ has a performance difference between singular and plural nouns. Similarly, LSTM-VQ performs best on the participle verb form and worse on the third person and root forms. Third person and root verbs are less frequent than participles, and plural nouns are less frequent than singulars. Hence, it may be the case that the codebook simply learns to encode frequent words better, and struggles with the less frequent word(form)s.

Table 6.5: Estimated model effects for our post-hoc testing of interaction effects and the results of Type III Wald χ^2 tests. LSTM and speaker 1 are the reference levels for VQ and Speaker id respectively. Plural and Participle are the Morphology reference levels for the noun and verb models respectively.

Effect	Estimate	Std. error	χ^2	p
		Nouns		
VQ	0.03	0.01	3.69	0.06
Word count:VQ	0.10	0.02	23.17	$<$ **0.001**
Morphology				
Singular	1.34	0.86	2.45	0.12
Singular:VQ	0.12	0.02	38.42	$<$ **0.001**
		Verbs		
VQ	-0.04	0.01	11.02	$<$ **0.001**
Word count:VQ	0.07	0.01	38.42	$<$ **0.001**
Morphology				
Root	−0.05	0.22		
Third	0.46	0.33	6.85	**0.03**
Root:VQ	−0.07	0.02		
Third:VQ	−0.002	0.02	30.86	$<$ **0.001**

To further investigate whether the VQ models recognises frequent words more accurately, we performed a post-hoc test where we refit the word recognition GLMM with an interaction between VQ and word count and between VQ and morphology. We fit separate GLMMs on the noun and verb targets, the results of which can be seen in Table 6.5. We find the expected interactions between VQ and morphology where recognition on the less frequent word forms (plural, third and root) is worse than on the more frequent forms (singular, participle) for the VQ network. Furthermore, we also find positive interactions between word count and VQ, further indicating that frequency of exposure has a greater effect on the LSTM-VQ models than on the LSTM models.

6.3.2 Word competition

The results of the GLMM for the word competition experiment are summarised in Table 6.6. Of the fixed effects of interest, neighbourhood density, gate number, word-initial cohort size and number of consonants have significant effects on word recognition performance. Furthermore, we found significant interaction effects between word count and neighbourhood density, and between VQ and gate number.

Table 6.6: Estimated model effects for the gating GLMM and the results of Type III Wald χ^2 tests. Plural, LSTM and speaker 1 are the reference levels for Morphology, VQ and Speaker id respectively.

Effect	Estimate	Std. error	χ^2	p
Intercept	−0.66	0.24	7.30	**0.007**
Lemma count	0.87	0.20	18.1	< **0.001**
Word count	0.06	0.14	0.17	0.68
#Vowels	−0.08	0.29	0.07	0.79
#Consonants	0.57	0.21	7.39	**0.007**
Density	0.51	0.20	6.60	**0.01**
Gate	0.27	0.07	12.66	< **0.001**
Initial cohort	−0.98	0.20	23.0	< **0.001**
Morphology	0.01	0.15	0.01	0.92
VQ	0.04	0.03	3.18	0.07
Speaker id	−0.11	0.07	2.36	0.12
Lemma count:density	0.19	0.13	2.09	0.15
Word count:density	−0.20	0.10	4.09	**0.04**
VQ:gate	−0.016	0.005	11.61	< **0.001**

As in the previous GLMM analysis, the number of consonants has a positive effect. The gate number (number of phones processed so far) also has a positive effect: unsurprisingly, the model is better able to recognise the target word as more of the word has been presented. This effect is modulated by the presence of VQ layers, where the negative coefficient indicates that the effect of gate is slightly smaller in the LSTM-VQ than in the LSTM models. There is a significant negative effect of word-initial cohort size. This means recognition performance is lower the more candidates there are. While neighbourhood density has an overall positive effect on word recognition, care should be taken in interpreting this effect in light of the negative interaction with word count. The positive effect would indicate that words with a higher neighbourhood density are recognised better, however the interaction indicates this effect decreases with higher word count and might become negative for the most frequent words.

6.3.3 Plurality

Using the plurality annotations of the visual referents for the noun target words, we test whether the VGS models actually differentiate between singular and plural nouns. That is, if we present it with a plural noun, does it return pictures with multiple visual referents? For this we first select only those target words which have both a plural and singular form. Then, we only keep those words

Table 6.7: Confusion matrices for singular and plural nouns indicating how many of the correctly retrieved images contained only one or multiple visual referents to the target word.

Model	#refs in image	Noun morphology	
LSTM		singular	plural
	one	3048 (57%)	2940 (51%)
	multiple	2281 (43%)	2881 (49%)
LSTM-VQ		singular	plural
	one	2857 (56%)	2631 (49%)
	multiple	2278 (44%)	2754 (51%)

which have at least ten images depicting a single visual referent and ten images with multiple visual referents. So, in theory the VGS models can achieve a perfect P@10 score on these words while also perfectly distinguishing between singular and plural nouns. This results in a final target word set of 28 nouns.

Table 6.7 shows the confusion matrices for the LSTM and LSTM-VQ models, with numbers of single- versus multiple-referent images returned when the model is presented with a singular versus plural target word. We see that both VGS models, when presented with singular nouns, more often return images with a single referent than with multiple referents. When presented with plural nouns, this difference decreases and, for LSTM-VQ, even reverses (LSTM: $\chi^2(1) = 49.8, p < 0.0001, N = 11150$; LSTM-VQ: $\chi^2(1) = 48.1,$ $p < 0.0001, N = 10520$).

Recognition of plural nouns critically depends on the plural suffix, as this is what indicates whether a target word is plural (although subtle prosodic cues might also be at play, see Kemps et al. 2005). Figure 6.3 shows the P@10 scores from the gating experiment as a function of the gate number (number of phones processed so far), averaged over words of the same length. Unsurprisingly, recognition scores tend to increase as more phones are processed. Interestingly, for the plural nouns, recognition scores tend to drop at the last phone which, except for 'men' and 'women', is the plural suffix /z/ or /s/. The average P@10 value for plural target words drops from .517 to .479 between the penultimate and final gate for the LSTM model and from .513 to .449 for the LSTM-VQ model. It seems both VGS models have difficulty processing this suffix, the LSTM-VQ model even more so than the LSTM model.

A possible explanation for the P@10 drop is that, although the plural suffix causes the model to retrieve fewer images with single visual referents and more images with multiple referents (see Table 6.7), the decrease in single-referent

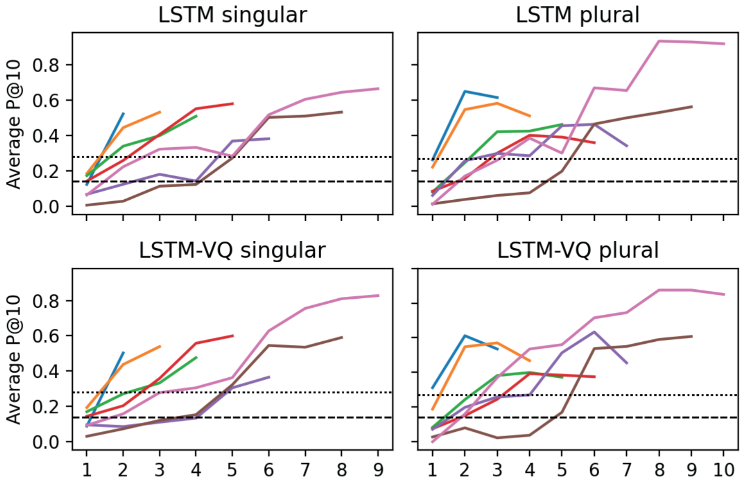

Figure 6.3: Recognition scores as a function of the gate number (the number of phones processed so far). The solid lines represent averaged P@10 scores over words with an equal number of phones (the length and colour of each line indicates the number of phones). The dotted and dashed lines represent the naive and random baseline scores, respectively.

Table 6.8: Confusion matrices for singular and plural nouns indicating how many of the correctly retrieved images contained only one or multiple visual referents to the target word. Here we show the counts at the penultimate phone and (parenthesised) the increase or decrease after having processed the final phone.

Model	#refs in image	Noun morphology	
LSTM		singular	plural
	one	2470 (578)	3339 (−399)
	multiple	1851 (430)	2694 (187)
LSTM-VQ		singular	plural
	one	2374 (483)	3171 (−540)
	multiple	1704 (574)	2565 (189)

images is greater than the increase in multiple-referent images. Table 6.8 shows the same confusion matrices as Table 6.7 but for the phone sequence up to the penultimate gate instead of the full word. The numbers between brackets indicate how the number of retrieved images changes upon processing the final phone. In case of plural nouns, the plural suffix is missing at the penultimate gate, so the model retrieves more images with a single referent, and fewer with multiple referents, than after also presenting the final phone. As can be seen in Table 6.8, and as hypothesised above, processing the plural suffix causes a drop in retrieval of single-referent images (—399) that is greater than the simultaneous increase in multiple-referent images (187), resulting in a drop in P@10 in Figure 6.3.

6.4 Discussion

In this study we investigated the recognition of isolated nouns and verbs in a Visually Grounded Speech model. We were interested in whether visual grounding allows the model to learn to recognise words as coherent linguistic units, even though our model is trained on full sentences and at no point receives explicit information about word boundaries or even that words exist at all. Havard et al. (2019) used synthetic speech to test word recognition in their VGS model; we used newly recorded real speech. We could have opted to extract the words from spoken captions in the test set but this has a few disadvantages. Firstly, words in a sentence context are often significantly reduced and reduced word forms are hard to recognise in isolation even though they are perfectly recognisable in their original sentence context (Ernestus et al., 2002). Secondly, due to co-articulation, we would not really be testing for single-word recognition unless the affected phones are removed, further reducing the word.

6.4.1 Word recognition

Our first goal was to investigate whether the VGS model can recognise words in isolation after being trained on full utterances only. Our word recognition results show that our VGS model is able to recognise isolated target nouns. We have even shown that the LSTM model recognises both plural and singular nouns equally well even though plurals occur less often in the training data than singulars. While our scores are lower than those reported by Havard et al. (2019), some difference was to be expected when working on real as opposed to synthetic speech. The average P@10 scores indicate that more than half of the top 10 retrieved images contain the visual referent and the models score well above the baselines. In fact, only four words (two in the LSTM model and two in the LSTM-VQ model) are not recognised at all, namely ‘river’ (in both models), ‘ball’ (LSTM) and ‘waves’ (LSTM-VQ). We saw that ‘river’ does return pictures of bodies of water (e.g., lakes or the ocean), and indeed it can be hard to discern the difference between a lake and a river from a picture. The fact that 'ball' is not recognised is a little baffling considering that 'basketball' has a P@10 score of .8 and 'football' a score of .4 (and pictures of either are also annotated as just 'ball').

We also tested whether models are able to recognise verbs in root, third person and participle form, the latter being the most common in the image descriptions. But even when we look only at the scores on the participle form, recognition scores for verbs are much lower than for nouns. In fact, most verbs are not recognised at all, and only 11 (LSTM) or 12 (LSTM-VQ) verbs have P@10 scores over .5. Looking at these words we see that many of them consistently occur together with an object (e.g. 'surfing', 'playing', 'skiing', 'holding' and 'racing') so the models might simply recognise the objects they co-occur with. This could be explained by our use of image features from ResNet-152, a network trained to recognise objects, not actions or body postures. However, it also recognises 'running', 'walking', 'jumping' and 'smiling', so the image features do seem to contain more information than simply the presence of a human in the image. Verb recognition in our model was far from good and this presents an interesting avenue for further research. We think it is possible for the VGS model to also learn to recognise actions, perhaps by fine-tuning parts of ResNet with the VGS model or training the visual side of the model from scratch as done by Harwath et al. (2020b).

6.4.2 Word competition

In our gating experiment, we investigated whether the model's word recognition is affected by word competition, as is the case in humans. The results show clear evidence of word-competition effects in our model. There is a strong effect of word-initial cohort size where recognition scores are lower when more words are possible given the current input sequence. We also find a positive effect of neighbourhood density that is modulated by a negative interaction with word count. This means that the effect of neighbourhood density is higher for lower-frequency words. This is in line with findings that, for humans, recognition of low-frequency words is facilitated by dense neighbourhoods whereas recognition of high-frequency words is facilitated by sparse neighbourhoods (Metsala, 1997; Goh et al., 2009).

The positive effect of neighbourhood density is contrary to what we may expect if we assume more word competition (i.e. a denser neighborhood) makes word recognition harder. Furthermore, given the strength of the interaction with word count, the neighbourhood density effect is only negative for highly frequent words. (Metsala, 1997) gives a possible explanation for the interaction between word count and neighbourhood density: during word learning, dense neighbourhoods have a positive effect on word recognition because hearing similar-sounding words facilitates learning. During word recognition, dense neighbourhoods have a negative effect because similar-sounding words compete for recognition. For infrequent words, the learning effect outweighs the competition effect, and vice versa. Our model may simply have been trained on too few of the most frequent words for the competition effect to outweigh the learning effect, explaining the overall positive effect of neighbourhood density. Together with the strong effect of initial cohort size, we argue that we do indeed see word-competition effects in our VGS model.

6.4.3 Plurality

We also investigated whether our VGS model learns the difference between singular and plural nouns. Our results show that not only is the model able to recognise target nouns in both forms but, to a limited extent, it also learns to differentiate between the two forms: when prompted with plural target nouns, the model retrieves more images with multiple referents and fewer with single referents than when prompted with single nouns (see Table 6.7). Thus, the model learns a meaningful difference between singular and plural nouns in terms of their visual representations.

P@10 scores from our gating experiment showed that words are recognised better when more of the word is processed. Yet, we also see that recognition scores are well above the baselines before word offset, which means that the model is able to recognise words from partial input. We take this to mean that the model not only recognises words, but is also able to encode useful sub-lexical information. However, at first glance, both models seemed to have trouble with the plural suffix. As shown by the results of the gating experiment, before the plural suffix recognition of plural target words is often more accurate than recognition of singulars. However, at the final phone, recognition scores of plural nouns drop and become equal or lower to that of singular nouns. While this seems to be evidence against the encoding of useful sub-lexical information, our results also show that presenting the model with plural nouns causes both models to retrieve *more* images with multiple visual referents and *fewer* images with a single referent. This indicates that the model encodes the plural suffix in a way that correctly affects recognition.

Using the recognition results from the gating experiment, we found that it is indeed only after the plural suffix that the distribution over single and multiple referents in the retrieved images shifts. At the gate just before the plural suffix (where the word is technically still singular), the model retrieves more single-referent images and fewer multiple-referent images than after the plural suffix. As previously said this is in contrast to human listeners, who are able to use subtle prosodic cues to recognise plural nouns (Kemps et al., 2005). It is not surprising that our current model, which is far from human performance in terms of word learning and recognition, is not able to exploit such cues, but this is an interesting avenue for further research.

Further analysis showed that after processing the plural suffix, the drop in single-referent images is larger than the increase in multiple-referent images. This may simply be caused by an imbalance in the test data; there are more annotations of single visual referents (3,864) than multiple visual referents (2,203). Further testing with a more balanced set of test images could show whether the performance drop seen in our gating experiment is indeed due to *correct* recognition of the plural suffix, as we would then expect the increase in retrieved multiple-referent images to outweigh the decrease in retrieved single-referent images.

6.4.4 Vector Quantisation

Our final research goal was to establish whether the addition of VQ layers to the VGS model aids in the discovery and recognition of words. Previous research had shown that VQ layers inserted into a VGS model learned a hierarchy of linguistic units; a phoneme-like inventory in the first layer, and a word-like inventory in the second layer (Harwath et al., 2020a). VQ layers discretise otherwise continuous hidden representations by mapping neighbouring speech frames to the same embedding in the codebook. We expected that this aids in the discovery of words and perhaps even allows the LSTM-VQ model to recognise words earlier in the gating experiment, as the model is forced to output discrete units from its word-like VQ layer at every time step. Moreover, the codebook size (2048) is smaller than the total number of unique words in Flickr8k so, if anything, one would expect the model to prioritise highly frequent words, of which we took the top 50 as our targets.

In all of the experiments, however, we found no evidence of the VQ layers aiding in the recognition of words: we showed that the LSTM-VQ model slightly outperforms the LSTM model on the training task (image-caption retrieval) so it cannot be the case that it is simply not a good VGS model. With regard to word recognition performance, the LSTM-VQ model recognises singular nouns better than the LSTM model, but it performs much worse at recognising plural nouns. Also noticeable is a gap between singular versus plural noun recognition that is not present in the LSTM model (when looking at the subset of words that have both a plural and singular form).

Furthermore, both GLMMs showed no main effect of the presence of VQ layers on recognition scores. We did find a negative interaction between VQ and gate number, indicating that the effect of gate is smaller for the LSTM-VQ model than for the LSTM model. Considering that final recognition performance is similar between the two models, the smaller effect of gate means the LSTM-VQ model performs better at early gates. That is, it recognises words *earlier* than the LSTM model. Together, these results indicate that the addition of VQ layers is neither beneficial nor detrimental to word recognition performance, although the LSTM-VQ model requires less of the input sequence for correct recognition. An interesting question for future research is which model performs more 'human-like', that is, which model recognises words closest to the point where humans do.

Finally, we did a post-hoc test for the interaction between VQ and morphology that shows the LSTM-VQ model has an advantage on the most frequent noun and verb forms, but performs worse on the less frequent forms. Perhaps this is due to the limited codebook size forcing the model to dedicate codes to the most frequent words in the training data.

6.4.5 Limitations

In this study, we trained and tested a model on real speech, as opposed to synthetic speech. As expected, overall recognition scores were lower than reported on synthetic speech, as natural speech is known to be more challenging for current models of speech recognition. However, the speech used in this study is read aloud speech, which is itself cleaner than spontaneous speech. In the interest of learning from data that is as natural as possible, spontaneous speech is preferred as this is the type of speech humans are most exposed to.

Furthermore, while we have shown that our model is capable of recognising words in isolation while only having seen those words in utterances, we selected only a small number of words. The small number mainly results from selecting only words with enough occurrences in the training data to reasonably expect the model to be able to learn to recognise the word, and enough occurrences of their visual referents in the test images in order to evaluate the recognition performance. On the other hand, given that the model was able to learn to recognise the words in this study after relatively little exposure, it is not unreasonable to expect the model to be able to learn more words if exposed to them.

Finally, our model depends on correlations between the speech signal and the images in order to learn to recognise meaningful constituents in utterances. Furthermore, our concept of 'recognition' of a word is defined as the retrieval of images containing its visual referent, limiting the model to 'visible' things, such as object nouns and action verbs (and not even all of those). As our results showed, the model especially struggles with verbs, even though we selected verbs with a visual referent (the actions referred to were definitely 'visible' as we were able to annotate their presence). As mentioned before, this may partly be due to the fact that we use a pretrained *object* recognition network. However, it should be mentioned that the inter-annotator agreement for verbs was lower than for nouns, so even for the annotators, it was harder to determine the presence of actions than the presence of objects. We have argued here that visual information is an important learning signal in learning language, however, still images are but a single possible source of visual information. Actions can be partly defined by the movements involved, and as such, video might be a more appropriate learning signal.

6.5 Conclusion

We investigated whether VGS models learn to discover and recognise words from natural speech. Our results show that our models learn to recognise nouns. To a lesser extent, they are capable of recognising verbs but future research should look into the image recognition side of the model to further improve this. Our models even learned to encode meaningful sub-lexical information, enabling it to interpret the visual difference signalled by the plural morphology. Contrary to what we expected based on previous research, our results show no evidence that vector quantisation aids in the discovery and recognition of words in speech. Importantly, we investigated the cognitive plausibility of the model by testing whether word competition influences our models' word recognition performance, as we know happens in humans. We have shown that two well-known measures of word competition predict word recognition in our models and found evidence in favour of a disputed interaction between word count and neighbourhood density found in human word recognition.

Taking inspiration from human learning processes, our research has shown that using multiple streams of sensory information allows our model to discover and recognise words without any prior linguistic information from a relatively small dataset of scenes and spoken descriptions. Using realistic and naturally occurring input is important for creating speech recognition models that are more cognitively plausible, and visual grounding is an important step in that direction.

7 Human sentence processing: recurrence or attention?

Recurrent neural networks (RNNs) have long been an architecture of interest for computational models of human sentence processing. The recently introduced Transformer architecture outperforms RNNs on many natural language processing tasks but little is known about its ability to model human language processing. We compare Transformer- and RNN-based language models' ability to account for measures of human reading effort. Our analysis shows Transformers to outperform RNNs in explaining self-paced reading times and neural activity during reading English sentences, challenging the widely held idea that human sentence processing involves recurrent and immediate processing and provides evidence for cue-based retrieval.

7.1 Introduction

Recurrent Neural Networks (RNNs) are widely used in psycholinguistics and Natural Language Processing (NLP). Psycholinguists have looked to RNNs as an architecture for modelling human sentence processing (for a recent review, see Frank et al. 2019). RNNs have been used to account for the time it takes humans to read the words of a text (Monsalve et al., 2012; Goodkind and Bicknell, 2018) and the size of the N400 event-related brain potential as measured by electroencephalography (EEG) during reading (Frank et al., 2015; Rabovsky et al., 2018; Brouwer et al., 2017; Schwartz and Mitchell, 2019).

Simple Recurrent Networks (SRNs; Elman 1990) have difficulties capturing long-term patterns. Alternative RNN architectures have been proposed that address this issue by adding gating mechanisms that control the flow of information over time; allowing the networks to weigh old and new inputs and memorise or forget information when appropriate. The best known of these are the Long

This chapter is based on: Danny Merkx and Stefan L. Frank. Human sentence processing: recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12-22. ACL, 2021.

Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997) and Gated Recurrent Unit (GRU; Cho et al. 2014) models.

In essence, all RNN types process sequential information by recurrence: each new input is processed and combined with the current hidden state. While gated RNNs achieved state-of-the-art results on NLP tasks such as translation, caption generation and speech recognition (Bahdanau et al., 2015; Xu et al., 2015; Zeyer et al., 2017; Michel and Neubig, 2018), a recent study comparing SRN, GRU and LSTM models' ability to predict human reading times and N400 amplitudes found no significant differences (Aurnhammer and Frank, 2019).

Unlike the LSTM and GRU, the recently introduced Transformer architecture is not simply an improved type of RNN because it does not use recurrence at all. A Transformer cell as originally proposed by Vaswani et al. (2017) consists of self-attention layers (Luong et al., 2015) followed by a linear feed forward layer. In contrast to recurrent processing, self-attention layers are allowed to 'attend' to parts of previous input directly.

Although the Transformer has achieved state-of-the art results on several NLP tasks (Devlin et al., 2019; Hayashi et al., 2019; Karita et al., 2019), not much is known about how it fares as a model of human sentence processing. The success of RNNs in explaining behavioural and neurophysiological data suggests that something akin to recurrent processing is involved in human sentence processing. In contrast, the attention operations' direct access to past input, regardless of temporal distance, seems cognitively implausible.

We compare how accurately the word surprisal estimates by Transformer- and GRU-based language models (LMs) predict human processing effort as measured by self-paced reading, eye tracking and EEG. The same human reading data was used by Aurnhammer and Frank (2019) to compare RNN types. We believe the introduction of the Transformer merits a similar comparison because the differences between Transformers and RNNs are more fundamental than among RNN types. All code used for the training of the neural networks and the analysis is available at `https://github.com/DannyMerkx/next_word_prediction`

7.2 Background

7.2.1 Human sentence processing

Why are some words more difficult to process than others? It has long been known that more predictable words are generally read faster and are more likely to be skipped than less predictable words (Ehrlich and Rayner, 1981). Predictability has been formalised as surprisal, which can be derived from LMs. Neural network LMs are trained to predict the next word given all previous words in the sequence. After training, the LM can assign a probability to a word: it has an expectation of a word w at position t given the preceding words $w_1, ..., w_{t-1}$. The word's surprisal then equals $-\log P(w_t|w_1, ..., w_{t-1})$.

Hale (2001) and Levy (2008) related surprisal to human word processing effort in sentence comprehension. In psycholinguistics, reading times are commonly taken as a measure of word processing difficulty, and the positive correlation between reading time and surprisal has been firmly established (Mitchell et al., 2010; Monsalve et al., 2012; Smith and Levy, 2013). The N400, a brain potential peaking around 400 ms after stimulus onset and associated with semantic incongruity (Kutas and Hillyard, 1980), has been shown to correlate with word surprisal in both EEG and MEG studies (Frank et al., 2015; Wehbe et al., 2014).

In this paper, we compare RNN and Transformer-based LMs on their ability to predict reading time and N400 amplitude. Likewise, Aurnhammer and Frank (2019) compared SRNs, LSTMs and GRUs on human reading data from three psycholinguistic experiments. Despite the GRU and LSTM generally outperforming the SRN on NLP tasks, they found no difference in how well the models' surprisal predicted self-paced reading, eye-tracking and EEG data.

7.2.2 Comparing RNN and Transformer

According to Levy (2008), surprisal acts as a 'causal bottleneck' in the comprehension process, which implies that predictions of human processing difficulty only depend on the model architecture through the estimated word probabilities. Here we briefly highlight the difference in how RNNs and Transformers process sequential information. The activation flow through the networks is represented in Figure 7.1.[1]

[1] Note that the figure only shows how activation is propagated through time and across layers, not how specific architectures compute the hidden states (see Elman 1990; Hochreiter and Schmidhuber 1997; Cho et al. 2014; Vaswani et al. 2017 for specifics on the SRN, LSTM, GRU and Transformer, respectively).

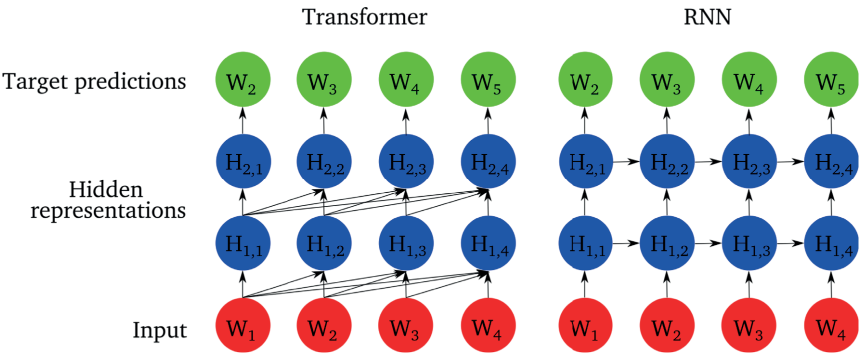

Figure 7.1: Comparison of sequential information flow through the Transformer and RNN, trained on next-word prediction.

In an RNN, incoming information is immediately processed and represented as a hidden state. The next token in the sequence is again immediately processed and combined with the previous hidden state to form a new hidden state. Across layers, each time-step only sees the corresponding hidden state from the previous layer in addition to the hidden state of the previous time-step, so processing is immediate and incremental. Information from previous time-steps is encoded in the hidden state, which is limited in how much it can encode so decay of previous time-steps is implicit and difficult to avoid. In contrast, the Transformer's attention layer allows each input to directly receive information from all previous time-steps.[2] This basically unlimited memory is a major conceptual difference with RNNs. Processing is not incremental over time: processing of word w_t is not dependent on hidden states H_1 through H_{t-1} but on the unprocessed inputs w_1 through w_{t-1}. Consequently, the Transformer cannot use implicit order information, rather, explicit order information is added to the input.

However, a uni-directional Transformer can also use implicit order information as long as it has multiple layers. Consider $H_{1,3}$ in the first layer which is based on w_1, w_2 and w_3. Hidden state $H_{1,3}$ does not depend on the order of the previous inputs (any order will result in the same hidden state). However, $H_{2,3}$ depends on $H_{1,1}, H_{1,2}$ and $H_{1,3}$. If the order of the inputs w_1, w_2, w_3 is different, $H_{1,3}$ will be the same hidden state but $H_{1,1}$ and $H_{1,2}$ will not, resulting in a different hidden state at $H_{2,3}$.

Unlike Transformers, RNNs are inherently sequential, making them seemingly more plausible as a cognitive model. Christiansen and Chater (2016) argue

[2] Language modelling is trivial if the model also receives information from future time-steps, as is commonly allowed in Transformers. Our Transformer is thus uni-directional, which is achieved by applying a simple mask to the input.

for a 'now-or-never' bottleneck in language processing; incoming inputs need to be rapidly recoded and passed on for further processing to prevent interference from the rapidly incoming stream of new material. In line with this theory, Futrell et al. (2020) proposed a model of lossy-context surprisal based on a lossy representation of memory. Recurrent processing, where input is forgotten as soon as it is processed and only available for subsequent processing through a bounded-size hidden state, is more compatible with these theories than the Transformer's attention operation.

7.3 Methods

We train LMs with Transformer and GRU architectures and compare how well their surprisal explains human behavioural and neural data. Although a state-of-the-art pretrained model can achieve higher LM quality, we opt to train our own models for several reasons. Firstly, the predictive power of surprisal increases with language model quality (Goodkind and Bicknell, 2018), so to separate the effects of LM quality from those of the architectural differences, the architectures must be compared at equal LM capability. We also need to make sure both models have seen the same sentences. Training our own models gives us control over training material, hyper-parameters and LM quality to make a fair comparison.

Perhaps most importantly, we test our models on previously collected human sentence processing data. Most popular large-scale pretrained models use efficient byte pair encodings (BPEs) as input rather than raw word tokens. This is a useful technique for creating the best possible LM, but also a crucial mismatch with how our test material was presented to the human subjects. It is not possible to directly compare the surprisal generated on BPEs to whole-word measures such as gaze durations and reading times.

7.3.1 Language Model architectures

We first trained a GRU model using the same architecture as Aurnhammer and Frank (2019): an embedding layer with 400 dimensions per word, a 500-unit GRU layer, followed by a 400-unit linear layer with a $tanh$ activation function, and finally an output layer with log-softmax activation function. All LMs used in this experiment use randomly initialised (i.e., not pretrained) embedding layers.

We implement the Transformer in PyTorch following Vaswani et al. (2017). To minimise the differences between the LMs, we picked parameters for the Transformer such that the total number of weights is as close as possible to the GRU model. We also make sure the embedding layers for the models share the same initial weights. The Transformer model has an embedding layer with 400 dimensions per word, followed by a single Transformer layer with 8 attention heads and a fully connected layer with 1024 units, and finally an output layer with log-softmax activation function. The total number of parameters for our single-layer GRU and Transformer models are 9,673,137 and 9,581,961 respectively.

We also train two-layer GRU and Transformer models. Neural networks tend to increase in expressiveness with depth (Abnar et al., 2019; Giulianelli et al., 2018) and a second layer allows the Transformer to use implicit order information, as explained above. While results (see Section 7.4.2) showed that the two-layer Transformer outperformed the single-layer Transformer in explaining the human reading data, the Transformer did not further benefit from an increase to four layers so we include only the single and two layer models. We did not see a performance increase in the two-layer GRU over the the single-layer GRU and therefore did not try to further increase its layer depth.

7.3.2 Language Model training

We train our LMs on Section 1 of the English Corpora from the Web (ENCOW 2014; Schäfer 2015), consisting of sentences randomly selected from the web. We first exclude word tokens containing numerical values or punctuation other than hyphens and apostrophes, and treat common contractions such as ‘don’t’ as a single token. Following Aurnhammer and Frank (2019) we then select the 10,000 most frequent word types from ENCOW. 134 word types from the test data (see Section 7.3.3) that were not covered by these most frequent words are added for a final vocabulary of 10,134 words. We select the sentences from ENCOW that consist only of words in the vocabulary and limit the sentence length to 39 tokens (the longest sentence in the test data). Our training data contains 5.9M sentences with a total of 85M tokens.

The LMs are trained on a standard next-word prediction task using cross-entropy loss. In the Transformer, we apply a mask to the upper diagonal of the attention matrix such that each position can only attend to itself and previous positions. To account for random effects of weight initialisation and data presentation order we train eight LMs of each type and share the random seeds between LM types so each random presentation order and embedding layer initialisation is present in both LM types. The LMs were trained for two epochs using stochastic gradient descent with a momentum of 0.9. Initial learning rates

(0.02 for the GRU and 0.005 for the Transformer) were chosen such that the language modelling performance of the GRU and Transformer models are as similar as possible. The learning rate was halved after $\frac{1}{3}$, $\frac{2}{3}$, and all sentences during the first epoch and then kept constant over the second epoch. LMs were trained on minibatches of ten sentences.

7.3.3 Human reading data

We use the self paced reading (SPR, 54 participants) and eye-tracking (ET, 35 participants) data from Frank et al. (2013) and the EEG data (24 participants) from Frank et al. (2015). In these experiments, participants read English sentences from unpublished novels. In the SPR and EEG experiments, the participants were presented sentences one word at a time. In the SPR experiment the reading was self paced while in the EEG experiment words had a fixed presentation time. In the ET experiment, participants were shown full sentences while an eye tracking device monitored which word was fixated. The SPR stimuli consist of 361 sentences, with the EEG and ET stimuli being a subset of the 205 shortest SPR stimuli. The experimental measures representing processing effort of a word are reading time for the SPR data (time between key presses), gaze duration for the ET data (time a word is fixated before the first fixation on a different word) and N400 amplitude for the EEG data (average amplitude at the centroparietal electrodes between 300 and 500 ms after word onset; Frank et al. 2015).

We exclude from analysis sentence-initial and -final words, and words directly followed by a comma. From the SPR and ET data we also exclude the word following a comma, and words with a reading time under 50 ms or over 3500 ms. From the EEG data we exclude datapoints that were marked by Frank et al. (2015) as containing artefacts. The numbers of data points for SPR, ET, and EEG were 136,727, 33,001, and 32,417, respectively.

7.3.4 Analysis procedure

At 10 different points during training (1K, 3K, 10K, 30K, 100K, 300K, 1M, 3M sentences and after the first and second epoch) we save each LM's parameters and estimate a surprisal value on each word of the 361 test sentences.

Linear Mixed Effects Regression

Following Aurnhammer and Frank (2019), we analyse how well each set of surprisal values predicts the human sentence processing data using linear mixed effects regression (LMER) models with the *MixedModels* package in Julia (Bates et al., 2019). For each dataset (SPR, ET, and EEG) we fit a baseline LMER model which takes into account several factors known to influence processing effort. The dependent variables for the SPR and ET models are log-transformed reading time and gaze duration, respectively; for the EEG model it is the size of the N400 response. All LMER models include log-transformed word frequency (taken from SUBTLEXus; Brysbaert and New 2009), word length (in characters) and the word's position in the sentence as fixed effects.

Spill-over occurs when word processing is not yet completed when the next word is read (Rayner, 1998). To account for spill-over in the SPR and ET data we include the previous word's frequency and length. For the SPR data, we include the previous word's reading time to account for the high correlation between consecutive words' reading times. For the EEG data, we include the baseline activity (average amplitude in the 100 ms before word onset). All fixed effects were standardised, and all LMER models include two-way interaction effects between all fixed effects, by-subject and by-item (word token) random intercepts, and by-subject random slopes for all fixed effects.

After fitting the baseline models, we include the surprisal values (for SPR and ET also the previous word's surprisal) as fixed effects, but no new interactions. For each LMER model with surprisal, we calculate the log-likelihood ratio with its corresponding baseline model, indicating the decrease in model deviance due to adding the surprisal measures. The more the surprisal values decrease the model deviance, the better they predict the human reading data. We call this log-likelihood ratio the goodness-of-fit between the surprisal and the data. Surprisal from the early stages of training often received a negative coefficient, contrary to the expected longer reading times and higher N400 amplitude for higher surprisal. This could be caused by collinearity, most likely between surprisal and the log-frequency, which was confirmed by their very high correlation ($> .9$) and Variance Inflation Factors (> 15) (Tomaschek et al., 2018). Apparently, the neural networks are very sensitive to word frequency before they learn to pick up on more complex relations in the data. We indicate affected goodness-of-fit scores by adding a negative sign and excluded these scores from the next stage of analysis.

Generalised Additive Modelling

As said before, it is well known that surprisal values derived from better LMs are a better fit to human reading data (Monsalve et al., 2012; Frank et al., 2015; Goodkind and Bicknell, 2018). We use generalised additive modelling (GAM) to assess whether the LMs differ in their ability to explain the human reading data at equal language modelling capability, that is, because of their architectural differences and not due to being a better LM. The log-likelihood ratios obtained in the LMER analyses are a measure of how well each LM explains the human reading data. We use each LM's average log probability over the datapoints used in the LMER analyses as a measure of the LM's language modelling capability. Separate GAMs are fit for each of the three datasets, using the R package *mgcv* by Wood (2004). LM type (single-layer GRU, two-layer GRU, etc.) is used as an unordered factor so that separate smooths are fit for each LM type. Furthermore, we add training repetition (i.e., the random training order and embedding initialisation) as a random smooth effect.

7.4 Results

7.4.1 LM Quality and Goodness-of-Fit

Figure 7.2 shows the goodness-of-fit values from the LMER models and the smooths fit by the GAMs. Overall we see the expected relationship where higher LM quality results in higher goodness of fit. The LM quality increases monotonically during training, meaning the clusters seen in the scatter-plots correspond to the points during training where the network parameters were stored. The models do seem to reach similar levels of LM quality at the end of training: the average log probability of the best LM (two-layer Transformer) is only 0.17 higher than the worst LM (two-layer GRU).

7.4.2 GAM comparisons

The bottom row of Figure 7.2 shows the estimated differences between the GAM curves in the middle row. The two-layer GRU does not seem to improve over the single-layer GRU. It outperforms the single-layer GRU only in the early stages of training on the EEG data, with the single-layer GRU outperforming it in the later stages and on the SPR data. The two-layer GRU also reaches lower LM quality on all datasets. For the Transformers we see the opposite, with the two-layer

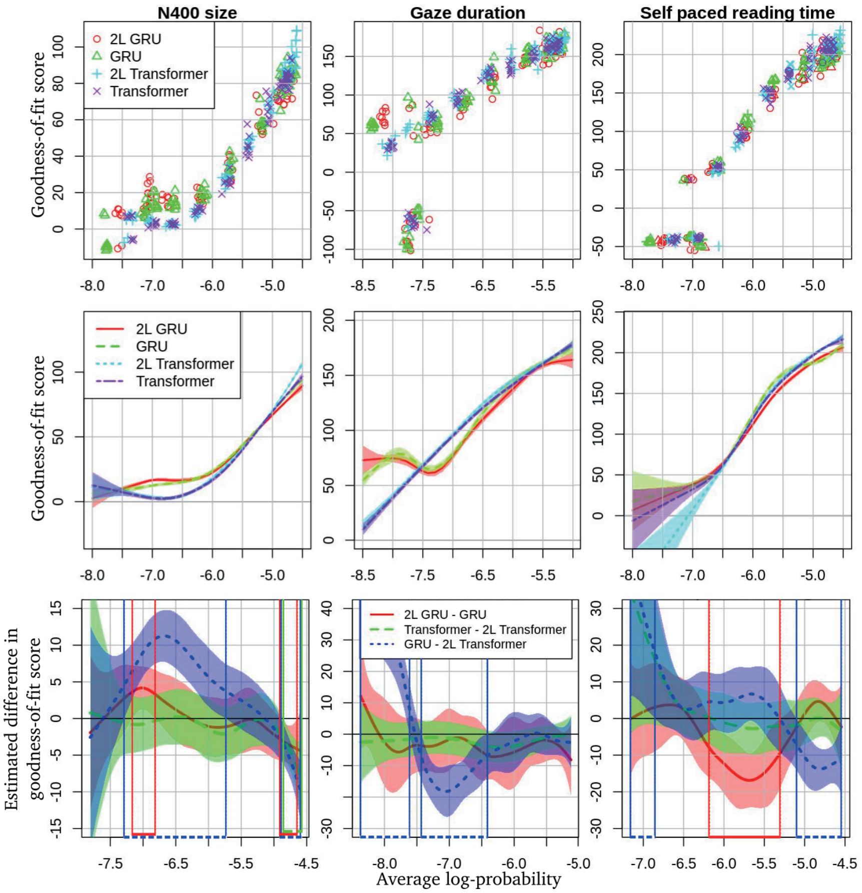

Figure 7.2: Top row: results of the linear mixed effects regression analysis grouped by LM type. These scatter-plots show the resulting goodness-of-fit values plotted against the average log-probability over the included test data. Negative goodness-of-fit indicates effects in the unexpected direction. Middle row: smooths resulting from the GAMs fitted on the goodness-of-fit data (excluding negative values), with their 95% confidence intervals. Bottom row: estimated differences in goodness-of-fit score. The markings on the x-axis and the vertical lines indicate intervals where zero is not within the 95% confidence interval. Each curve represents a comparison between two models, with an estimated difference above zero meaning the first model performed better and vice versa for differences below zero.

Transformer outperforming the single-layer Transformer on the N400 data at the end of training and never being outperformed by its shallower counterpart. The two-layer Transformer reaches a higher maximum LM quality on all datasets.

For the comparison between architectures, we only compare the best model of each type, i.e., the single-layer GRU and two-layer Transformer. The GRU outperforms the Transformer in the early stages of training (3K-300K sentences) on the N400 data, but the Transformer outperforms the GRU at the end of training on both the SPR and N400 data. On the gaze duration data, there are some performance differences with the Transformers and GRUs outperforming each other at different points during training but there are no differences in the later stages of training.

7.4.3 Shorter and longer sentences in SPR

The SPR data contains a subset of sentences longer (in number of characters) than those in the EEG/ET data. As the Transformer has unlimited memory of past inputs, the presence of longer sentences could explain why it outperformed the GRU on the SPR data. We repeated the analysis of the single- and two-layer GRUs and Transformers but only on those sentences from the SPR data that also occurred in the EEG/ET data. On these shorter sentences, there are no notable performance differences between any of the LM architectures (Figure 7.3). When we test on only those sentences that were not included in the EEG/ET experiments (i.e., the longer sentences), the Transformers outperform the GRUs as they did on the complete SPR dataset.

7.5 Discussion

We trained several language models based on Transformer and GRU architectures to investigate how well these neural networks account for human reading data. At equal LM quality, the Transformers generally outperform the GRUs. It seems that their attention-based computation allows them to better fit the self-paced reading and EEG data. This is an unexpected result, as we considered the Transformer's unlimited memory and access to past inputs implausible given current theories on human language processing.

Notably, the Transformer outperformed the GRU on the two datasets where sentences were presented to participants word by word (SPR and EEG). Neurophysiological evidence suggests that natural (whole sentence) reading places different demands on the reader than word-by-word reading, leading to different encoding and reading strategies (Metzner et al., 2015). Metzner et al. speculate that word-by-word reading places greater demand on working memory, leading

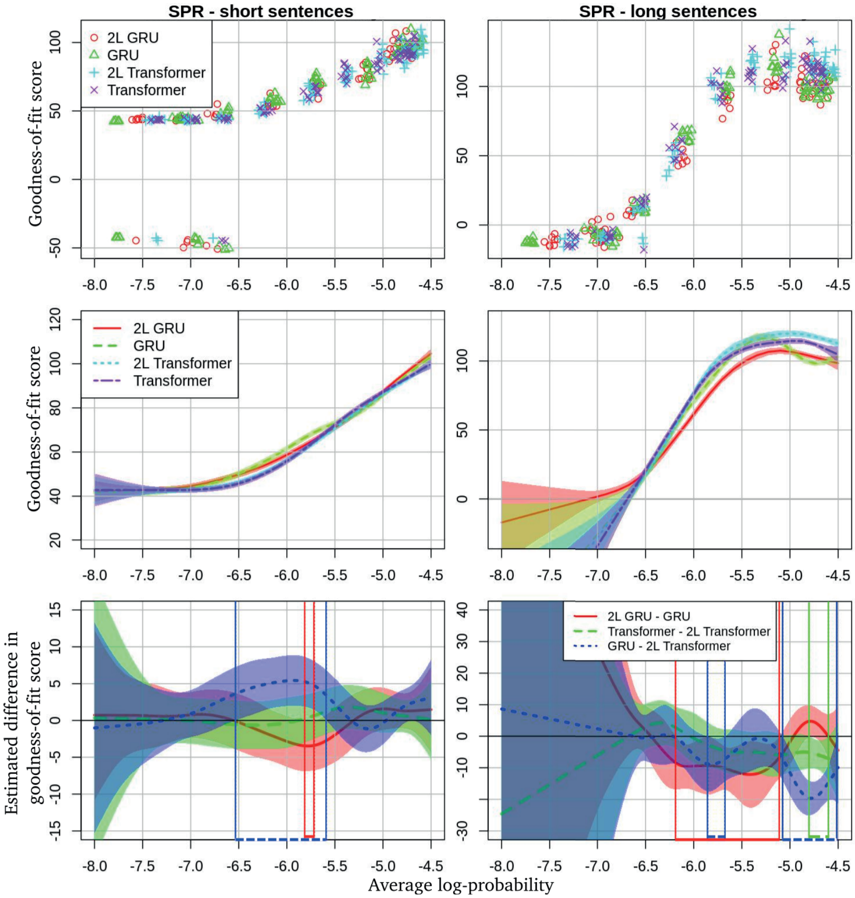

Figure 7.3: Top row: the results of the linear mixed effects regression analysis on the SPR data, where the data is split by whether the sentences were present in the ET/EEG experiment or not. These scatter-plots show the resulting goodness-of-fit values plotted against the average surprisal over the included test data. Middle row: the smooths resulting from the GAMs fitted on the goodness-of-fit data, with their 95% confidence intervals. Bottom row: the estimated differences in goodness-of-fit score with intervals where 0 is not within the 95% confidence interval marked by vertical lines and markers on the x-axis. Each curve represents a comparison between two models, with an estimated difference above zero meaning the first model performed better and vice versa for differences below zero.

to faster retrieval of previously processed material. This seems to be supported by our results; the Transformer has direct access to previous inputs and hidden states and is better at explaining the RT and N400 data from the word-by-word reading experiments. However, when we split the SPR data by sentence length, the results suggest that the Transformer's advantage is mainly due to performing better on the longer SPR sentences. On the other hand, the Transformer did outperform the GRU on the EEG dataset which contains only the shorter subset of sentences. The question remains whether the Transformer's unlimited memory is an advantage on longer sentences only, or if it could also explain why it performs better on data presented word-by-word. This question could be resolved with new data gathered in experiments where the same set of stimuli is used in SPR and EEG. Furthermore, future research could do a more specific error analysis to identify on which sentences the Transformer performs better, and perhaps even on which sentences the GRU performs better. Such an analysis may reveal the models are sensitive to certain linguistic properties allowing us to form testable hypotheses.

Surprisingly, adding a GRU layer did not improve performance, and even hurt it on reading time data, despite previous research showing that increasing layer depth in RNNs allows them to capture more complex patterns in linguistic data (Abnar et al., 2019; Giulianelli et al., 2018). The Transformers did show improvement when adding a second layer but did not improve much with four layers. As explained in 7.2, a single-layer Transformer cannot make use of implicit order information in the sequence. When adding a single layer to our Transformer, the second layer operates no longer on raw input embeddings but on contextualised hidden states allowing the model to utilise implicit input order information. Further layers increase the complexity of the model but do not make such a fundamental difference in how input is processed. In future work we could investigate how powerful this implicit order information is, and whether multi-layer Transformer LMs no longer require the additional explicit order information.

Our results raise the question how good recurrent models are as models of human sentence processing if they are outperformed by a cognitively implausible model. However, one could also interpret the results in favour of Transformers (and the attention mechanism) being plausible as a cognitive model. While unlimited working memory is certainly implausible, some argue that the capacity of working memory is even smaller than often thought (only 2 or 3 items) and that language processing depends on rapid direct-access retrieval of items from storage (McElree, 1998; Lewis et al., 2006). Cue-based retrieval theory posits that items are rapidly retrieved based on how well they match the cue (Parker et al., 2017). This is compatible with the attention mechanism used in Transformers which, simply put, weighs previous inputs based on their similarity to the current input. Cue-based retrieval models do have a recency bias due to decaying activation of memory representations but it is possible to implement a similar mechanism in Transformers (Peng et al., 2021).

Interestingly, Lewis et al. (2006) claim that serial order information is retrieved too slowly to support sentence comprehension. However, our two-layer Transformer outperforms the single layer Transformer, presumably due to order information implicitly arising as a natural result from the attention operation being performed. The use of serial order information could be compatible with cue-based retrieval models if the order information can naturally arise from the retrieval operations.

In conclusion, we investigated how the Transformer architecture holds up as a model of human sentence processing compared to the GRU. Our Transformer LMs are better at explaining the EEG and SPR data which contradicts the widely held idea that human sentence processing involves recurrent and immediate processing with lossy retrieval of previous input and provides evidence for cue-based retrieval in sentence processing.

8 Discussion and conclusion

In this dissertation I presented a model that creates visually grounded linguistic representations. I investigated two versions of the VGM: a text-based model and a speech-based model. In this final chapter I summarise the results of the studies presented in this dissertation, draw general conclusions about the VGM as a model for learning linguistic representations and suggest avenues for future research.

8.1 Summary of the results

8.1.1 Text model

In **Chapter 2** I introduced my text-based implementation of the Visual Grounding Model (VGM). The main goal was to investigate its viability as a cognitive model because caption-image retrieval models had so far been used primarily as NLP systems. I investigated whether the visually grounded sentence embeddings captured aspects of semantic sentence similarity without requiring prior lexical knowledge in the form of pretrained word embeddings. I showed that the distances between resulting sentence embeddings correlate well with human semantic similarity judgements. By using sentences from a wide range of communicative domains, I showed that it generalises well to language that is very different from the image descriptions my model is trained on. Importantly, I compared the model to InferSent, at the time a state-of-the-art sentence embedding model, and showed that the VGM performs on par with it and even outperforms it on several sentence subsets. This indicates that the VGM learns to create a semantic embedding space, and, unlike InferSent, does so without requiring pretrained word embeddings. These results are a crucial first step to testing the VGM's viability as a cognitive model as 1) visual grounding seems helpful in learning language, and 2) the model learns meaningful sentence representations without treating word learning as a separate and prior process, which I argue is a more plausible order for learning language.

After having shown that the model learns meaningful sentence representations from visual and textual input without requiring prior word-level knowledge, in **Chapter 3** I investigated whether the model learns meaningful word representations. Furthermore, while the results of **Chapter 2** suggest that visual information is helpful to the model, the experiments in **Chapter 3** were specifically aimed at uncovering the contribution of the visual grounding to the semantic knowledge captured in the word representations. I introduced changes to the VGM and presented a method to create good quality word embeddings from the model. Through two experiments using different types of human behavioural data I showed that the VGM can be used to create word representations that reflect cognitive aspects of word meaning. Furthermore, by accounting for the information captured by purely text-based methods, I showed that visual grounding allows the model to capture information that is complementary to what can be learned from text alone. In the experiment involving semantic similarity judgements, I showed that the visually grounded embeddings can outperform text-based embeddings, and not just on databases concerning concrete words that might be depicted in images. The results of the experiment involving semantic priming data shows similar results, with the grounded embeddings having more predictive value than the text-based embeddings. More importantly, in both experiments I showed that the grounded embeddings explain a unique portion of variance in the human behavioural data even after accounting for text-based embeddings pretrained on billions of tokens of text. This study shows that we cannot create representations that fully capture human word knowledge if we ignore the wider range of human sensory experience.

8.1.2 Speech model

Having created a functioning text-based model and shown that it learns to create meaningful word and sentence embeddings, I presented the speech-based model in **Chapter 4**. The speech-based model is based on my own text-based VGM and previous speech-based VGMs (e.g., Harwath and Glass 2015; Harwath et al. 2016; Chrupała et al. 2017). I implemented several improvements to its training and architecture that made it the top-performing model at the time. These improvements concern combining and adapting several state-of-the-art techniques from other types of deep learning models such as cyclical learning rates, importance sampling and vectorial self-attention. An important difference with the text-based model is that the speech-based model embeds utterances without explicit clues as to its constituent units (e.g., words). A probing experiment, where I trained simple classifiers to detect word presence in the sentence embeddings, showed that word presence can be decoded from the sentence embeddings. This shows that even though the model has no idea that words even exist, it does encode word presence, showing that it finds meaningful constituents in utterances. The results of this initial study show that the model learns to create useful sentence-level embeddings, and also picks up on the existence of meaningful sub-sentence units, two important prerequisites for my further investigations.

In **Chapter 5**, I investigated whether the VGM learns to capture sentence semantics. As no suitable evaluation data for this purpose existed, I collected spoken versions of the sentences in the evaluation toolbox used in **Chapter 2**. The results showed that the the speech-based model indeed learns to capture cognitive aspects of sentence similarity. I furthermore investigated whether there is more to creating good quality semantic embeddings besides database size. Since starting this dissertation, several, increasingly larger, databases have been released, and my best performing embedding model is trained on the smallest of these datasets. In an experiment, I created five 'new' datasets using subsets of the largest database (MSCOCO), where I vary the number of image descriptions per image, while keeping the total number of image descriptions the same. I showed that models with more descriptions per image perform better, even though the total number of descriptions is the same (and the total number of *images* thus lower), showing that paraphrases contain an important learning signal for the model.

In **Chapter 6** I showed that the VGM learns to recognise words presented in isolation, even though the model has only been exposed to full sentences during training. By collecting spoken nouns and verbs and new image annotations, I showed that the model learns to recognise these words, as it is able to use the embedded input words to retrieve images containing the objects and actions these words refer to. Furthermore, the model's word recognition performance is influenced by word competition from the word-initial cohort and neighbourhood density, two competition effects known to influence human word recognition. This shows that the representations learned by the model encode phonetic information. In an experiment, I showed that the embeddings encode information which allows the model to differentiate between singular and plural nouns, albeit not perfectly, showing that the model encodes a meaningful visual distinction between the two.

8.1.3 Working memory in language processing

Lastly, in **Chapter 7** I investigated the processing mechanisms involved in the RNN and the recently introduced Transformer. I argued that the main difference between these models is in the way they process and 'remember' previous inputs. When used as cognitive models, these mechanisms are perhaps best understood as representing the 'working memory' involved in language processing. The RNN 'forgets' input immediately after processing it and propagates past information through a single, lossy representation of what has been seen so far, whereas the Transformer has direct and lossless access to previous input. I analysed language models based on both architectures and the results showed that the Transformer best predicts human sentence processing effort as measured by reading times and brain potentials. Even though, at face value, the Transformer's processing mechanism seems a cognitively implausible analogy to working memory, I argued that it is actually quite compatible with cue-based theories of working memory retrieval (Parker et al., 2017).

8.2 Overall conclusions

The results presented above show that the VGM is able to create representations at the sentence and word levels that capture cognitive aspects of word and sentence meaning. Importantly, the VGM is able to do so without using pretrained word embeddings, and even encodes information that is complementary to such pretrained embeddings from relatively little exposure to language. The speech-based model goes even further, as it does not even receive the prior information that words exist. It learns to create meaningful sentence embeddings while at the same time learning to recognise words and encoding word-level semantics. The results show that incorporating the visual modality into the learning process has multiple benefits. Firstly, the VGM is able to exploit the correlations between the spoken and visual inputs to identify patterns (i.e., words) in the speech signal. Secondly, the model is able to incorporate visual information into its representations, capturing aspects of word meaning that purely text-based embeddings do not. I showed for instance in **Chapter 3** that the model learns information that is complementary to word embeddings created from huge corpora and in **Chapter 6** I showed that the model learns the visual distinction between singular and plural nouns.

Throughout this dissertation, I mentioned that my presentation of the VGM as a cognitive model is inspired by two theories, namely the usage-based theory of language and the embodied cognition theory. I reiterate that the VGM is not a full computational account of child language acquisition, usage-based language theory or the embodied cognition theory. However, I tested here a model that incorporates key aspects of these theories: from the usage-based theory of language, the idea that the utterance is the basic unit of language, from which smaller units follow through use and pattern finding; and from embodied cognition theory the idea that all our sensory experiences contribute to these linguistic units.

So, what has the work in this dissertation taught us about these theories besides converging evidence for their main tenets? As mentioned in the introduction, Tomasello considers pattern finding and intention reading the two most important cognitive skills that allow children to learn language (Tomasello, 2009). Pattern finding amounts to learning to identify slots in utterances and learning which utterance constituents can be ‘cut and pasted’ to create novel utterances (Pine and Lieven, 1993; Tomasello, 2000). VGMs show that visual information can play an important role in pattern finding, by exploiting the correlations between the spoken and visual input. Furthermore, joint visual attention between parent and child might play an important role in child language acquisition as shown for instance by the VGM work by Räsänen and Khorrami (2019). The work presented here revealed an important connection between the usage-based theory of language and embodied cognition theory. That is, the two most important skills required for language learning, according to the usage-based theory, might critically depend on the fact that human language processing utilises multiple sensory modalities.

8.3 Future work

In this dissertation I used a multi-modal model that combines vision with text or speech. However, the VGM used in this dissertation is limited to learning from still images, and descriptions thereof. While I showed that we can to a certain extent learn language from this data, there is still much to gain in terms of the ‘naturalness’ of the data. One direction that research is already beginning to explore is using video rather than images, as for example by Palaskar et al. (2018) and Nikolaus et al. (2022). In **Chapter 6**, I concluded that while the model is capable of learning to recognise nouns, its recognition of verbs was bad enough to not warrant further investigation in the following experiments. While some actions (e.g., skateboarding) have strong visual correlates even in still images, an action is mostly defined by the movements involved. It makes sense that in order to learn about actions, one needs to consider visual information that captures these movements. Furthermore, I considered here only read speech that was recorded in relatively noiseless environments. More 'natural' unprepared casual speech sounds different from read speech, and research suggests we cannot ignore these differences if we want to understand language processing (Tucker and Ernestus, 2016).

Even though vision is probably the most important of our senses, the human sensory experience is of course richer than that. What's more, people are capable of learning language without vision, showing that our other senses do not just complement vision, but can be sufficient to learn language. Furthermore, according to embodied cognition theory, *all* our sensory experiences contribute to meaning. In this dissertation, the auditory information stream consisted only of linguistic input (spoken words). However, auditory information associated with words entails more than just pronunciation. A dog barks, a fridge hums, and one could say a tune is more defined by its sound than any of its other properties. Kiela and Clark (2015) present a multi-modal learning model that combines words with associated sounds, (e.g., 'dog' and barking), but this type of multi-modal grounding has not received a lot of attention since. Other senses, such as smell and taste, have received even less attention as a secondary source of information in language learning models.

In the introduction I mentioned the intention to move away from the definition of linguistic units as 'the things between spaces' in alphabetic languages. Indeed, one of the premises of the usage-based theory of language is that linguistic units can be for instance multi-word expressions. Ironically, throughout this thesis, I have still depended on the 'the things between spaces' definition of words in the chapters concerning word representations. An important reason for this is simply that the available evaluation datasets, such as the priming database used in Chapter 3, all depend on this definition of a word as well. If you want to test linguistic units that are not 'the things between spaces', you first need a hypothesis about what these units are instead. So, there is a need for alternative definitions and evaluation data to go with these. In a study of written language processing, Yang et al. (2022) used the Less is Better (LiB) model (Yang et al., 2020) to derive such hypotheses and validate them in an eye-tracking experiment. Briefly put, the LiB model tries to find linguistic units in an unsupervised manner by 1) trying to minimise the number of units in a text, while 2) trying to minimise the number of unique tokens in the lexicon. This creates a trade-off between storing multi-word units and limiting the total number of units, where multi-word expressions will only be stored in the lexicon if they are frequent (e.g., treating entire sentences as single units reduces the number of units per text, but results in a large amount of unique units in the lexicon). Importantly, Yang et al. (2020) test the cognitive validity of these units, and show that the units found by the LiB model better predict eye fixation during reading than the words as defined as 'the things between spaces'.

Furthermore, in this dissertation I offer a critical note on the trend in deep learning to focus on larger models and larger datasets as a means of improving models. Even if cognitive plausibility is not an issue (as in NLP applications), my results show that by simply considering a different, complementary source of information, the model is able to do a lot with a little. Throughout this dissertation, I mostly used Flickr8k, the first and smallest of the well known image-caption databases, even though larger ones have been released since. In **Chapters 2** and **5** I directly compared models trained on different databases and found that the models trained on larger databases did not even perform that much better if they performed better at all. Even still, the observation that bigger models perform better is not that interesting and only serves to increase the start-up costs of doing research. While others have moved on to bigger databases, there is more to get out of Flickr8k and I think it is safe to say that such innovations, unsurprisingly, generalise to bigger datasets as well. I hope the future trend in deep learning will be the realisation that there are aspects that define a database besides its size.

Lastly, in **Chapter 7** I investigated the Transformer as a computational model of human sentence processing. Since starting this dissertation, the Transformer has caused quite a stir in the area of NLP and deep learning by breaking several performance records. It has more recently also begun to gain some recognition in the field of cognitive modelling. While I have tried to create a VGM implementation based on the Transformer, I could not get better results or even comparable results to the models based on RNNs. Whether that is a problem with my implementation or training settings remains to be seen, but the Transformer is an obvious candidate for further improvements in the architecture of VGMs.

8.4 Closing remarks

While many computational linguistic models assume prior knowledge of linguistic units, few models actually try to explain where these units come from and how they are learned. The representations learned by distributional semantics models, while useful, are cognitively implausible. In this dissertation, I argued that in order to create better linguistic representations for cognitive models, we need models that are informed by how humans learn language. Visually grounded models of language learning take a step in this direction by 1) learning from speech rather than text, as humans do, 2) considering the wider range of sensory experiences humans have beyond linguistic exposure and 3) treating sentence and word learning as a single end-to-end process rather than isolated consecutive processes. Furthermore, the fact that the VGM presented in this study learns from relatively little linguistic exposure and is still competitive with and complementary to models trained on databases far larger than any human can read in their lifetime shows that learning language is about more than simply the quantity of linguistic exposure.

This realisation is important for cognitive models and even for NLP applications. I believe that even though increasingly large training databases and deeper, more complex neural networks might improve model performance for now, the marginal benefits of adding a few billion more sentences or network units will decrease without ever closing the gap between model and human performance. In order to understand the human capacity for language, we need to understand the linguistic representations involved. We should not forget that human performance is still a golden standard in many machine learning tasks; we need to create models of the mental lexicon that are informed by how humans learn language not just for the sake of being more cognitively plausible, but because cognitively plausible linguistic representations will be better representations. The human experience is rich and varied, and we need to look beyond text data when trying to learn about language.

Bibliography

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. Word Embeddings have Complementary Roles in Decoding Brain Activity. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66, 2018.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203. ACL, 2019.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2009)*, pages 19–27, 2009.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval*, pages 385–393. ACL, 2012.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. Semeval-2013 shared task: Semantic textual similarity. In *SemEval*, pages 32–43. ACL, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval*, pages 81–91. ACL, 2014.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015

Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *SemEval*, pages 252–263. ACL, 2015.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval*, pages 497–511. ACL, 2016.

Christoph Aurnhammer and Stefan L. Frank. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 112–118, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

Lawrence W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59(1): 617–645, 2008.

Douglas Bates, Phillip Alday, Dave Kleinschmidt, José Bayoán Santiago Calderón, Andreas Noack, Tony Kelman, Milan Bouchet-Valat, Yakir Luc Gagnon, Simon Babayan, Patrick Kofod Mogensen, Morten Piibeleht, Michael Hatherly, Elliot Saba, and Antoine Baldassari. Juliastats/mixedmodels.jl: v2.2.0. *Juliastats.org*, 2019.

Helen Benedict. Early lexical development: Comprehension and production. *Journal of Child Language*, 6(2), 1979.

Yoshua Bengio, Nicholas Léonard, and C. Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv: 1308.3432*, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124, 2016.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146., 2017.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642. ACL, 2015.

Martin D. S. Braine and Melissa Bowerman. Children's first word combinations. *Monographs of the Society for Research in Child Development*, 41(1):1–104, 1976.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017.

Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41:1318–1352, 2017.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2013.

Marc Brysbaert and Boris New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990, 2009.

Daniel Cer, Mona Diab, Eneko E. Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *SemEval*, pages 1 – 14, 2017.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv: 1504.00325*, pages 1–7, 2015.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Morten H. Christiansen and Nick Chater. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:E62, 2016.

Grzegorz Chrupała. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 22:673–707, 2022.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 613–622. ACL, 2017.

Grzegorz Chrupała, Lieke Gelderloos, Ákos Kádár, and Afra Alishahi. On the difficulty of a distributional semantics of spoken language. In *Proceedings of the Society for Computation in Linguistics*, volume 2, pages 167–173, 2018.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4146–4156. ACL, 2020.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Workshop on Deep Learning*, 2014.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. Imagined visual representations as multimodal embeddings. In *The 31st Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, pages 4378–4384, 2017.

Alexis Conneau and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Language Resources and Evaluation Conference (LREC)*, 2018.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. ACL, 2017.

Suzanne Cotton and François Grosjean. The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics*, 35 (1):41–48, 1984.

Cedric de Boom, Steven van Canneyt, Steven Bohez, Thomas Demeester, and Bart Dhoedt. Learning semantic similarity for very short texts. In *International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234. IEEE, 2015.

Simon De Deyne, Amy Perfors, and Daniel J. Navarro. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (ICJAI)*, pages 4806–4810. AAAI Press, 2017.

Simon De Deyne, Danielle J. Navarro, Guillem Collell, and Andrew Perfors. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922, 2021.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 Computer Vision and Pattern Recognition Conference (CVPR)*, pages 248–255. IEEE, 2009.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 260–270. ACL, 2018.

Steven Derby, Paul Miller, and Barry Devereux. Analysing word representation from the input and output embeddings in neural network language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 442–454, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 4171–4186, 2019.

Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20, 2018.

Jennifer Drexler and James Glass. Analysis of audio-visual features for unsupervised speech recognition. In *2017 International Workshop on Grounding Language Understanding, GLU*, pages 57–61, 2017.

Giovanni B. F. D'Arcais, Robert Schreuder, and Ge Glazenborg. Semantic activation during recognition of referential words. *Psychological Research*, 45(1): 39–49, 1985.

Susan F. Ehrlich and Keith Rayner. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20:641–655, 1981.

Frank Eisner and James M. McQueen. *Stevens' handbook of experimental psychology, fourth edition*, volume 3 Language & thought, chapter Speech perception, pages 1–47. John Wiley, fourth edition, 2018.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Jeffrey L. Elman. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582, 2009.

Jeffrey L Elman and James L McClelland. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2):143–165, 1988.

Karen D. Emmorey and Victoria A. Fromkin. *The mental lexicon*, volume 3, page 124–149. Cambridge University Press, 1988.

Mirjam Ernestus, Harald Baayen, and Rob Schreuder. The recognition of reduced word forms. *Brain and Language*, 81:162–173, 2002.

Mark Everingham, Luc Van Gool, Christpher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008. URL `http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html`.

Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. VSE++: improved visual-semantic embeddings. *ArXiv preprint ArXiv:1707.05612*, 2017.

Radek Fer, Pavel Matejka, Frantisek Grezl, Oldrich Plchot, Karel Vesely, and Jan Honza Cernocky. Multilingually trained bottleneck features in spoken language recognition. *Computer Speech & Language*, 46(Supplement C):252–267, 2017.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

Lucia Foglia and Robert A Wilson. Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(3):319–325, 2013.

Stefan L. Frank and Roel M. Willems. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203, 2017.

Stefan L. Frank, Irene F. Monsalve, Robin L. Thompson, and Gabriella Vigliocco. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190, 2013.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11, 2015.

Stefan L. Frank, Padraic Monaghan, and Chara Tsoukala. Neural network models of language acquisition and processing. In Peter Hagoort, editor, *Human Language: from Genes and Brains to Behavior*, pages 277–291. Cambridge, MA: The MIT Press, 2019.

Richard Futrell, Edward Gibson, and Roger P. Levy. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3), 2020.

Victoria M. Garlock, Amanda C. Walley, and Jamie L. Metsala. Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language*, 45(3):468–492, 2001.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248. ACL, 2018.

Arthur M. Glenberg. Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Journal of Experimental Psychology*, 69(2):165–171, 2015.

Winston D. Goh, Lidia Suáres, Melvin J. Yap, and Seok Hui Tan. Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects. *Psychonomic Bulletin & Review*, 16(5):882–887, 2009.

Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, 2018.

Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A Search Space Odyssey. *Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8, 2001.

David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE, 2015.

David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems 29*, pages 1858–1866, 2016.

David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *ICLR 2020 The Ninth International Conference on Learning Representations*, pages 1–22, 2020a.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, 128:620–641, 2020b.

William Havard, Laurent Besacier, and Jean-Pierre Chevrot. Catplayinginthesnow: Impact of prior segmentation on a model of visually grounded speech. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 291–301. ACL, 2020.

William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Word recognition, competition, and activation in a model of visually grounded speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 339–348. ACL, 2019.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. Findings of the third workshop on neural generation and translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4): 665–695, 2015.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1367–1377. ACL, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899, 2013.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*. NeurIPS, 2020.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot Ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Keith A. Hutchison, David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer, and Erin Buchanan. The semantic priming project. *Behaviour Research Methods*, 45:1099–1114, 2013.

Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 11th International Conference on Neural Information Processing Systems (NIPS)*, pages 487–493, 1998.

Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. In *The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018.

Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, pages 3677–3681, 2017a.

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic keyword spotting by learning from images and speech. *arXiv preprint arXiv:1710.01949*, 2017b.

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 27(1):89–98, 2019.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson E. Y. Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456, 2019.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

Rachèl J. J. K. Kemps, Mirjam Ernestus, Robers Schreuder, and R. Harald Baayen. Prosodic cues for morphological complexity: The case of dutch plural nouns. *Memory & Cognition*, 33:430–446, 2005.

Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2461–2470, 2015.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In *Proceedings of NAACL-HLT 2018*, pages 408–418. ACL, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–15, 2015.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302, 2015.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using Fisher Vectors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

Xaver Koch and Esther Janse. Speech rate effects on the processing of conversational speech across the adult life span. *The Journal of the Acoustical Society of America*, 139(4), 2016.

Marta Kutas and Steven A. Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(11):203–206, 1980.

Kenneth Leidal, David Harwath, and James Glass. Learning modality-invariant representations for speech and images. In *The 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 424–429. IEEE, 2017.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3): 1126–1177, 2008.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10:447–454, 2006.

Elena Lieven, Heike Behrens, Jennifer Speares, and Michael Tomasello. Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30 (2):333–370, 2003.

Margery Lucas. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4):618–630, 2000.

Paul A. Luce and David B. Pisoni. Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, 19:1–36, 1998.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. ACL, 2013.

L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2623–2631. IEEE, 2015.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. The role of image representations in vision to language tasks. *Natural Language Engineering*, 24(3): 415–439, 2018.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78, 2017.

William D. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25(1):71–102, 1987.

Brian McElree. Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory and Language*, 38(2):225–252, 1998.

Danny Merkx and Stefan L. Frank. Learning semantic sentence representations from visually grounded language without lexical knowledge. *Natural Language Engineering*, 25(4):451–466, 2019.

Danny Merkx and Stefan L. Frank. Human sentence processing: recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. ACL, 2021.

Danny Merkx, Stefan Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pages 1841–1845, 2019.

Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Semantic sentence similarity: size does not always matter. In *INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, pages 4393–4397, 2021.

Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–11. ACL, 2022.

Danny Merkx, Sebastiaan Scholten, Stefan L. Frank, Mirjam Ernestus, and Odette Scharenborg. Modelling word learning and recognition using visually grounded speech. *Cognitive Computation*, forthcoming.

Jamie L. Metsala. An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition*, 25(1): 47–56, 1997.

Paul Metzner, Titus von der Malsburg, Shravan Vasishth, and Frank Rösler. Brain responses to world knowledge violations: A comparison of stimulus- and fixation-triggered event-related potentials and neural oscillations. *Journal of Cognitive Neuroscience*, 27(5):1–10, 2015.

Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553. ACL, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv: 1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, page 3111–311, 2013b.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 196–206, 2010.

Irene F. Monsalve, Stefan L. Frank, and Gabriella Vigliocco. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 398–408, 2012.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya

Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*, 2017.

Qouc Bao Nguyen, Jonas Gehring, Markus Müller, Sebastian Stücker, and Alex Waibel. Multilingual shifting deep bottleneck features for low-resource asr. In *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 5607–5611. IEEE, 2014.

Mitja Nikolaus, Afra Alishahi, and Grzegorz Chrupała. Learning english with peppa pig. *arXiv preprint arxiv:2202.12917*, 2022.

Dennis Norris. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234, 1994.

Dennis Norris and James McQueen. Shortlist b: A bayesian model of continuous speech recognition. *Psychological review*, 115:357–95, 2008.

Dennis Norris, James M. McQueen, and Anne Cutler. Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5):1209, 1995.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multimodal speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5774–5778, 2018.

Dan Parker, Michael Shvartsman, and Julie A. Van Dyke. *Language processing and disorders*, chapter The cue-based retrieval theory of sentence comprehension: New findings and new challenges, pages 121–144. Cambridge Scholars Publishing, 2017.

Raj Nath Patel, Prakash B. Pimpale, and M Sasikumar. Recurrent Neural Network based Part-of-Speech Tagger for Code-Mixed Social Media Text. *arXiv preprint arXiv: 1611.04989*, 2016.

Diane Pecher, René Zeelenberg, and Jeroen Raaijmakers. Does pizza prime coin? perceptual priming in lexical decision and pronunciation. *Psychological Research*, 45(4):339–354, 1984.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *The Ninth International Conference on Learning Representations*, 2021.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. ACL, 2014.

Marco A. Petilli, Fritz Günther, Alessandra Vergallito, Marco Ciapparelli, and Marco Marelli. Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language*, 117, 2021.

Julian M. Pine and Elena Lieven. Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech. *Journal of Child Language*, 20:551–571, 1993.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, pages 1–4. IEEE Signal Processing Society, 2011.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of NAACL-HLT 2018*, pages 529–535. ACL, 2018.

Milena Rabovsky, Steven S. Hansen, and James L. McClelland. Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2:693–705, 2018.

Okko Räsänen and Khazar Khorrami. A computational model of early language acquisition from audiovisual experiences of young infants. *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pages 3594–3598, 2019.

Okko Räsänen and Heikki Rasilo. A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological review*, 122 (4):792, 2015.

Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

Judith Rispens, Anne Baker, and Iris Duinmeijer. Word recognition and nonword repetition in children with language disorders: The effects of neighborhood density, lexical frequency, and phonotactic probability. *Journal of Speech, Language, and Hearing Research*, 58(1):78–92, 2015.

Armand S. Rotaru, Gabriella Vigliocco, and Stefan L. Frank. Modeling the Structure and Dynamics of Semantic Processing. *Cognitive Science*, pages 1–28, 2018.

Deb Roy and Alex Pentland. Learning words from natural audio-visual input. In *5th International Conference on Spoken Language Processing*, pages 1279–1282, 1998.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the Association for Computing Machinery*, 8(10): 627–633, 1965.

Roland Schäfer. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora*, pages 28–34, 2015.

Odette Scharenborg. Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America*, 127 (6):3758–3770, 2010.

Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark A. Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the “speaking rosetta” jsalt 2017 workshop. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4979–4983, 2018.

Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:964–975, 2020.

Sebastiaan Scholten, Danny Merkx, and Odette Scharenborg. Learning to recognise words using visually grounded speech. In *Proceedings of the IEEE International Conference on Circuits and Systems*, pages 1–5. IEEE, 2021.

Robert Schreuder, Giovanni B. F. D'Arcais, and Ge Glazenborg. Effects of perceptual and conceptual similarity in semantic priming. *Journal of Memory and Language*, 38(4):401–418, 1998.

Dan Schwartz and Tom Mitchell. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 43–57. ACL, 2019.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 959–962. Association for Computing Machinery, 2015.

Scharolta Katharina Sienčnik. Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243, 2015.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017.

Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.

Lynn S. Snyder, Elizabeth Bates, and Inge Bretherton. Content and context in early lexical development. *Journal of Child Language*, 8(3), 1981.

Robert Speer and Joshua Chin. An Ensemble Method to Produce High-Quality Word Embeddings. *arXiv preprint arXiv: 1604.01692*, 2016.

Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. Fine-grained grounding for multimodal speech recognition. In *Findings of Empirical Methods in Natural Language Processing (EMNLP) 2020*, pages 2667–2677, 2020.

Louis ten Bosch, Lou Boves, Benjamin Tucker, and Mirjam Ernestus. Diana: towards computational modeling reaction times in lexical decision in north american english. In *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, pages 1576,1580, 2015.

Fabian Tomaschek, Peter Hendrix, and R. Harald Baayen. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71:249–267, 2018.

Michael Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1/2):61–82, 2000.

Michael Tomasello. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press, 2009.

Benjamin V. Tucker and Mirjam Ernestus. Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *Mental Lexicon*, 11:375–400, 2016.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc., 2017.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. In *INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, pages 4836–4840, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aiden N. Gomes, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-Embeddings of Images and Language. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–12, 2016.

Michael S. Vitevitch and Paul A. Luce. Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, 2:75–94, 2016.

Xinsheng Wang, Tian Tian, Jihua Zhu, and Odette Scharenborg. Learning fine-grained semantics in spoken language using visual grounding. In *Proceedings of the IEEE International Conference on Circuits and Systems*, pages 1–5, 2021.

Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig. A tree-based decoder for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Zhong-Qiu Wang and DeLiang Wang. A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(4):796–806, 2016.

Andrea Weber and Odette Scharenborg. Models of processing: lexicon. *WIREs Cognitive Science*, pages 387–401, 2012.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, 2014.

Jônatas Wehrmann, Anderson Mattjie, and Rodrigo C. Barros. Order embeddings and character-level convolutions for multimodal alignment. *Pattern Recognition Letters*, 102:15–22, 2018.

Simon N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

Haoyan Xu, Brian Murphy, and Alona Fyshe. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021, 2016.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell:

Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 169–176, 2015.

Jinbiao Yang, Stefan L. Frank, and Antal van den Bosch. Less is better: A cognitively inspired unsupervised model for language segmentation. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 33–45. ACL, 2020.

Jinbiao Yang, Antal van den Bosch, and Stefan L. Frank. Unsupervised text segmentation predicts eye fixations during reading. *Frontiers in Artificial Intelligence*, 5:1–13, 2022.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, 2018.

Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schluter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2462–2466, 2017.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using Places database. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International Conference on Computer Vision*. IEEE, 2015.

Samenvatting

Hoewel bijna elk cognitief model van taalverwerking aanneemt dat mensen mentale representaties van woorden hebben, zijn er maar weinig modellen die verklaren waar die representaties vandaan komen. Makers van modellen gaan ervan uit dat deze representaties er gewoon zijn, en gebruiken dan bijvoorbeeld zogenaamde woordvectoren die al door een ander model geleerd zijn. Hoewel deze representaties hun nut veelvoudig bewezen hebben, zijn ze voor een cognitief model niet heel erg plausibel; het leren van zulke woordvectoren is afhankelijk van geschreven taal, en vergt meer data dan een mens ooit kan verwerken. Kinderen leren hun moedertaal terwijl ze aan veel minder taal worden blootgesteld dan onze computermodellen en bovendien leren ze om te praten en luisteren lang voordat ze kunnen lezen en schrijven. Centraal in dit proefschrift staat het idee dat mensen dit kunnen omdat we informatie van al onze zintuigen gebruiken bij het leren, en dan met name ons zicht. Het doel van dit proefschrift was om meer cognitief plausibele taalrepresentaties te ontwikkelen, met een model dat taal en visuele informatie combineert. Om dat doel te bereiken, mocht het model geen voorkennis van taal meekrijgen, en moest het cognitief plausibele hoeveelheden data gebruiken. Daarnaast werd er in de tweede helft van het proefschrift geen tekstdata meer gebruikt, maar leert het model, net als mensen, direct van spraak.

In **Hoofdstuk 2** introduceerde ik het model. Dit model leert zinsrepresentaties maken aan de hand van plaatjes met bijbehorende beschrijvingen. Ik onderzocht hoe het model aspecten van zinssemantiek leerde encoderen in deze representaties, en toonde aan dat de representaties correleren met menselijke intuïtie over de gelijkheid in betekenis van zinsparen. Het model doet dit, anders dan veel vergelijkbare modellen, zonder voorkennis over de betekenis van woorden nodig te hebben. Het model behandelt het leren van woord- en zinsbetekenis dus niet als twee aparte en opeenvolgende processen, wat een plausibelere manier is om taal te leren.

De experimenten met menselijke gedragsdata in **Hoofdstuk 3**, toonden aan dat het model gebruikt kan worden om woordrepresentaties te maken die correleren met cognitieve aspecten van woordbetekenis. Cruciaal aan deze resultaten was dat de combinatie van taal en visuele informatie een unieke deel van de variantie in de gedragsdata verklaarde, bovenop wat er verklaart kon worden met woordvectoren getraind op miljarden woorden aan geschreven tekst. Dit hoofdstuk laat zien dat de woordrepresentaties van onze modellen nooit de menselijke taalkennis kunnen benaderen zolang we niet kijken naar het hele scala van de menselijke zintuiglijke ervaring.

In **Hoofdstuk 4** presenteerde ik het model dat gebaseerd is op spraak. Het spraakmodel leert van hele zinnen, en krijgt daarbij geen expliciete informatie dat een zin uit meerdere betekenisvolle segmenten bestaat (zoals woorden). Ook al krijgt het model geen informatie over het bestaan van woorden, de aanwezigheid van woorden kon wel uit de resulterende representaties gedecodeerd worden. Dit experiment laat zien dat het model betekenisvolle segmenten in zinnen leert coderen zonder expliciete supervisie.

In **Hoofdstuk 5** onderzocht ik wat het spraakmodel leert over zinssemantiek. Omdat hier nog geen beschikbare spraakdata voor bestond, heb ik deze evaluatiedata verzameld. De resultaten lieten zien dat de representaties van het spraakmodel, net als eerder aangetoond voor het tekstmodel, correleren met menselijke intuïtie over zinsbetekenis. In reactie op de recente trend in de datawetenschap om modellen te verbeteren door telkens weer grotere databases te gebruiken, heb ik onderzocht welke database eigenschappen, behalve de hoeveelheid data, nog meer belangrijk zijn voor het maken van goede semantische representaties. De resultaten toonden aan dat modellen getraind op meer beschrijvingen per plaatje beter presteren, ook al is het totale aantal beschrijvingen gelijk (en het aantal plaatjes dus minder). Dit laat zien dat parafrasen een belangrijk leersignaal voor het model vormen.

De experimenten in **Hoofdstuk 6** toonden aan dat het spraakmodel woorden leert herkennen, en dat de woordherkenning wordt beïnvloed door twee bekende woordcompetitie-effecten die ook bij mensen de woordherkenning beïnvloeden. Verder toonde ik aan dat de woordrepresentaties onderscheid maken tussen enkel- en meervoudige zelfstandige naamwoorden. Het model leert het visuele verschil en laat bijvoorbeeld bij het horen van ‘hond’ plaatjes met een enkele hond zien, en bij het horen van ‘honden’ plaatjes met meerdere honden.

Ten slotte heb ik in **Hoofdstuk 7** de taalverwerkingsmechanismen van de RNN en de recent geïntroduceerde Transformer vergeleken. Ik heb taalmodellen gebaseerd op beide architecturen gebruikt om de leestijden en hersenpotentialen van mensen tijdens het lezen van zinnen te voorspellen, en heb aangetoond dat de Transformer dit het beste doet. Ondanks dat op het eerste gezicht het taalverwerkingsmechanisme van de Transformer niet cognitief plausibel lijkt, betoog ik dat het juist goed past in de ‘cue-based’ theorie van het menselijk werkgeheugen (Parker et al., 2017).

Concluderend heb ik in dit proefschrift een pleidooi gehouden voor cognitief plausibele methoden om taalrepresentaties te creëren, door beter te kijken naar hoe de mens taal leert. Modellen die taal en visuele informatie combineren zetten een stap in de goede richting door 1) net als mensen te leren van spraak, en niet van tekst, 2) het model bloot te stellen aan meer informatie dan slechts taal, omdat bij mensen alle zintuigen bijdragen aan onze mentale representaties en 3) het leren van zinnen en woorden in een geïntegreerd proces in plaats van losse opeenvolgende processen. Ten slotte toont het feit dat het model getraind is op relatief weinig data en toch vergelijkbaar presteert met modellen die meer data nodig hebben dan een mens ooit kan verwerken aan dat het leren van taal draait om meer dan slechts kwantiteit. Mensen hebben een rijke belevingswereld en we moeten verder kijken dan slechts tekstdata om meer te leren over taal.

Summary

While nearly all cognitive models of language processing assume that we have mental representations of words, surprisingly few actually deal with where they come from. Often, these representations are assumed given, for instance as pretrained word embeddings. While these representations have proven useful, they are not cognitively plausible as creating high-quality embeddings requires more data than any human can ever digest and is often dependent on text. Children can learn language from relatively little linguistic exposure (compared to our computational models) and learn to communicate long before they are able to read. A central idea in this dissertation is that humans are able to do so because all our sensory experiences are involved in learning language, most prominently our visual experience. The goal of this dissertation was to create linguistic representations in a more cognitively plausible way, using a model that combines visual and linguistic information. I aimed to create high-quality representations without any prior information, from cognitively plausible amounts and types of data, the latter meaning that in the second half of the dissertation I focus on learning directly from spoken input instead of text.

In **Chapter 2** I introduced a model that learns to create sentence representations from images with corresponding captions. I investigated whether the sentence representations captured aspects of semantic sentence similarity and showed that they correlate well with human semantic similarity judgements. The model learns meaningful sentence representations without requiring pretrained word embeddings and thus without treating word learning as a separate and prior process, which I argue is a more plausible order for learning language.

In **Chapter 3**, experiments with human behavioural data showed that the model can be used to create word representations that reflect cognitive aspects of word meaning. More importantly, the experiments showed that by combining linguistic and visual information, the model's representations explain a unique portion of variance in the human behavioural data even after accounting for text-based embeddings pretrained on billions of tokens of text. This chapter shows that we cannot create representations that fully capture human word knowledge if we ignore the wider range of human sensory experience.

I presented the speech-based model in **Chapter 4**. An important difference with the text-based model is that the speech-based model embeds utterances without explicit clues as to its constituent units (e.g., words). A probing experiment showed that word presence can be decoded from the sentence embeddings. Even though the model has no idea that words even exist, it does encode word

presence, showing that visual information can guide the model in finding meaningful constituents in utterances.

In **Chapter 5**, I investigated whether the speech-based model learns to capture sentence semantics. As no suitable evaluation data for this purpose existed, I collected spoken evaluation data. The results showed that the speech-based model indeed learns to capture cognitive aspects of sentence similarity. Responding to the trend in deep learning research to use ever larger databases to improve models, I investigated whether there is more to creating good-quality semantic embeddings besides database size. I showed that models with more descriptions per image perform better, even though the total number of descriptions is the same (and the total number of *images* thus lower), showing that paraphrases contain an important learning signal for the model.

In **Chapter 6** I showed that the model learns to recognise words and that word recognition performance is influenced by word competition from the word-initial cohort and neighbourhood density, two competition effects known to influence human word recognition. In an experiment, I showed that model encodes information which allows it to differentiate between singular and plural nouns, albeit not perfectly, showing that the model encodes a meaningful visual distinction between the two.

Lastly, in **Chapter 7** I investigated the processing mechanisms involved in the RNN and the recently introduced Transformer. I analysed language models based on both architectures and the results showed that the Transformer best predicts human sentence processing effort as measured by reading times and brain potentials. Even though, at face value, the Transformer's processing mechanism seems a cognitively implausible analogy to working memory, I argued that it is actually quite compatible with cue-based theories of working memory retrieval (Parker et al., 2017).

Overall, in this dissertation I argued that in order to create better linguistic representations for cognitive models, we need models that are informed by how humans learn language. Models combining visual and linguistic information take a step in this direction by 1) learning from speech rather than text, as humans do, 2) considering the wider range of sensory experiences humans have beyond linguistic exposure and 3) treating sentence and word learning as a single end-to-end process rather than isolated consecutive processes. Furthermore, the fact that the model presented in this study learns from relatively little linguistic exposure and is still competitive with, and complementary to, models trained on databases far larger than any human can read in their lifetime shows that learning language is about more than simply the quantity of linguistic exposure. The human experience is rich and varied, and we need to look beyond text data when trying to learn about language.

Acknowledgements

First of all, I would like to thank my supervisors Mirjam Ernestus and Stefan Frank. Your ideas and advice are why this dissertation has resulted in these exciting and interesting studies (allowing me to visit several great conferences). From curbing my tendency to implement now and think about possible research questions later to your helpful feedback on the papers, I learned a lot from you. Stefan's door was always open and right at the end of the hall, which resulted in many informal interruptions of your own work over the years. I am very grateful to both of you for your complete understanding when the last few months of my project were interrupted in the most unexpected way. I spent that time with my family without any worries about the fact that the dissertation was on hold.

Next, I would like to thank the manuscript committee, who read and evaluated this dissertation, and undoubtedly are preparing some difficult questions for the public defence. I look forward to seeing you at the defence and discussing these questions.

To the thesis students I supervised, thank you for taking an interest in the topics that I work on and delivering such interesting and high quality work. Special thanks to Sebastiaan, whose work led to a conference paper which we expanded into the journal paper that has just been accepted for publication.

As a part of the Language in Interaction (LiI) consortium, I got a lot of feedback on my projects through the presentations at our meetings. It was nice to regularly meet a large group of PhDs, all working on different aspects of language research. LiI provided a lot of opportunities to organise both social and educational events for the consortium PhDs. Similarly I would like to thank the International Max Plank Research school (IMPRS), which provided most of the education I followed during these four years. And of course thanks to the IMPRS coordinator Kevin for organising all the events and checking up on all of us and making sure we were getting by during the pandemic.

A special thanks to Odette, my former Master thesis supervisor. I had no particular ideas for my thesis project; knowing next to nothing about speech recognition and linguistic models, I responded to your project idea and you got me enthusiastic for the problem, and turned me into a fledgling computational linguist in little over a year. I also joined you as a research intern at the JSALT workshop you set up, which was a great opportunity to grow. The work we did there was the start of this four year project.

Bedankt Kristel, dat je begrip had voor de lange dagen die ik regelmatig heb gemaakt, me motiveerde als het tegenzat en altijd geloofd hebt dat het ging lukken. Lieve Fenna, die we pas na het afronden van mijn PhD hadden moeten ontmoeten maar er eerder was. Je kruipt ondertussen lekker het hele huis door, bedankt dat je zo'n blij meisje bent.

Bedankt pap, Jen en mam, het even heeft geduurd en het einde leek soms ver weg, maar uiteindelijk heb ik mijn studies afgerond en ben ik nu ook gepromoveerd. Op de universiteit heb ik eindelijk mijn draai gevonden, maar jullie hebben altijd al geloofd dat ik beter kon dan ik op school liet zien.

Biography

Danny Merkx was born on the 26th of May 1990 in ’s-Hertogenbosch, the Netherlands. He studied Public Administration and Artificial Intelligence at the Radboud University in Nijmegen. After obtaining both Bachelor degrees, he decided his main interest was in the cognitive science aspects of Artificial Intelligence and pursued a Master in Computation in Neural and Artificial Systems, obtaining his Master’s degree in 2017. He specialised in computational linguistics during his Master’s, culminating in a thesis on speech recognition and an internship as a research assistant at the Jelinek Summer Workshop on Speech and Language Technology at Carnegie Mellon University. During that internship, on speech recognition for unwritten languages, he worked with the multi-modal machine learning techniques that would become the basis for his PhD project. The PhD position at the Language in Interaction Consortium, with the aim of investigating the representations of the mental lexicon, was the perfect opportunity to further study this multi-modal learning approach. After finishing his PhD, he worked as a post-doc at the TU Delft. Danny is currently working as an actuarial data scientist at CZ health insurance.

Publications

- Danny Merkx, Sebastiaan Scholten, Stefan L. Frank, Mirjam Ernestus, and Odette Scharenborg. Modelling word learning and recognition using visually grounded speech. *Cognitive Computation*, forthcoming.

- Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–11. ACL, 2022.

- Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Semantic sentence similarity: size does not always matter. In *INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, pages 4393–4397, 2021.

- Danny Merkx and Stefan L. Frank. Human sentence processing: recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. ACL, 2021.

- Sebastiaan Scholten, Danny Merkx, and Odette Scharenborg. Learning to recognise words using visually grounded speech. In *Proceedings of the IEEE International Conference on Circuits and Systems*, pages 1–5. IEEE, 2021.

- Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang and Emmanuel Dupoux. Speech technology for unwritten languages. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 28:964–975, 2020.

- Danny Merkx, Stefan Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pages 1841–1845, 2019.

- Danny Merkx and Stefan L. Frank. Learning semantic sentence representations from visually grounded language without lexical knowledge. *Natural Language Engineering*, 25(4):451–466, 2019.

- Danny Merkx and Odette Scharenborg. Articulatory feature classification using convolutional neural networks. In *Proceedings of Interspeech 2018 - the 19th Annual Conference of the International Speech Communication Association*, pages 2142-2146, 2018.

- Odette Scharenborg and Danny Merkx. The role of articulatory feature representation quality in a computational model of human spoken-word recognition. In *Proceedings of the Machine Learning in Speech and Language Processing Workshop (MLSLP)*, pages 1-3, 2018.

- Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark A. Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang and Emmanuel Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "speaking rosetta" jsalt 2017 workshop. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4979–4983, 2018.

Research data management

This section gives an overview of the data used and created for the purpose of this dissertation.

Existing data

This section lists the existing databases I used, a url to where they can be found (last checked 09-03-2022) or, if a url is not available, describes how I obtained the data.

- Flickr8k: `https://forms.illinois.edu/sec/1713398`
- MSCOCO: `https://cocodataset.org/#download`
- Resnet-152: `https://pytorch.org/hub/pytorch_vision_resnet/`
- STS: `http://ixa2.si.ehu.eus/stswiki/index.php/Main_Page`
- SentEval: `https://github.com/facebookresearch/SentEval`
- Flickr8k audio captions: `https://groups.csail.mit.edu/sls/downloads/flickraudio/downloads.cgi`
- Multi-lingual bottleneck features: `https://github.com/lucasondel/multilingual-bottleneck-features`
- ENCOW: `https://www.webcorpora.org/register/`
- SUBTLEX-US: `https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus`
- EEG, Eye-tracking and reading time data from (Frank et al., 2013, 2015): requested from authors.
- Places205: `http://places.csail.mit.edu/`
- Places Audio: `https://groups.csail.mit.edu/sls/downloads/placesaudio/downloads.cgi`
- SpokenCOCO: `https://groups.csail.mit.edu/sls/downloads/placesaudio/downloads.cgi`
- Wordsim353: `https://gabrilovich.com/resources/data/wordsim353/wordsim353.html`

- SimLex999: `https://fh295.github.io/simlex.html`
- MEN: `https://aclweb.org/aclwiki/MEN_Test_Collection_(State_of_the_art)`
- RareWords: `https://nlp.stanford.edu/~lmthang/morphoNLM/`
- Semantic Priming Project: `https://www.montana.edu/attmemlab/spp.html`
- FastText vectors: `https://fasttext.cc/docs/en/pretrained-vectors.html`
- Google News Word2Vec vectors: `https://code.google.com/archive/p/word2vec/`
- Glove 840B vectors: `https://nlp.stanford.edu/projects/glove/`

New data

This section lists the new data I created, how I dealt with privacy issues and where the databases can be found. I collected a database called SpokenSTS for **Chapter 5** and a small collection of recorded nouns and verbs for the word recognition experiments in **Chapter 6**

Overview

(**SpokenSTS**) I collected spoken data from 4 native English speakers. These speakers were found through our network (colleagues and exchange interns). This data-set consists of approximately 750 sentence pairs (1500 sentences) taken at random (but balanced across STS subsets) from the Semantic Textual Similarity database which are pronounced by the participants and recorded at the CLS lab. The original recordings comprise of 11 recordings (1 per recording session 2-3 recording sessions were required per participant) stored in the .wav format. I have approximately 3 hours of recordings per participant. This is critical data concerning the recorded voices of participants. We do not store other information about the participants except for the consent forms. An assistant who annotated the raw data had secure access to the raw data.

I also created Text-to-Speech (TTS) versions of all the written STS sentences using Google TTS in 6 different voices.

(**Word recognition stimuli**) I collected spoken data from two native English speakers. Finding such speakers locally was difficult due to the Covid situation at the time, so I used my personal network to recruit two acquaintances in the US who were willing to record the script.

The script concerned 100 nouns and 150 verbs. These come from Flickr8k, selected as the most frequent words in their category. Recording the script takes about 10 minutes. Participants were asked to record the script twice, as getting back to them to re-record a few errors takes more time than simply having a back-up recording. This resulted in four 10-minute recordings stored in .wav format. This is critical data concerning the recorded voices of participants. I did not store other information about the participants except for the consent forms.

I furthermore had an existing collection of images annotated for occurrence of these words. The images to be annotated came from the Flickr8k database. This data has no privacy risks or ethics involved. The assistants involved in annotating this data had no access to the speech recordings.

Privacy

I list the types of data and the privacy concerns involved and then discuss how I ensured that I did not collect more data than needed and protected the participants privacy as much as possible.

Data types Raw data: this data is critical and is the biggest potential privacy risk. The scripts consisted of sentences and words that were read from a script and are not associated with the participants feelings, opinions or preferences in any way, but something they said to the experimenter during recording (e.g., ‘I feel tired today’) might. This data should be kept safe and encrypted at all times and does not need to be provided to others (i.e., our study can be reproduced without the raw data). I did not require any questionnaires or information from the participants other than on the consent forms.

Processed data: (**SpokenSTS**) segmented sentences (750 pairs × 4 participants = 6000 audio files) and information on which audio files form a pair and a link to their ground truth semantic similarity rating. This data is also critical. The segmented sentences will be made available to other researchers as a dataset for evaluating sentence embedding models.

(**Word recognition stimuli**) segmented words (250 words × 2 participants = 500 audio files). This data is also critical. The segmented words will be available to other researchers as a dataset for evaluating multimodal speech recognition models. The collected image annotations consist of two Excel files, one for verbs one for nouns. The files contain a row for each image and a column for each word. The cells contain a marker if the word occurred in the image. This is standard data and is not a privacy risk.

These audio files were further processed into MFCC features for the purpose of modelling. These processed features need not be permanently stored and made available (creation of said features is perfectly reproducible given parameter settings and common/free software modules) and forms no privacy risk over and above the audio data.

Analysed data: (**SpokenSTS**) the analysed data consists of the outcome of statistical tests which indicate the performance of the sentence embeddings models such as correlation coefficients. (Word recognition stimuli) The analysed data consisted of the outcome of statistical tests which indicate the performance of the models such as recognition accuracy. These data need not be permanently stored and made available (this can be perfectly reproduced using common/free software modules). This is standard data.

Dealing with privacy concerns I required recorded sentences (**SpokenSTS**) and words (**Word recognition stimuli**) which do not reflect participants' feelings, opinions or personal information. I dealt with these privacy concerns by cleaning the raw data of anything but contents of the scripts such that the processed data no longer contains information about the participant other than their voice. The associated papers had to mention that the participants are adult native speakers of English and balanced with 2 male and 2 female speakers (**SpokenSTS**) and 1 male and 1 female speaker (**Word recognition stimuli**) to establish the properties of the data (but I did not indicate which speakers are male and female).

Given that the data concerns recordings of the participants voice, I cannot make the data more anonymous than by processing it into segmented audio files which are not linked to their name or traits nor contain their personal opinions or feelings. Distorting the speech to make the voices unrecognisable would render the data unusable for our purposes. So, inherently, the data cannot be fully anonymous in the sense that their voice can be recognised. Participants were informed of the privacy risks, such as that people who know them might recognise their voice. The risk of third parties identifying the speaker or their traits using speech technology is unlikely unless the speakers disclose recordings of their voice in conjunction with personal information elsewhere. Participants were informed that identification of the speakers will not be supported by me and is less likely to happen if they prevent sharing of recordings of their voice in conjunction with personal information to a third party.

I did not collect any other information such as behaviour during recording or questionnaires. I did not need to make notes during the recording that could indicate the participants identity. The only other data collected were informed consent forms, which are obviously not shared and only securely stored at Radboud University for the purpose of scientific integrity.

Informed consent

I required informed consent from the participants. These forms are securely stored at Radboud University and informed the participants about the following:

Participants are fully informed about our research goals. Participants are informed that there are no foreseeable safety or discomfort risks but may still notify us of any discomfort without risking their compensation. Participants are informed that the data will be used in articles, and may be presented to others. Participants are informed that the data concerns audio, but we will make it as anonymous as possibly by not storing addition personal information in conjunction. Participants are informed that the data will be shared for replication or further development of our embedding models. Furthermore participants are made aware of what we see as the biggest privacy risk, namely speaker identification. We inform them how they themselves can help mitigate this risk if they so choose to. Participants are told they participate on a voluntary basis and may retract participation and their data at any point during or up until 1 week after making the recordings. Participants receive compensation for their participation at a rate of 10 euros in gift cards an hour. Participants are given the necessary information to contact me with questions about the study. Participants are told they can contact the lab manager with complaints.

Ethics

(**SpokenSTS**) Concerning ethics, this study needed approval by the ethics committee. Ethics approval could not be given as the recording started before the approval process. This error has been discussed by the committee and they decided this was an incidental slip that merits no further consequences.
(**Word recognition stimuli**) Recordings were not made in the CLS lab and participants were not recruited through SONA, so according to the rules of the Humanities Ethics Committee, ethics approval was not required.

Data security

(**SpokenSTS**) During all phases of the data collection and processing, the critical data resided on Radboud University servers that make regular backups and are protected by a user account and strong password.

(**Word recognition stimuli**) Due to the Covid situation at the time, participants recorded the script remotely in a setting of their own choosing and on their own hardware. This means that I had no control over the security of the data on their platform. To aid the participants in keeping their data safe, I provided them a way to transfer the data through a secure and encrypted cloud storage and informed them as soon as I had stored the data on the secure Radboud University servers so that they could delete the files from their own hardware. Whether they have done so cannot be confirmed.

During all further phases of the data processing, the critical data resided on Radboud University servers that make regular backups and are protected by a user account and strong password.

Data sharing

(**SpokenSTS**) The processed speech recordings and the TTS data were published as open access data in wav and flac format through DANS: `https://doi.org/10.17026/dans-z48-3ev6`

The shared data also includes instructions given to the annotator and speakers, the recording script and descriptions of our collection and processing methods and data structure. All data is also securely stored at Radboud University for the purpose of scientific integrity.

Annotations and raw data are not shared and only securely stored at Radboud University for the purpose of scientific integrity.

(**Word recognition stimuli**) The processed speech recordings (in wav and flac format) and image annotations were published as open access data through DANS (https://doi.org/10.17026/dans-22n-xh47).

The shared data also includes instructions given to the speakers, instructions to the image annotators, the recording script and descriptions of our collection and processing methods and data structure. All data is also securely stored at Radboud University for the purpose of scientific integrity.

Raw speech data are not shared and only securely stored at Radboud University for the purpose of scientific integrity.

Trained models

In all chapters of this dissertation, I trained PyTorch based deep learning models. These models have not been made publicly available, as I share the code that can be used to reproduce all experiments. All trained models used for this dissertation are securely stored at Radboud University for the purpose of scientific integrity and will be shared upon request.

Code

All the code produced for this dissertation has been made publicly available. All code is also securely stored at Radboud University for the purpose of scientific integrity. Per chapter, there is a branch in the github repository containing the code as it was at the time of submitting the associated paper. The links to these repositories in order of the chapters are:

- `https://github.com/DannyMerkx/caption2image`
- `https://github.com/DannyMerkx/speech2image/tree/CMCL2022`
- `https://github.com/DannyMerkx/speech2image/tree/Interspeech19`
- `https://github.com/DannyMerkx/speech2image/tree/Interspeech21`
- `https://github.com/DannyMerkx/speech2image/tree/CogComp2022`
- `https://github.com/DannyMerkx/next_word_prediction`

Papers

All papers associated with the chapters in this dissertation were published open access and I provide links to the pdfs here. The LaTeX source code of all papers are securely stored at the Radboud University.

- Learning semantic sentence representations from visually grounded language without lexical knowledge. `https://repository.ubn.ru.nl/handle/2066/205977`

- Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge. `https://aclanthology.org/2022.cmcl-1.1.pdf`

- Language learning using speech to image retrieval. `https://repository.ubn.ru.nl/handle/2066/208191`

- Semantic sentence similarity: size does not always matter. `https://repository.ubn.ru.nl/handle/2066/235108`

- Modelling human word learning and recognition using visually grounded speech. `https://arxiv.org/abs/2203.06937`

- Human sentence processing: recurrence or attention? `https://repository.ubn.ru.nl/handle/2066/235107`