*Natural Language Engineering* (2019), **25**, pp. 451–466
doi:10.1017/S1351324919000196

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

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

Danny Merkx* and Stefan L. Frank

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
*Corresponding author. Emails: d.merkx@let.ru.nl; s.frank@let.ru.nl

**Abstract**

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 multimodal 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 datasets: Microsoft Common Objects in Context (MSCOCO) and Flickr8k. We evaluate the semantic content of the resulting sentence embeddings using the data from the Semantic Textual Similarity (STS) benchmark task and show that the multimodal 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 multimodal 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.

**Keywords:** representation learning; semantic representations; multi-modal learning; semantic relatedness

## 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, Dumais, Furnas *et al.* 1990). Rubenstein and Goodenough (1965) 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, Chen, Corrado *et al.* 2013) and GloVe (Pennington, Socher, and Manning 2014) have shown meaningful clusters and correlations with human similarity judgements (De Deyne, Perfors, and Navarro 2017), and have become widely used features that boost performance in several natural language processing (NLP) tasks such as machine translation (Qi, Sachan, Felix *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, Zhu, Salakhutdinov *et al.* 2015), FastSent (Hill, Cho, and Korhonen 2016) and Paragraph-Vector (Le and Mikolov 2014), which are related to the Skip-Gram word model by Mikolov *et al.* (2013). Recently, there have also been successful sentence encoder models which

are trained on a supervised task and then transferred to other tasks (e.g. Conneau, Kiela, Schwenk *et al*. 2017; Yang, Yuan, Cer *et al*. 2018; Kiela, Conneau, Jabri *et al*. 2018).

So far, existing sentence embedding methods often require (pretrained) word embeddings (Conneau, Kiela, Schwenk *et al*. 2017; Kiela *et al*. 2018), large amounts of data (Hill *et al*. 2016) or both (Boom, Canneyt, Bohez *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, Behrens, Speares *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 multimodal data has proven successful on the level of word embeddings (see for instance Collell, Zhang, and Moens 2017; Derby, Miller, Murphy *et al*. 2018). For sentence semantics, the multimodal 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, Harwath, and Glass 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 (STS), 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 STS. This is to the best of our knowledge the first evaluation of the sentence level semantics of a multimodal 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.

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

We release the code of our preprocessing pipeline, models and evaluation on github as open source: https://github.com/DannyMerkx/caption2image.

## 2. Sentence embeddings

### 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 regard to transfer learning and semantic content. Conneau *et al.* (2017) trained a recurrent neural network (RNN) sentence encoder on the Stanford Natural Language Inference database (Bowman, Angeli, Potts *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 STS judgements and perform well on various transfer tasks like sentiment analysis and subjectivity/objectivity detection.

### 2.2 Multimodal methods

Image-caption retrieval is a multimodal 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, Lu, Shang *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. Kiela, Conneau, Jabri *et al.* (2015) converted the captions to Fisher vectors (Jaakkola and Haussler 1999) 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. Then they 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 a 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, Kiros, Fidler *et al.* 2016; Faghri, Fleet, Kiros *et al.* (2017); Wehrmann, Mattjie, and Barros 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

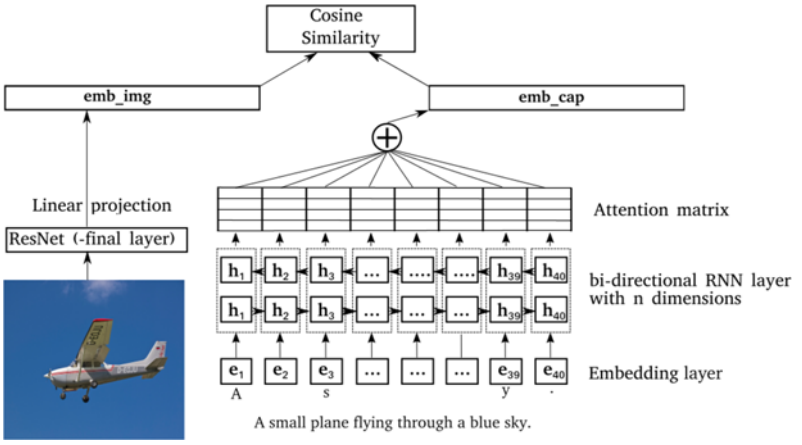


**Figure 1.** Model architecture: The model consists of two branches with the image encoder on the left and the caption encoder on the right. The character embeddings are denoted by $\mathbf{e}_t$ and the RNN hidden states by $\mathbf{h}_t$. Each hidden state has *n* features which are concatenated for the forward and backward RNN into 2*n* 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**).

a model similar to Dong, Li, and Snoek (2018), that is, an RNN 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 the 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 multimodal training data. In contrast, Kiela *et al.* (2018) augmented their grounded representations by combining them with non-grounded (Skip-Thought) representations. Finally, we probe the semantic content of our sentence representations more directly by evaluating the caption encoder on the STS 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.

## 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.

### *3.1 Encoder architectures*

#### *3.1.1 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 1.

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

The image features are extracted by a pretrained image recognition model trained on ImageNet (Deng, Dong, Socher *et al.* 2009). For this we used ResNet-152 (He, Zhang, Ren *et al.* 2016), a residualised network with 152 layers from which we take the activations of the penultimate fully connected layer[a]. 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.

### *3.1.2 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:

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

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 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 Chung, Gulcehre, Cho *et al.* 2014 and Greff, Srivastava, Koutník *et al.* 2017 for detailed descriptions of these RNNs). The GRU is a recurrent layer that is widely used in sequence modelling (e.g. Bahdanau, Cho, and Bengio 2015; Zhu, Kiros, Zemel *et al.* 2015; Patel, Pimpale, and Sasikumar 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.

[a]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, Wang, and Specia (2018) document that the features of the penultimate layer yield better transfer learning results than the object classification layer.

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

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 $\alpha$. The hinge loss $L$ as a function of the network parameters $\theta$ 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 $(cap, img) \neq (cap', img')$.

### 3.2 Training data

The multimodal embedding approach requires paired captions and images for which we use two popular image-caption retrieval benchmark datasets: Flickr8k (Hodosh, Young, and Hockenmaier 2013) and MSCOCO (Chen, Fang, Lin *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 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 $224 \times 224$ 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.

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 at the end of each caption.

### 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

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:

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

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, Li, Pleiss *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 $10^{-4}$.

### *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 STS analysis, SentEval includes the STS and Sentences Involving Compositional Knowledge (SICK) datasets which we briefly review here. After training our multimodal encoder network, we simply discard the image encoder, and the caption encoder is used to encode the test sentences in SentEval.

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, Banea, Cardie *et al.* 2015, p. 254, see also for a full description of the annotator instructions). The evaluation performed on the STS 2012–2016 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, Diab, Agirre *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 1 gives an overview of the datasets. For full descriptions of each dataset, see Agirre *et al.* (2012; 2013; 2014; 2015; 2016).

SICK is a database created for a shared task at SemEval-2014 with the purpose of testing compositional distributional semantics models (Bentivogli, Bernardi, Marelli *et al.* 2016). The dataset consists of 10,000 sentence pairs which were generated using sentences taken from Flickr8k and the STS 2012 MSRvid dataset. 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

**Table 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, Van Gool, Williams *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 | |

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 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% of the data as training set, 5% as development set and 50% as test set.

## 4. Results and discussion

### *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.

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

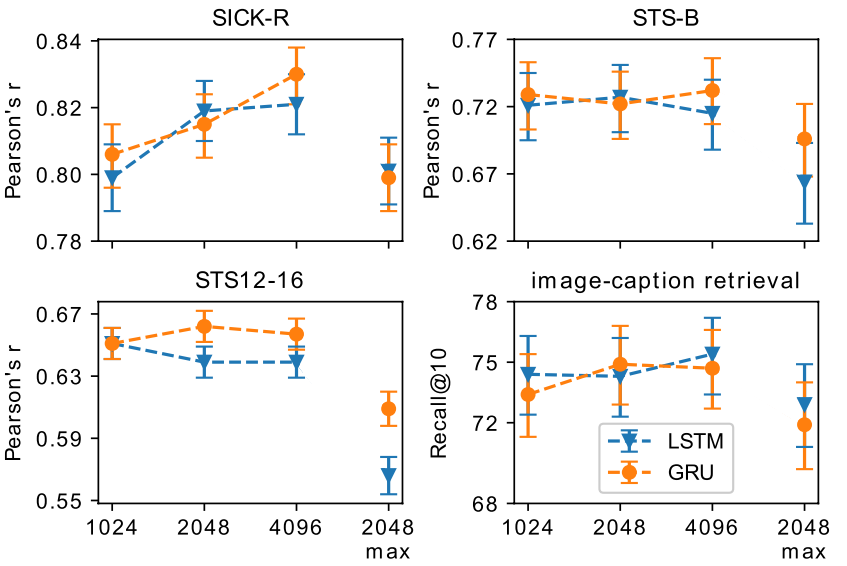


**Figure 2.** Model performance on the semantic (SICK-R, STS-B and STS12-16) and training task (image-caption retrieval) measures including the 95% confidence interval. Training task performance is measured in recall@10. The semantic performance measure is Pearson's *r*. The horizontal axis shows the embedding size with 'max' indicating the max pooling model.

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 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.

### *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 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

**Table 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% 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 |

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

### *4.3 Semantic evaluation*

We now look at the semantic properties of the sentence embeddings in more detail and compare our models with the previous work. Figure 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.

#### *4.3.1 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

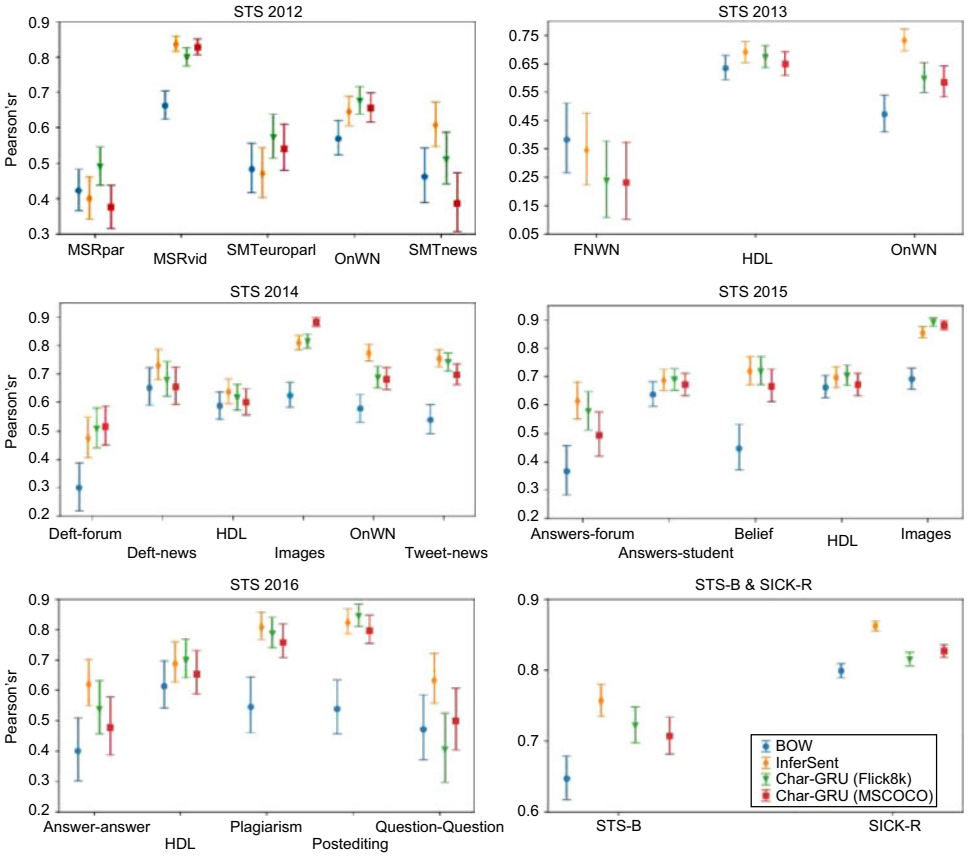


**Figure 3.** Semantic evaluation task results: Pearson correlation coefficients with their 95% confidence interval for the various subtasks (see Table 1). BOW is a bag of words approach using GloVe embeddings and InferSent is the model reported by Conneau *et al*. (2017). A supplement with a table of the results shown here is included in the github repository.

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 the Flickr8k database is only about 5% of the size of MSCOCO and about 1% of what InferSent is trained on.

### *4.3.2 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 3 shows that the BOW model performs

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

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 multimodal 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 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.

#### *4.3.3 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 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 towards an unrepresentative context when learning the embeddings. The MSRpar dataset contains many proper nouns for which no embedding might exist, and it is a 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 FrameNet-WordNet (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 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.

#### *4.3.4 Trade-off between training task and transfer task performance*

We further investigate how prone our model is to overspecialising on image descriptions. Figure 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

**Table 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<br>We are proud often here to represent the citizens of Europe |
| MSRpar | 4.0 | South Asia follows, with 1.1 million youths infected – 62% of them are female<br>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<br>Lacking in specific resources, qualities or substances |
| Question–Question | 4.0 | How do I make a height adjustable desk?<br>How can I build a wall mounted adjustable height desk? |

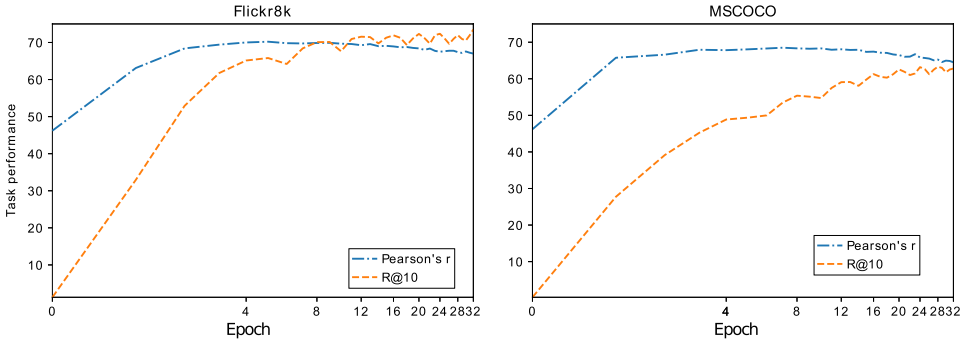


**Figure 4.** The training task performance (R@10) and the semantic task performance (Pearson's $r \times 100$) as they develop over training, with the number of epochs on a logarithmic scale. For MSCOCO (right) we show the training task performance on the 5,000 item test set.

and semantic task performance happens in the first four epochs. After that the training task performance still increases by 12.8% and 28.5% 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% 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 poses a challenge to the search for universal sentence embeddings.

## 5. Conclusion

We investigated whether sentence semantics can be captured in sentence embeddings without using (prior) lexical knowledge. We did this using a multimodal encoder which grounds language in vision using image-caption pairs. Harwath and Glass (2015) have claimed that this method produces a multimodal 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

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

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 3 are hard to capture in image). Future work could combine the visual grounding approach with text-only methods in order to learn from more diverse data. In such a multitask 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, Torralba, and Glass 2016; Chrupała, Gelderloos, and Alishahi 2017), we have barely scratched the surface of transfer learning using spoken input.

**Author ORCIDs.** Danny Merkx, 0000-0002-5829-5214; Stefan Frank, 0000-0002-7026-711X

**Acknowledgements.** The research presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. We would like to thank Mirjam Ernestus for commenting on an earlier version of this paper.

## References

**Agirre E., Banea C., Cardie C., Cer D., Diab M., Gonzalez-Agirre A., Guo W., Lopez-Gazpio I., Maritxalar M., Mihalcea R., Rigau G., Uria L. and Wiebe J.** (2015). SemEval-2015 task 2: Semantic textual similarity, English, Spanish and Pilot on interpretability. In *SemEval*. Denver, Colorado: ACL.

**Agirre E., Banea C., Cardie C., Cer D., Diab M., Gonzalez-Agirre A., Guo W., Mihalcea R., Rigau G. and Wiebe J.** (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval*. Dublin, Ireland: ACL.

**Agirre E., Banea C., Cer D., Diab M., Gonzalez-Agirre A., Mihalcea R., Rigau G. and Wiebe J.** (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval*. San Diego, California: ACL.

**Agirre E., Cer D., Diab M., Gonzalez-Agirre A. and Guo W.** (2013). *SEM 2013 shared task: Semantic textual similarity. In *SemEval*. Atlanta, Georgia: ACL.

**Agirre E., Diab M., Cer D. and Gonzalez-Agirre A.** (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval*. Montréal, Canada: ACL.

**Bahdanau D., Cho K. and Bengio Y.** (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*. San Diego, California: ICLR. pp. 1–15.

**Bentivogli L., Bernardi R., Marelli M., Menini S., Baroni M. and Zamparelli R.** (2016). SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *LRE* **50**(1).

**Boom C.D., Canneyt S.V., Bohez S., Demeester T. and Dhoedt B.** (2015). Learning semantic similarity for very short texts. In *ICDMW*. Atlantic City, New Jersey: IEEE.

**Bowman S.R., Angeli G., Potts C. and Manning C.D.** (2015). A large annotated corpus for learning natural language inference. In *EMNLP*. Lisbon, Portugal: ACL.

**Braine M.D.S. and Bowerman M.** (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development* **41**(1).

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

**Cer D., Diab M., Agirre E.E., Lopez-Gazpio I. and Specia L.** (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *SemEval*. Vancouver, Canada.

**Chen X., Fang H., Lin T.-Y., Vedantam R., Gupta S., Dollar P. and Zitnick C. L.** (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. 1–7. arXiv:1504.00325.

**Chrupała G., Gelderloos L. and Alishahi A.** (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the ACL*. Vancouver, Canada: ACL, pp. 613–622

**Chung J., Gulcehre C., Cho K. and Bengio Y.** (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Paper presented at the NIPS Workshop on Deep Learning*. Montreal, Canada: NIPS, pp. 1–9.

**Collell G., Zhang T. and Moens M.-F.** (2017). Imagined visual representations as multimodal embeddings. In *AAAI*. San Francisco, California: Association for the Advancement of Artificial Intelligence.

**Conneau A. and Kiela D.** (2018). SentEval: An evaluation toolkit for universal Sentence representations. In *LREC*. Miyazaki, Japan: LREC. pp. 1699–1704.

**Conneau A., Kiela D., Schwenk H., Barrault L. and Bordes A.** (2017). Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*. Copenhagen, Denmark: ACL.

**De Deyne S., Perfors A. and Navarro D.J.** (2017). Predicting human similarity judgments with distributional models: The value of word associations. In *ICJAI*. Melbourne, Australia: AAAI Press, pp. 4806–4810

**Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R.** (1990). Indexing by latent semantic analysis. *Journal of the Association for Information Science* **41**(6).

**Deng J., Dong W., Socher R., Li L., Li K. and Fei-Fei L.** (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*. Miami, Florida: IEEE.

**Derby S., Miller P., Murphy B. and Devereux B.** (2018). Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In *CoNLL*. Brussels, Belgium: ACL.

**Dong J., Li X. and Snoek C.G.M.** (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* **20**(12).

**Everingham M., Van Gool L., Williams C.K.I., Winn J. and Zisserman A.** (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results.

**Faghri F., Fleet D.J., Kiros R. and Fidler S.** (2017). VSE++: improved visual-semantic embeddings. 1–13. CoRR abs/1707.05612.

**Greff K., Srivastava R.K., Koutník J., Steunebrink B.R. and Schmidhuber J.** (2017). LSTM: A search space odyssey. *Transactions on Neural Networks and Learning Systems* **28**(10).

**Harwath D. and Glass J.** (2015). Deep multimodal semantic embeddings for speech and images. In *ASRU*. Scottsdale, Arizona: IEEE.

**Harwath D., Torralba A. and Glass J.** (2016). Unsupervised learning of spoken language with visual context. In *NIPS*.

**He K., Zhang X., Ren S. and Sun J.** (2016). Deep residual learning for image recognition. In *CVPR*. Las Vegas, Nevada: IEEE. pp. 770–778.

**Hill F., Cho K. and Korhonen A.** (2016). Learning distributed representations of sentences from unlabelled data. In *NAACL HLT*. San Diego, California: ACL.

**Hochreiter S. and Schmidhuber J.** (1997). Long short-term memory. *Neural Computation* **9**(8).

**Hodosh M., Young P. and Hockenmaier J.** (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* **47**(1).

**Huang G., Li Y., Pleiss G., Liu Z., Hopcroft J.E. and Weinberger K.Q.** (2017). Snapshot ensembles: Train 1, get M for free. In *ICLR*. Vancouver, Canada: ICLR. pp. 1–14.

**Jaakkola T. and Haussler D.** (1999). Exploiting generative models in discriminative classifiers. In *NIPS*.

**Karpathy A. and Fei-Fei L.** (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*. Boston, MA: IEEE. pp. 664–676.

**Kiela D., Conneau A., Jabri A. and Nickel M.** (2018). Learning visually grounded sentence representations. In *NAACL-HLT*. New Orleans, Louisiana: New Orleans, Louisiana: ACL.

**Kingma D.P. and Ba J.** (2015). Adam: A method for stochastic optimization. In *ICLR*. San Diego, California: ICLR. pp. 1–15.

**Kiros R., Zhu Y., Salakhutdinov R.R., Zemel R., Urtasun R., Torralba A. and Fidler S.** (2015). Skip-thought vectors. In *NIPS*.

**Klein B., Lev G., Sadeh G. and Wolf L.** (2015). Associating neural word embeddings with deep image representations using Fisher vectors. In *CVPR*. Boston, MA: IEEE.

**Le Q. and Mikolov T.** (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, vol. 32. Beijing, China: PMLR.

**Leidal K., Harwath D. and Glass J.** (2017). Learning modality-invariant representations for speech and images. In *ASRU*. Okinawa, Japan: IEEE.

**Lieven E., Behrens H., Speares J. and Tomasello M.** (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language* **30**(2).

**Ma L., Lu Z., Shang L. and Li H.** (2015). Multimodal convolutional neural networks for matching image and sentence. In *ICCV*. Santiago, Chile: IEEE.

https://doi.org/10.1017/S1351324919000196 Published online by Cambridge University Press

**Madhyastha P., Wang J. and Specia L.** (2018). The role of image representations in vision to language tasks. *NLE* **24**(3).
**Mikolov T., Chen K., Corrado G. and Dean J.** (2013). Efficient Estimation of Word Representations in Vector Space. 1–12. arXiv:1301.3781.
**Patel R.N., Pimpale P.B. and Sasikumar M.** (2016). Recurrent Neural Network based Part-of-Speech Tagger for Code-Mixed Social Media Text. 1–7. arXiv:1611.04989.
**Pennington J., Socher R. and Manning C.** (2014). Glove: Global vectors for word representation. In *EMNLP*. Doha, Qatar: ACL.
**Pine J.M. and Lieven E.** (1993). Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech. *Journal of Child Language* **20**.
**Qi Y., Sachan D.S., Felix M., Padmanabhan S.J. and Neubig G.** (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *NAACL-HLT*. New Orleans, Louisiana: ACL.
**Rubenstein H. and Goodenough J.B.** (1965). Contextual correlates of synonymy. *Communications of the ACM* **8**(10).
**Smith L.N.** (2017). Cyclical learning rates for training neural networks. In *WACV*. Santa Rosa, California: IEEE.
**Tomasello M.** (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* **11**(1/2).
**Vendrov I., Kiros R., Fidler S. and Urtasun R.** (2016). Order-embeddings of images and language. In *ICLR*. San Juan, Puerto Rico: ICLR. pp. 1–12.
**Wehrmann J., Mattjie A. and Barros R.C.** (2018). Order embeddings and character-level convolutions for multimodal alignment. *Pattern Recognition Letters* **102**.
**Yang Y., Yuan S., Cer D., Kong S.-y., Constant N., Pilar P., Ge H., Sung Y.-H., Strope B. and Kurzweil R.** (2018). Learning semantic textual similarity from conversations. In *RepL4NLP*. Melbourne, Australia: ACL.
**Zhu Y., Kiros R., Zemel R., Salakhutdinov R., Urtasun R., Torralba A. and Fidler S.** (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*. Santiago, Chile: IEEE. pp. 19–27.

**Cite this article:** Merkx D and Frank SL. Learning semantic sentence representations from visually grounded language without lexical knowledge. *Natural Language Engineering* **25**, 451–466. https://doi.org10.1017/S1351324919000196