SANDI: Story-and-Images Alignment

Sreyasi Nag Chowdhury **Simon Razniewski** **Gerhard Weikum**
Max Planck Institute for Informatics, Saarbrücken, Germany
(sreyasi, srazniew, weikum)@mpi-inf.mpg.de

Abstract

The Internet contains a multitude of social media posts and other stories where text is interspersed with images. In these contexts, images are not simply used for general illustration, but are judiciously placed in certain spots of a story for multimodal descriptions and narration. In this work we analyze the problem of text-image alignment, and present SANDI, a methodology for automatically selecting images from an image collection and aligning them with text paragraphs of a story. SANDI combines visual tags, user-provided tags and background knowledge, and uses an Integer Linear Program to compute alignments that are semantically meaningful. Experiments show that SANDI can select and align images with texts with high quality of semantic fit.

1 Introduction

It is well-known (and supported by studies Lester (2013); Messaris and Abraham (2001)) that the most powerful messages are delivered with a combination of words and pictures. On the Internet, such multimodal content is abundant in the form of news articles, social media posts, and personal blog posts where authors enrich their stories with carefully chosen and placed images. As an example, consider a vacation report, to be posted on a blog site or online community. The backbone of the travel report is a textual narration, but the user typically places illustrative images in appropriate spots, carefully selected from her photo collection from this trip. These images can either show specific objective highlights such as waterfalls, mountain hikes or animal encounters, or may serve to depict the thematic mood of the trip, e.g., by showing nice sunsets. Another example is brochures for various organizations. Here, the text describes the mission, achievements and ongoing projects, and is accompanied with judiciously selected and placed photos

Figure 1: The story-and-images alignment problem.

of buildings, people, products and images depicting the subjects of interest, e.g., galaxies or telescopes for research in astrophysics.

The generation of such multimodal stories requires substantial human judgement and reasoning, and is thus time-consuming and labor-intensive. In particular, the effort on the human side includes selecting the right images from a pool of story-specific photos (e.g., the traveler's own photos) and possibly also from a broader pool for visual illustration (e.g., from Pinterest). Even if the set of photos were exactly given, there is still considerable effort to place them within appropriate paragraphs, paying attention to the semantic coherence between surrounding text and image. In this paper, we set out to automate this human task, formalizing it as a *Story-AND-Images (SANDI) alignment* problem.

Problem Statement. Given a story-like text document and a set of images, the problem is to automatically decide where individual images are placed in the text. Figure 1 depicts this task. The problem comes in different variants: either all images in the given set need to be placed, or a subset of given cardinality must be selected and aligned with text paragraphs. Formally, given n paragraphs and $m \geq n$ images, assign $b \leq n$ of these images to a subset of the paragraphs, such that each paragraph has at most one image.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 989–999
April 19 - 23, 2021.

Prior Work and its Inadequacy. There is ample literature on computer support for multimodal content creation, most notably, on generating image captions. Closest to our problem is work on *Story Illustration* (Joshi et al., 2006; Schwarz et al., 2010) where the task is to select illustrative images from a large pool. However, the task is quite different from ours, making prior approaches inadequate for the setting of this paper. First, unlike story illustration, we need to consider the text-image alignments jointly for all pieces of a story, rather than making context-free choices one piece at a time. Second, prior work assumes that each image in the pool has an informative caption or set of tags, by which the selection algorithm computes its choices. Our model harnesses visual tags from deep neural network based object-detection frameworks and incorporates background knowledge, as automatic steps to enrich the semantic interpretation of images.

Our Approach – SANDI. We present a framework that casts the story-images alignment task into a combinatorial optimization problem. The objective function, to be maximized, captures the semantic coherence between each paragraph and the co-located image. To this end, we consider a suite of features – the visual tags associated with an image (automatically detected tags as well as user-defined tags when available), text embeddings, and also background knowledge. The optimization is constrained by the number of images that the story should be enriched with. As a solution algorithm, we devise an integer linear program (ILP) and employ the Gurobi ILP solver for computing the exact optimum. Experiments show that SANDI produces semantically coherent alignments. A demonstration of SANDI (Nag Chowdhury et al., 2020) can be viewed at `https://youtu.be/k5gu2pNxdNU`.

Contributions. To the best of our knowledge, this is the first work to address story-images alignment. Our salient contributions are:

1. We introduce and define the problem of story-images alignment.
2. We analyze two real-world datasets of stories with rich visual illustrations, and derive insights on alignment decisions and quality measures.
3. We devise relevant features, formalize the alignment task as a combinatorial optimization problem, and develop an exact-solution algorithm using integer linear programming.
4. We compare our method against baselines that use multimodal embeddings.

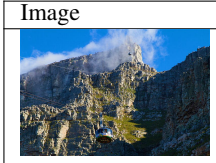
2 Related Work

Existing work on associations between text and images can be categorized into the following areas.

Image Attribute Recognition. High level concepts in images lead to better results in Vision-to-Language problems (Wu et al., 2016). Traditionally image tagging was based on community input (Gupta et al., 2010). Modern deep-learning based tools detect objects (Hoffman et al., 2014; Redmon and Farhadi, 2017; Ren et al., 2015) and scenes (Zhou et al., 2014) in images. Inter-concept incoherence can also be refined using background knowledge (Nag Chowdhury et al., 2018). We leverage some frameworks from this category in our model to detect visual concepts in images.

Story Illustration. Prior work finds suitable images from annotated image collections to illustrate personal stories (Joshi et al., 2006; Ravi et al., 2018) or news posts (Schwarz et al., 2010; Delgado et al., 2010). The results are presented as clusters of related images (Guan et al., 2011), or an illustrated article (Jhamtani et al., 2016). Story illustration only addresses the problem of image selection, whereas we solve two problems simultaneously: image selection and image placement – making a joint decision on all pieces of long complex stories. This makes our problem distinct. There is no way to systematically compare our full-blown model with prior works on story illustration alone.

Multimodal Embeddings. A popular method of semantically comparing images and text has been to map textual and visual features into a common space of multimodal embeddings (Frome et al., 2013; Vendrov et al., 2016; Faghri et al., 2018; Wu et al., 2019; Wang et al., 2019; Liu et al., 2019). Visual-Semantic-Embeddings (VSE) has been used for generating captions for whole images (Faghri et al., 2018), or to associate text with image regions (Karpathy and Li, 2015). Color, geometry, aspect-ratio have been used to align image regions to nouns ("chair"), attributes ("big"), and pronouns ("it") in corresponding text (Kong et al., 2014). Recent work train on document-level co-occurrences and predict links between images and sentences in a document (Hessel et al., 2019; Chu and Kao, 2017). However, alignment of small image regions to text snippets or linking images to single sentences play little role in jointly interpreting the correlation between images and a larger body of text. We focus on the latter in this work.

Image	Ground Truth Paragraph
	. . . Table Mountain Cableway. The revolving car provides 360 degree views as you ascend this mesmerising 60-million-year-old mountain. From the upper cableway station. . .
	. . . On the east flank of the hill is the old Muslim quarter of the Bo-Kaap; have your camera ready to capture images of the photogenic pastel-painted colonial period homes. . .

(a) Sample image and paragraph from Lonely Planet

Image	Ground Truth Paragraph
	. . . If you are just looking for some peace and quiet or hanging out with other students...library on campus, a student hangout space in the International College building. . . .
	. . . I was scared to travel alone. But I quickly realized that there's no need to be afraid. Leaving home and getting out of your comfort zone is an important part of growing up. . . .

(b) Sample image and paragraph from Asia Exchange

Figure 2: Text-image pairs from our datasets.

Image Caption Generation. Most prior works generate factual captions (Xu et al., 2015; Tan and Chan, 2016; Lu et al., 2017; Wang et al., 2018), while some recent architectures venture into producing stylized captions (Gan et al., 2017) and stories (Zhu et al., 2015; Krause et al., 2017). An image caption can be considered as a precise focused description of an image without much superfluous or contextual information. However, in a multimodal story, the paragraphs surrounding an image contain detailed thematic descriptions. We try to capture the thematic indirection between an image and surrounding text, thus making the problem distinct.

Commonsense Knowledge for Story Understanding. One of the earliest applications of Commonsense Knowledge (CSK) to interpret text-image associtions is a photo agent which automatically annotated images from user's multi-modal (text-image) emails or web pages, while also inferring additional CSK concepts (Lieberman and Liu, 2002). Subsequent works used CSK reasoning to infer causality in stories (Williams et al., 2017). We enhance automatically detected objects and scenes in image with relevant CSK concepts from ConceptNet (Speer et al., 2017). This often helps to capture more context about an image.

3 Dataset and Problem Analysis

3.1 Dataset

To the best of our knowledge, there is no experimental dataset for text-image alignment. We therefore compile two datasets from two blogging sites:

- Lonely Planet[1]: 2178 travel stories containing 20 paragraphs and 4.5 images per story. Most images are accompanied by captions. The images come from the author's personal archives and adhere strictly to the content of the story.
- Asia Exchange[2]: 200 stories about education opportunities in Asia, with 13.5 paragraphs and 4 images per story. Some stories contain generic stock images complying with the abstract theme. Most images have captions.

Figure 2 shows image-paragraph examples from the datasets.

Text-Image Semantic Coherence. To understand human judgments behind text-image pairing, we analyze 50 randomly chosen images and their corresponding paragraphs from the Lonely Planet dataset. We identify six possibly overlapping concept classes that appear in images as well as in their corresponding paragraphs: (i) natural named objects such as Mt. Everest (ii) human activities such as biking (iii) generic objects such as cars (iv) general nature scenes such as forest (v) specific man-made entities such as monuments (vi) geographic locations such as Rome. The outcome of this analysis is shown in Table 1.

Concept Classes	% of text-images pairs with shared concepts
Natural named objects	9%
Human activities	12%
Generic objects	15%
General nature scenes	20%
Man-made named objects	21%
Geographic locations	29%

Table 1: Reasons for text-image semantic coherence.

3.2 Image Tags

Based on the analysis in Table 1, we consider the following kinds of tags for describing images:

Visual Tags (CV). We use three state-of-the-art computer-vision methods for object and scene detection. First, deep convolutional neural networks based architectures like LSDA (Hoffman et al., 2014) and YOLO (Redmon and Farhadi, 2017), are used to detect objects like *person*, *frisbee* or *bench*, that denote "Generic objects" from Table 1.

[1] www.lonelyplanet.com/blog

[2] www.asiaexchange.org

For stories, general scene descriptors like *restaurant* or *beach* play a major role, too. Therefore, our second asset is scene detection from the MIT Scenes Database (Zhou et al., 2014). These constitute "General nature scenes" from Table 1. Thirdly, since stories often abstract away from explicit visual concepts, a framework that incorporates abstractions into visual detections – VISIR (Nag Chowdhury et al., 2018) – is also leveraged. For e.g., the concept "hiking" is supplemented with the concepts "walking" (Hypernym of "hiking" from WordNet) and "fun" (from ConceptNet (Speer et al., 2017) assertion ⟨ *hiking, has property, fun*⟩).

User Tags (MAN). Owners of images often have additional knowledge about content and context – for e.g., activities or geographical information ("hiking near Lake Placid"), which, from Table 1 play a major role in text-image alignment. For experiments, we use nouns and adjectives from image captions from our datasets as user tags. In downstream applications, images can be selected either from web repositories or from a personal collection. In the former case, explicit tags or words from captions/titles serve as user tags. In the latter case, location details like names of places can be easily inferred from metadata like GPS coordinates associated with "raw" phone/camera images.

Big-data Tags (BD). Big data and crowd knowledge allow to infer additional context that may not be visually apparent. We utilize the Google reverse image search API[3] to incorporate such tags. This API allows to search by image, and suggests tags based on visually similar images in the vast web image repository. These tags depict popular places, such as "Savarmati Ashram", or "Mexico City insect market", and thus constitute "Natural names objects", "Man-made named objects", as well as "Geographic locations" from Table 1.

To further improve the semantic characterization of an image, we extend the tag set of an image by related commonsense knowledge concepts.

Commonsense Knowledge (CSK). CSK can bridge the gap between visual and textual concepts (Nag Chowdhury et al., 2016). CV, BD, and MAN tags are enriched with CSK from the following ConceptNet relations – *used for, has property, causes, at location, located near, conceptually related to*. E.g., for the left image in Figure 3, we add CSK concept "show talent" from CV tag "stage" from the assertion ⟨stage, *used for*, show talent⟩. CSK concepts cover multiple classes from Table 1. Owing to the noise and subjectivity in ConceptNet, only concepts which are informative for a given image are retained. If the top-10 web search results of a CSK concept are semantically similar to the image tags (CV/MAN/BD), the CSK concept is considered to be *informative* for the image. Cosine similarity between the mean vectors (from word2vec) of the image context and the search results is used as a measure of semantic similarity.

Figure 3 shows examples of the image tags. In use cases all features are not always available – user tags may not exist or may not be retained during web distribution, big data requires access to paid APIs , and visual tags are error-prone. We will thus study the features in isolation and jointly.

CV: person, sunglasses, stage
MAN: Globe Theatre
BD: Shakespeare's Globe
CSK: show talent, attend concert, entertain audience

CV: terra cotta, village
MAN:tiled rooftops
BD: Languedoc Roussillon
CSK: colony, small town

Figure 3: Types of image tags. CV – visual objects/scenes, MAN and BD – nuanced descriptions, locations, CSK - high-level thematic concepts.

4 Model for Story-Images Alignment

Our *story-image alignment* model constitutes an Integer Linear Program (ILP) which jointly optimizes the placement of selected images within a story. The main ingredient for this alignment is the pairwise similarity between images and units of text. We consider a paragraph as a text unit.

Text-Image Pairwise Similarity. Given an image, each of the three kinds of descriptors of Section 3.2 gives rise to a bag of features. We use these features to compute *text-image semantic relatedness scores* $srel(i, t)$ for an image i and a paragraph t.

$$srel(i, t) = cosine(\vec{i}, \vec{t}) \quad (1)$$

where $\vec{i}$ and $\vec{t}$ are the mean word embeddings for the image tags and the paragraph respectively. For images, we use all detected tags. For paragraphs, we consider only the top 50% of concepts w.r.t. their TF-IDF ranking over the entire dataset. We use word embeddings from word2vec trained on Google News Corpus. $srel(i, t)$ scores from Equation 1 serve as weights for variables in the ILP.

[3] www.google.com/searchbyimage

Tasks. Our problem has two distinct tasks: 1. *Image Selection* – to select relevant images from an image pool. 2. *Image Placement* – to place selected images in the story. These two components are modelled into one ILP where Image Placement is achieved by maximizing an objective function, while Image Selection is styled by constraints. In the following subsections we discuss two flavors of our model consisting of one or both the tasks.

4.1 Complete Alignment

Complete Alignment constitutes the problem of aligning *all* images in a given image collection with relevant text units of a story. Hence, only *Image Placement* is applicable. For a story with $|T|$ text units and an associated image album with $|I|$ images, the alignment of images $i \in I$ to text units $t \in T$ can be modeled as an Integer Linear Program (ILP) with the following definitions:

Decision Variables. The following binary decision variables are introduced: $X_{it} = 1$ if image i should be aligned with text unit t, 0 otherwise.

Objective. Select image i to be aligned with text unit t such that the semantic relatedness over all text-image pairs is maximized:

$$max\left[\sum_{i \in I}\sum_{t \in T} srel(i,t)X_{it}\right] \quad (2)$$

where $srel(i,t)$ is the text-image semantic relatedness from Equation 1.

Constraints. We make two assumptions for text-image alignments – no image may be repeated in the story (Constraint 3), and no paragraph may be aligned with multiple images (Constraint 4). The former is a trivial observation from multimodal presentations on the web such as in blog posts, newswire, brochures. The latter is made based on the nature of our datasets, and it is designed as a hard constraint in order to facilitate a fair evaluation.

$$\sum_{i} X_{it} \leq 1 \forall t \quad (3) \qquad \sum_{t} X_{it} = 1 \forall i \quad (4)$$

4.2 Selective Alignment

Selective Alignment is the flavor of the model which selects a certain number of thematically relevant images from a big image pool, and places them within the story. Hence, it constitutes both tasks – *Image Selection* and *Image Placement*. Along with the constraint in (3), Image Selection entails the following additional constraints:

$$\sum_{t} X_{it} \leq 1 \forall i \quad (5) \qquad \sum_{i}\sum_{t} X_{it} = b \quad (6)$$

where b is the budget for the number of images in the story. b may be simply defined as the number of paragraphs in the story, following our assumption that each paragraph may be associated with a maximum of one image. (5) is an adjustment to (4) which implies that not all images from the image pool need to be aligned with the story. The objective function from (2) rewards the selection of best fitting images from the image pool.

5 Quality Measures

In this section we define metrics for automatic evaluation of the text-image alignment problem. The two tasks involved – *Image Selection* and *Image Placement* – call for separate evaluation metrics as discussed below.

5.1 Image Selection

Representative images for a story are selected from a big pool of images. There are multiple conceptually similar images in our image pool since they have been gathered from blogs of the domain "travel". Hence evaluating the results on strict precision (based on exact matches between selected and ground-truth images) does not necessarily assess true quality. We therefore define a relaxed precision metric (based on semantic similarity) in addition to the strict metric. Given a set of selected images I and the set of ground truth images J, where $|I| = |J|$, the precision metrics are:

$$RelaxedPrecision = \frac{\sum_{i \in I, j \in J} max(cosine(\vec{i},\vec{j}))}{|I|} \quad (7)$$

$$StrictPrecision = \frac{|I \cap J|}{|I|} \quad (8)$$

5.2 Image Placement

For each image in a multimodal story, the ground truth (GT) paragraph is assumed to be the one following the image in our datasets. To evaluate the quality of SANDI's text-image alignments, we compare the GT paragraph and the paragraph assigned to the image by SANDI (henceforth referred to as "aligned paragraph"). We propose the following metrics for evaluating the quality of alignments:

BLEU and ROUGE. BLEU and ROUGE are classic n-gram-overlap-based metrics for evaluating

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve
RandomAlign	3.1	6.9	75.1	50.0	50.0
VSE++	11.0	9.5	84.6	59.1	55.2
VSE++ ILP	12.6	11.2	84.0	58.1	48.0
SANDI-CV	18.2	17.6	86.3	63.7	54.5
SANDI-MAN	**45.6**	**44.5**	**89.8**	72.5	**77.4**
SANDI-BD	26.6	25.1	84.7	61.3	61.2
SANDI∗	44.3	42.9	89.7	**73.2**	76.3

Table 2: Complete Alignment: Lonely Planet.

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve
RandomAlign	6.8	8.9	70.8	50.0	50.0
VSE++	19.4	17.7	85.7	51.9	48.0
VSE++ ILP	23.5	20.1	86.0	52.6	46.1
SANDI-CV	21.5	20.6	87.8	58.4	52.0
SANDI-MAN	**35.2**	**32.2**	89.2	61.5	61.5
SANDI-BD	24.1	22.3	86.7	56.0	53.6
SANDI∗	33.4	31.5	**89.7**	**62.4**	**62.5**

Table 3: Complete Alignment: Asia Exchange.

machine translation and text summarization. Although known to be limited insofar as they do not recognize synonyms and semantically equivalent formulations, they are in widespread use. We consider them as basic measures of concept overlap between GT and aligned paragraphs.

Semantic Similarity. To alleviate the shortcoming of requiring exact matches, we consider a metric based on embedding similarity. We compute the similarity between two text units t_i and t_j by the average similarity of their word embeddings, considering all unigrams and bigrams as words.

$$SemSim(t_i, t_j) = cosine(\vec{t_i}, \vec{t_j}) \tag{9}$$

where $\vec{x}$ is the mean vector of words in x. For this calculation, we drop uninformative words by keeping only the top 50% with regard to their TF-IDF weights over the whole dataset.

Average Rank of Aligned Paragraph. We associate each paragraph in the story with a ranked list of all the paragraphs on the basis of semantic similarity (Eq. 9), where rank 1 is the paragraph itself. Our goal is to produce alignments ranked higher with the GT paragraph. The average rank of alignments by a model is computed as follows:

$$ParaRank = 1 - \left[\left(\frac{\sum_{t \in T'} rank(t)}{|I|} - 1\right) \Big/ \Big(|T| - 1\Big)\right] \tag{10}$$

where $|I|$ is the number of images and $|T|$ is the number of paragraphs in the story. $T' \subset T$ is the set of paragraphs aligned to images. Scores are normalized between 0 and 1; 1 being the perfect alignment and 0 being the worst alignment.

Order Preservation. Most stories either follow a time-line or storyline. Images placed at meaningful spots within the text would ideally adhere to this sequence. Hence the measure of pairwise ordering provides a sense of maintaining or respecting the storyline. It can be defined as the number of order preserving image pairs in the alignment (i_m, i_n) normalized by the total number of ordered image pairs in the ground truth.

$$OrderPreserve = \frac{|(i_m, i_n)|}{(|I|(|I|-1)/2)} \tag{11}$$

Correlation between Metrics. Table 4 shows the pairwise correlation between our evaluation metrics – BLEU, ROUGE, *SemSim*, *ParaRank*, *Orderpreserve* – computed on a random alignment of images to paragraphs in 100 stories. Not surprisingly, BLEU and ROUGE correlate nearly perfectly, and show reasonable correlation to $SemSim$ and $ParaRank$. $OrderPreserve$ works at a different level and exhibits virtually no correlation to the other metrics. This illustrates that order-preserving alignments are not necessarily semantically meaningful, and vice versa.

	ROUGE	SemSim	ParaRank	Order Preserve
BLEU	0.98	0.32	0.39	-0.23
ROUGE		0.33	0.40	-0.23
SemSim			0.29	0.08
ParaRank				-0.06

Table 4: Correlation between evaluation metrics.

6 Experiments and Results

We evaluate the two flavors of SANDI – *Complete Alignment* and *Selective Alignment* – based on the quality measures from Section 5.

6.1 Setup

Tools. Deep learning based architectures – LSDA (Hoffman et al., 2014), YOLO (Redmon and Farhadi, 2017), VISIR (Nag Chowdhury et al., 2018) and Places-CNN (Zhou et al., 2014) are used as sources of *Visual tags (CV)*. Google reverse image search tag suggestions are used as *Big-data tags (BD)*. We use the Gurobi Optimizer for solving the ILP. A Word2Vec (Mikolov et al., 2013) model trained on the Google News Corpus encompasses a large cross-section of domains, and hence is used as a source of word embeddings.

SANDI Variants. The variants of our text-image alignment model are based on the use of image descriptors described in Section 3.2.

- SANDI-CV, SANDI-MAN, and SANDI-BD use CV, MAN, and BD tags respectively.
- SANDI∗ combines CV, MAN, and BD tags.

6.2 Complete Alignment

We evaluate our Complete Alignment model (defined in Section 4.1), which places *all* images from a given image pool within a story.

Baselines. To the best of our knowledge, there is no existing work on story-image alignment. Hence we modify methods on joint visual-semantic-embeddings (VSE) (Kiros et al., 2014; Faghri et al., 2018) to serve as baselines, henceforth referred to as VSE++. We compare SANDI with:

- RandomAlign: a simple baseline with random image-text alignments.
- VSE++: for an image, VSE++ is adapted to produce a ranked list of paragraphs from the given story. The top paragraph is considered as an alignment, with a greedy constraint that one paragraph can be aligned to at most one image.
- VSE++ ILP: using cosine similarity scores between image and paragraph from the joint visual-semantic embedding space, we solve an ILP as described in Section 4.

Since there are no existing story-image alignment datasets, VSE++ has been trained on the MSCOCO captions dataset (Lin et al., 2014), which contains 330K images with 5 captions per image.

Evaluation. Tables 2 and 3 show the performance of the baselines and the SANDI variants on the Lonely Planet and Asia Exchange datasets respectively. SANDI outperforms the baselines on all evaluation metrics to various degrees. While VSE++ looks at each image in isolation, SANDI captures context better by considering all text units of the story and all images from the corresponding album at once in a constrained optimization problem. VSE++ ILP, although closer to SANDI in methodology, does not outperform SANDI. This can be attributed to the fact that SANDI is less tied to a particular dataset, relying only on word2vec embeddings that are trained on a much larger corpus than MSCOCO. On Lonely Planet, SANDI-MAN is the best configuration – this is expected since user tags (MAN) contain concepts most specific to the story. SANDI∗ marginally outperforms it on Asia Exchange – recall that images in this dataset are sometimes generic thematic illustrations, hence a combination of all features capture more context. The consistency of scores across both datasets highlight the robustness of SANDI.

Role of Commonsense Knowledge (CSK). We observe that CSK helps improve performance of SANDI-CV. This is intuitive because CV tags denote only explicit objects and scenes, which do not capture high-level concepts of the images. CSK alleviates this to some extent. For example, in the first image in Figure 3 – CSK (*show talent, attend concert, entertain audience*) appends a more meaningful context to the CV tags (*person, sunglasses, stage*); MAN and BD tags already capture a broader context. Table 5 compares SANDI-CV and SANDI-CV-CSK on Asia Exchange.

	BLEU	ROUGE	SemSim	ParaRank	Order Preserve
SANDI-CV	21.5	20.6	87.8	58.4	52.0
SANDI-CV -CSK	19.9	19.4	**88.3**	**62.4**	50.0

Table 5: Role of Commonsense Knowledge: Asia Exchange

6.3 Selective Alignment

This variation of our model, as defined in Section 4.2, solves two tasks – *Image Selection* and *Image placement*.

6.3.1 Image Selection

Setup. In addition to the setup described in Section 6.1, some additional requirements are:

- Image pool – We pool images from stories in our dataset. Since stories from a particular domain (e.g. travel blogs) are largely quite similar, images in the pool may also be very similar in content – e.g., stories on *hiking* contain images containing *mountain, person, backpack*.
- Image budget – For each story, the number of images in the ground truth is considered as the image budget b (Equation 4.2).

Baselines. We compare SANDI with:

- RandomAlign: a baseline of randomly selected images from the pool.
- NN: nearest neighbors from a common embedding space of images and paragraphs. Images are represented as mean vectors of their tags, and paragraphs are represented as mean vectors of their distinctive words. The basic word vectors are obtained from Word2Vec trained on Google News Corpus.

Tag Space	Precision	Random	NN	VSE++	SANDI
CV	*Strict*	0.4	2.0	1.14	**4.18**
	Relaxed	42.16	52.68	29.83	**53.54**
MAN	*Strict*	0.4	3.95	-	**14.57**
	Relaxed	37.14	42.73		**49.65**
BD	*Strict*	0.4	1.75	-	**2.71**
	Relaxed	32.59	37.94		**38.86**
$*$	*Strict*	0.4	4.8	-	**11.28**
	relaxed	43.84	50.06		**54.34**

Table 6: Selective Alignment-Image Selection:Lonely Planet.

Tag Space	Precision	Random	NN	VSE++	SANDI
CV	*Strict*	0.45	0.65	0.44	**0.79**
	Relaxed	55.0	**57.64**	30.05	57.2
MAN	*Strict*	0.45	0.78	-	**3.42**
	Relaxed	40.24	52.0		**52.87**
BD	*Strict*	0.45	0.82	-	**0.87**
	Relaxed	31.12	**33.27**		33.25
$*$	*Strict*	0.45	1.04	-	**1.7**
	relaxed	55.68	58.1		**58.2**

Table 7: Selective Alignment-Image Selection:Asia Exchange.

	BLEU	ROUGE	SemSim	ParaRank
RandomAlign	0.31	0.26	69.18	48.16
VSE++	1.04	0.8	79.18	53.09
VSE++ ILP	1.23	1.03	79.04	53.96
SANDI-CV	1.70	1.60	83.76	61.69
SANDI-MAN	**8.82**	**7.40**	82.95	66.83
SANDI-BD	1.77	1.69	**84.66**	**76.18**
SANDI$*$	6.82	6.57	84.50	75.84

Table 8: Selective Alignment-Image Placement:Lonely Planet.

	BLEU	ROUGE	SemSim	ParaRank
RandomAlign	2.06	1.37	53.14	58.28
VSE++	2.66	1.39	58.00	64.34
VSE++ ILP	2.78	1.47	57.65	64.29
SANDI-CV	1.04	1.51	60.28	75.42
SANDI-MAN	**3.49**	**2.98**	61.11	**82.00**
SANDI-BD	1.68	1.52	**76.86**	70.41
SANDI$*$	1.53	1.84	64.76	80.57

Table 9: Selective Alignment-Image Placement:Asia Exchange.

- VSE++: a joint visual-textual embeddings method presented in (Faghri et al., 2018) is adapted to retrieve the top-b images for a story.

Evaluation. We evaluate *Image Selection* by the measures in Section 5.1. Table 6 shows the results for SANDI and the baselines on a pool of 500 images from Lonely Planet. NN and SANDI both use Word2Vec for text-image similarity. SANDI's better scores are attributed to the joint optimization over the entire story, as opposed to greedy selection in case of NN. VSE++ uses a joint text-image embeddings space for similarity scores. Our evaluation metric $RelaxedPrecision$ (Eq. 7) factors in the semantic similarity between images based on the image descriptors (Section 3.2). Hence we compute results on the different image tag spaces, where '$*$' refers to the combination of CV, MAN, and BD. The baseline VSE++ however, operates only on visual features; hence we report its performance only for CV. Results on Asia Exchange are similar (Table 7). Recall from Section 3.1 that the Asia Exchange dataset often has stock images for generic illustration rather than only story-specific images. Hence the average relaxed precision on image selection is comparitively higher. Figure 4 shows image selection results for one story. The original story contains 17 paragraphs; only the main concepts from the story have been retained for readability. SANDI is able to retrieve 2 ground-truth (GT) images out of 7, while the baselines retrieve 1 each. Note that SANDI's non-exact matches are thematically similar to GT – images in the 4th column of both GT and SANDI feature a yellow train in a backdrop of mountains, images in the 5th column show sunset. This can be attributed to the wider space of concepts that SANDI explores through the image tags from Section 3.2.

6.3.2 Image Placement

Having selected thematically related images from a big image pool, SANDI places them within contextual paragraphs of the story. Note that SANDI integrates the *Image Selection* and *Image Placement* stages into joint inference on selective alignment, whereas the baselines operate in two steps.

We evaluate the alignments by the measures from Section 5.2. Note that the measure *OrderPreserve* does not apply to Selective Alignment since the images are selected from a pool of mixed images which cannot be ordered. From Table 8 and 9 we observe that SANDI outperforms the baselines by a clear margin, harnessing its more expressive pool of tags. We show anecdotal evidence of the diversity of our image tags in Figure 3 and Table 10.

6.4 Role of Model Components

Image Descriptors. The wide variety of image tags that SANDI leverages (CV, BD, MAN) capture special characteristics of the images. These are unavailable to baselines such as VSE++, attributing to their poor performance.

Embeddings. The nature of embeddings is decisive towards alignment quality. Joint visual-semantic-embeddings trained on MSCOCO (used by VSE++) fall short in capturing high-level se-

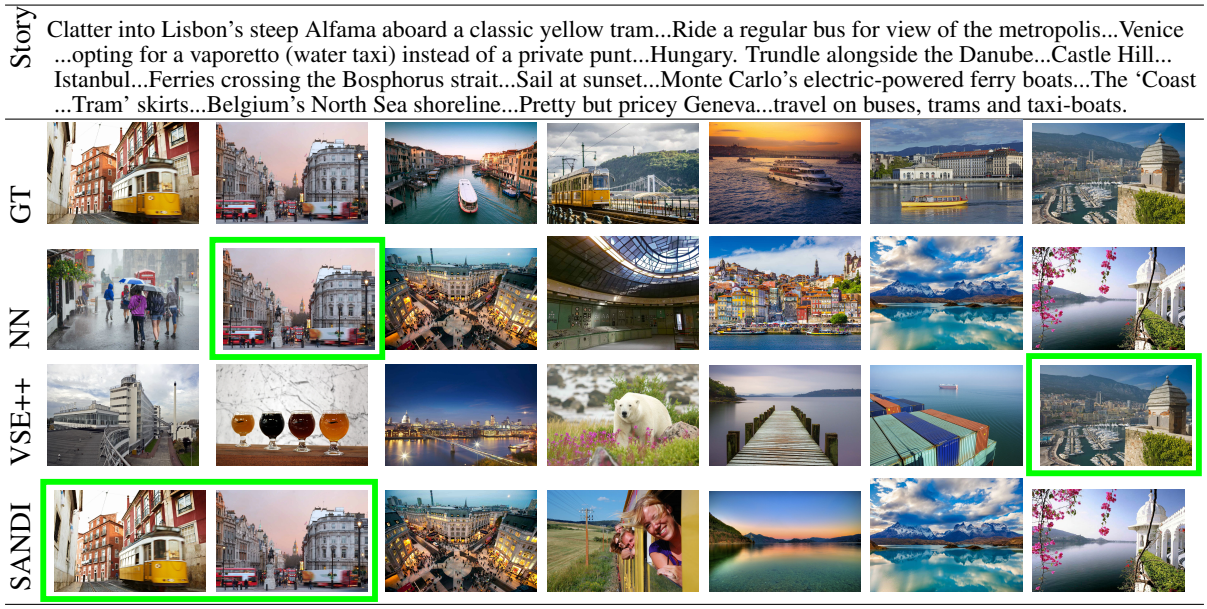


Figure 4: Image Selection. Images within green boxes are exact matches with ground truth (GT). SANDI retrieves more exact matches than the baselines (NN, VSE++). SANDI's non-exact matches are also much more thematically similar to the GT.

Image and detected concepts	SANDI-CV	SANDI-MAN	SANDI-BD
 CV: snowy mountains, massif, alpine glacier, mountain range MAN: outdoor lover, New Zealand, study destination, BD: New Zealand	New Zealand produced the first man to ever climb Mount Everest and also the creator of the bungee-jump. Thus, it comes as no surprise that this country is filled with adventures and adrenaline junkies.	Moreover, the wildlife in New Zealand is something to behold. Try and find a Kiwi! (The bird!) They are nocturnal creatures so it is quite a challenge. New Zealand is also home to the smallest dolphin species. Lastly, take the opportunity to search for the beautiful yellow-eyed penguin.	Home to hobbits, warriors, orcs and dragons. If you're a fan of the famous trilogies, Lord of the Rings and The Hobbit, then choosing New Zealand should be a no-brainer.

Table 10: Example alignments. Highlighted texts show similar concepts between image and aligned paragraphs.

mantics between images and story. Word2Vec embeddings trained on a much larger and domain-independent Google News corpus better represents high-level image-story interpretations.

ILP. As observed in Tables 6 and 7, Combinatorial optimization (SANDI) outperforms greedy optimization approaches (NN), both methods using the same embedding space.

7 Conclusion

In this paper we introduced the problem of story-images alignment – selecting and placing a set of representative images within a story. We analyzed features towards meaningful alignments from real-world multimodal datasets – Lonely Planet and Asia Exchange blogs – and defined various evaluation measures. We presented SANDI, a methodology for automating such alignments by a constrained optimization problem maximizing semantic coherence between text-image pairs jointly for the entire story. Evaluations show that SANDI produces semantically meaningful alignments. Nevertheless, some follow-up questions arise.

Additional Features. Our feature space covers most natural aspects. In addition, GPS locations where available may provide cues for geographic named entities, while timestamps may capture temporal aspects of a storyline.

Abstract and Metaphoric Relations. We do not address stylistic elements like metaphors and sarcasm in text, which would entail more challenging alignments. For example, the text "the news was a dagger to his heart" should not be paired with a picture of a dagger. Although user provided tags may provide some cues towards such abstract relationships, a deeper understanding of semantic coherence is desired.

The proposed text-image alignment system is available at https://sandi.mpi-inf.mpg.de, and a video of the demonstration can be viewed at https://youtu.be/k5gu2pNxdNU.

References

Wei-Ta Chu and Ming-Chih Kao. 2017. Blog article summarization with image-text alignment techniques. In *ISM*.

Diogo Delgado, João Magalhães, and Nuno Correia. 2010. Automated illustration of news stories. In *ICSC*.

Fartash Faghri, David J. Fleet, Jamie Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *CVPR*.

Genliang Guan, Zhiyong Wang, Xian-Sheng Hua, and David Dagan Feng. 2011. *StoryImaging*: a media-rich presentation system for textual stories. In *ACM MM*.

Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. *SIGKDD Explorations*.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP*.

Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: large scale detection through adaptation. In *NIPS*.

Harsh Jhamtani, Shubham Varma, Midhun Gundapuneni, and Siddhartha Kumar Dutta. 2016. A supervised approach for text illustration. In *ACM MM*.

Dhiraj Joshi, James Ze Wang, and Jia Li. 2006. The story picturing engine - a system for automatic text illustration. *TOMCCAP*.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *CVPR*.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*.

Paul Martin Lester. 2013. *Visual communication: Images with messages*. Cengage Learning.

Henry Lieberman and Hugo Liu. 2002. Adaptive linking between text and photos using common sense reasoning. In *AH*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM MM*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Paul Messaris and Linus Abraham. 2001. The role of images in framing news stories. In *Framing public life*. Routledge.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Sreyasi Nag Chowdhury, William Cheng, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. 2020. Illustrate your story: Enriching text with images. In *WSDM*.

Sreyasi Nag Chowdhury, Niket Tandon, Hakan Ferhatosmanoglu, and Gerhard Weikum. 2018. VISIR: visual and semantic image label refinement. In *WSDM*.

Sreyasi Nag Chowdhury, Niket Tandon, and Gerhard Weikum. 2016. Know2look: Commonsense knowledge for visual search. In *AKBC*.

Hareesh Ravi, Lezi Wang, Carlos Muñiz, Leonid Sigal, Dimitris N. Metaxas, and Mubbasir Kapadia. 2018. Show me a story: Towards coherent neural story illustration. In *CVPR*.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *CVPR*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*.

Katharina Schwarz, Pavel Rojtberg, Joachim Caspar, Iryna Gurevych, Michael Goesele, and Hendrik P. A. Lensch. 2010. Text-to-video: Story illustration from online photo collections. In *KES*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Ying Hua Tan and Chee Seng Chan. 2016. phi-lstm: A phrase-based hierarchical LSTM model for image captioning. In *ACCV*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *ICLR*.

Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching images and text with multi-modal tensor fusion and re-ranking. In *ACM MM*.

Ziwei Wang, Yadan Luo, Yang Li, Zi Huang, and Hongzhi Yin. 2018. Look deeper see richer: Depth-aware image paragraph captioning. In *ACM Multimedia*, pages 672–680. ACM.

Bryan Williams, Henry Lieberman, and Patrick H. Winston. 2017. Understanding stories with large-scale common sense. In *COMMONSENSE*.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *ACM MM*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *NIPS*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story like visual explanations by watching movies and reading books. In *ICCV*.