

# Illustrate Your Story: Enriching Text with Images

Sreyasi Nag Chowdhury<sup>\*</sup> William Cheng<sup>†</sup> Gerard de Melo<sup>†</sup> Simon Razniewski<sup>\*</sup> Gerhard Weikum<sup>\*</sup>

sreyasi, srazniew, weikum@mpi-inf.mpg.de

chengwill97@gmail.com

gdm@demelo.org

<sup>\*</sup> Max Planck Institute for Informatics, Saarbrücken

<sup>†</sup> Rutgers University, New Jersey

## ABSTRACT

Human perception is known to be predominantly visual. As modern web infrastructure promoted the storage of media, the web-data paradigm shifted from text-only documents to those containing text and images. A multitude of blog posts, news articles, and social media posts exist on the Internet today as examples of multimodal stories. The manual alignment of images and text in a story is time-consuming and labor intensive. We present a web application for automatically selecting relevant images from an album and placing them in suitable contexts within a body of text. The application solves a global optimization problem that maximizes the coherence of text paragraphs and image descriptors, and allows for exploring the underlying image descriptors and similarity metrics. Experiments show that our method can align images with texts with high semantic fit, and to user satisfaction.

### ACM Reference Format:

Sreyasi Nag Chowdhury, William Cheng, Gerard de Melo, Simon Razniewski, Gerhard Weikum 2020. Illustrate Your Story: Enriching Text with Images. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20)*, February 3–7, 2020, Houston, TX, USA. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3336191.3371866>

## 1 INTRODUCTION

**Motivation and Problem.** The role of visual communication has been a long-standing topic of scientific as well as philosophical discourse [2]. Studies establish that the most powerful and meaningful messages are delivered with a careful combination of words and pictures [10]. Not surprisingly, over the years, data on the Internet has become predominantly multi-modal, consisting of text punctuated with images [1]. Articles ranging from commercial websites, personal blog accounts, newswire, educational material, all contain a fair proportion of text and images.

The generation of such multimodal stories constitutes fine-grained human reasoning: selecting images relevant to the textual content, assigning meaningful tags and captions to the images, and placing them at suitable contextual locations within the text for a coherent

narration. Work on Image Selection [8] traditionally relied on manually annotated image repositories – e.g. of web images. Advances in Computer Vision have enabled automatic image tagging [12] and captioning [3] to a fair extent, enabling searching of images without prior manual annotations. However, aligning individual images to snippets in a larger body of text requires a deeper understanding of the underlying semantics.

This task – selection and placement of relevant images within a story – which is fairly simple for humans, offers a challenge for automated systems. The SANDI approach [4] relies on combinatorial optimization and computes exact solutions. We build an end-to-end system based on this model. We also additionally propose a functionality that ensures greater visual appeal of the generated multimodal content. An automated system that can successfully select and align images to text will be useful to a multitude of end users like journalists, bloggers, authors, and commercial enterprises.

**Related Work.** Interesting problems at the intersection of Computer Vision and Natural language Processing include image tagging [9, 12] and captioning [3], multimodal embeddings [5, 6], as well as story illustration [8, 11]. Most modern methods addressing such multimodal tasks leverage deep neural networks to learn a joint visual–textual embedding space. A multimodal story exhibits high-level indirections between images and aligned paragraphs – e.g., a paragraph about the environment may be suitably aligned with an image of a dying polar bear without explicitly mentioning “polar bear”. Joint visual semantic embeddings learned from a large dataset of multimodal stories may capture such indirect links. However, such a dataset is not yet available. SANDI [4] solves unsupervised story–image alignment using a combinatorial optimization approach that is robust, since the underlying model is not data-dependent and only relies on word embeddings trained on a large text repository. In this paper we build a tool based on this model. Our tool can be accessed at <https://sandi.mpi-inf.mpg.de>.

**Contributions.** We implement a working system for the Image Selection and Placement model proposed in [4]. To the best of our knowledge, this is the first implementation of the story–image alignment problem. Additionally:

- (1) We argue that spacing of images within a story is important for the visual appeal of the generated multimodal content, and propose a new model to that effect.
- (2) We also study the inherent subjectivity of the task in a suitable user study.

## 2 METHOD

The problem proposed by [4] consists in solving two distinct tasks – Image Selection and Image Placement. Their Integer Linear Program (ILP) based model makes a joint decision on image selection and

<sup>†</sup> William Cheng now works at Google, Mountain View.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WSDM '20, February 3–7, 2020, Houston, TX, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-6822-3/20/02...\$15.00

<https://doi.org/10.1145/3336191.3371866>

placement by optimizing the pairwise semantic similarity between images and text paragraphs. Given a story with  $|T|$  paragraphs and an image collection with  $|I|$  images, the ILP is defined as follows:

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

$$\max \left[ \sum_{i \in I} \sum_{t \in T} \text{srel}(i, t) X_{it} \right] \quad (1)$$

where  $\text{srel}(i, t) = \cos(\theta_{\vec{i}, \vec{t}})$ ,  $\vec{i}$  and  $\vec{t}$  being the mean vectors of image concepts and paragraph concepts, respectively, and  $\theta_{\vec{i}, \vec{t}}$  denoting the angle between them.  $X_{it}$  are binary decision variables, where  $X_{it} = 1$  if image  $i$  should be placed within paragraph  $t$ , 0 otherwise.

Two types of alignments are proposed in [4] – Complete Alignment and Selective Alignment. We propose a third type of alignment – Spacing-aware Alignment – that considers uniform image spacing as an additional criterion.

- **Complete Alignment** – placement of *all* images from an image collection within relevant paragraphs of a story. This constitutes the following constraints to the objective in (1):

$$\sum_i X_{it} \leq 1 \forall t \quad (2) \quad \sum_t X_{it} = 1 \forall i \quad (3)$$

Constraint (2) guarantees that no two images are aligned with the same paragraph. Constraint (3) ensures that no image is repeated in the story and each image from the collection is used.

- **Selective Alignment** – selection of a few relevant images from an image collection and placement of the selected images within the given story. Along with the constraint in (2), the following additional constraints apply:

$$\sum_t X_{it} \leq 1 \forall i \quad (4) \quad \sum_i \sum_t X_{it} = b \quad (5)$$

where  $b$  is the number of images that should be selected for the story. The constraint in (4) implies that not all images from the collection need to be placed within the story.

- **Spacing-aware Alignment** – We observe from the datasets used in [4] – Lonely Planet\* and Asia Echange† – that images are usually uniformly distributed throughout the story. For more visual appeal, we add placement constraints that spread images as evenly as possible across the story.

For a story with  $T$  paragraphs and  $I$  images, the number of paragraphs between successive images,  $m$ , is bounded as follows:

$$\left\lfloor \frac{T+1}{I+1} \right\rfloor \leq m \leq \left\lceil \frac{T-1}{I-1} \right\rceil \quad (6)$$

Additional constraints for the ILP restrict the distances between neighboring images:

$$\sum_{i=0}^{u-1} \sum_{s=0}^{l-1} X_{i,t+s} \geq 1 \forall t \quad (7) \quad \sum_{i=0}^{u-1} \sum_{s=0}^{l-1} X_{i,t+s} \leq 1 \forall t \quad (8)$$

where  $u = \lceil \frac{T-1}{I-1} \rceil$  and  $l = \lfloor \frac{T+1}{I+1} \rfloor$ .

The constraint in (7) ensures that there are at most  $u$  paragraphs between two images, while the constraint in (8) ensures that there are at least  $l$  paragraphs between two images.

**Evaluating Spacing-aware Alignments.** [4] proposes various metrics for automatic evaluation of text-image alignments. BLEU and ROUGE are metrics based on n-gram overlaps between *reference* and *target* texts. In the context of text-image alignment, for each image, the corresponding paragraph from the original story (ground truth) serves as a reference, and the paragraph selected by a model acts as the target. *SemSim* measures the cosine similarity between ground truth (GT) and aligned paragraphs. *ParaRank* is a measure of how far semantically the aligned paragraph is from the GT paragraph. *OrderPreserve* checks how much of the GT image ordering is maintained in the alignment. To evaluate our Spacing-aware Alignment model (from Section 2), we define yet another automatic evaluation metric.

Ideal alignments would space out images evenly within a story. We propose to measure the deviation from such an ideal alignment as an indication of visual appeal. For a story with  $T$  paragraphs and  $I$  (selected) images, the ideal relative distance between two images would be  $1/(|I| - 1)$ .  $\mathbb{T} = 1, 2, \dots, t$  is the set of ordered paragraphs, where  $|\mathbb{T}| = t$ . We define the allocation of images to paragraphs via a function  $f: I \rightarrow \mathbb{T}$  defined as  $f(i) = j$  if and only if  $X_{ij} = 1$ .  $X_{ij}$  is a binary random variable from (1). We now look at the co-domain of the function  $\mathbb{T}' \in \mathbb{T}$ , and sort its elements  $1 \leq j_1 \leq j_2 \leq \dots \leq j_{|I|} \leq t$ . The average image spacing variance of a story is then defined as:

$$\text{Spacing}(X) = 1 - \left[ \frac{1}{|I| - 1} \sum_{k=1}^{|I|-1} \left| \frac{j_{k+1} - j_k}{t - 1} - \frac{1}{|I| - 1} \right| \right] \quad (9)$$

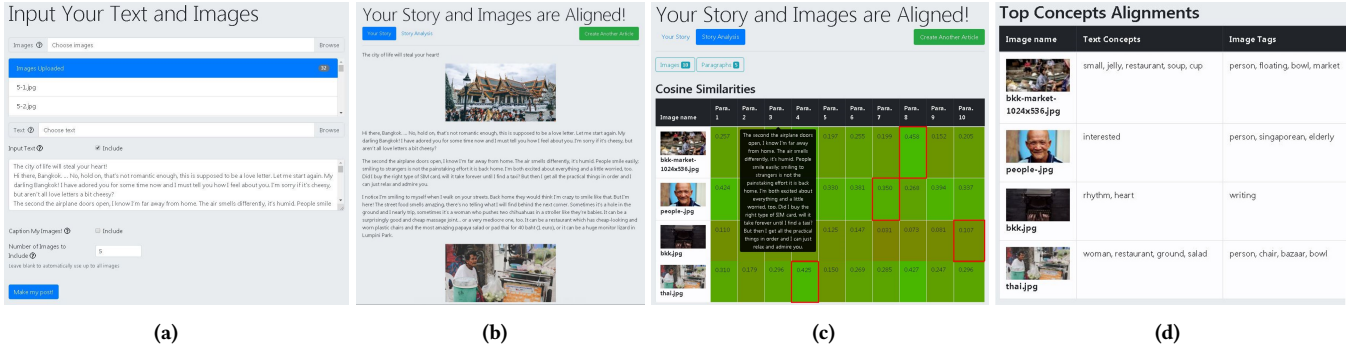
We compare Spacing-aware Alignment with Complete Alignment and baselines from [4] (VSE++ [5] ILP and a random alignment) on all metrics defined in [4], as well as on *Spacing* as defined in Eq. 9. The results can be seen in Table 1; all values are scaled to  $[0, 100]$ . The Spacing-aware Alignment ensures a more uniform image spacing in the story, while sacrificing slightly on semantic coherence of image-paragraph pairs. This can be justified as an acceptable trade-off since images are often just used for general illustration of the overall theme of the story (as observed by [4]), without maintaining a tight semantic fit with surrounding paragraphs. The *Spacing* of the ground truth alignments from the Lonely Planet dataset is 81.03.

**Table 1: Influence of enforced image-spacing on the Lonely Planet dataset: slightly reduced image-text semantic coherence.**

	BLEU	ROUGE	SemSim	ParaRank	OrderPreserve	Spacing
Random	3.1	6.9	75.1	50.0	50.0	78.6
VSE++ [5] ILP	12.6	11.2	84.0	58.1	47.9	79.1
Complete Alignment	45.6	44.5	89.8	72.5	77.4	84.0
Spacing-aware Alignment	43.3	41.5	89.5	71.5	75.5	89.1

\*<https://www.lonelyplanet.com>

†<https://www.asiaexchange.org>



**Figure 1: Our system takes an image collection and a body of text as input (a), and generates a multimodal story (b). Story analysis: the matrix in (c) shows cosine similarities between images and paragraphs, highlighting the alignments with red boxes, hovering over the column headers show the corresponding paragraphs. Similar concepts from aligned images and paragraphs can be seen in (d).**

### 3 SYSTEM OVERVIEW

Following [4], the computational model of our system (SANDI) consists of multiple image descriptors, word embeddings to represent concepts from images and paragraphs, and a combinatorial optimization problem solver.

**Input.** SANDI takes the following mandatory user inputs – an image collection and a body of text.

**Output.** After computation of the image–paragraph alignments, SANDI displays the generated multimodal story.

#### 3.1 Model Components

**Image Descriptors.** In order to detect the contents of the input images, several image descriptors are used similar to [4].

- **Visual Tags:** The Convolutional Neural Network based object detection framework YOLO [12], and the scene detection framework PlacesCNN [13] are invoked.
- **Big-data Tags:** Google Reverse Image Search<sup>‡</sup> tag suggestions often provide information not visually detectable in an image – e.g., locations such as “Shakespeare’s Globe Theatre”.
- **User Tags:** Optionally, if our system fails to automatically identify image contents, user-specified image tags are used.

**Word Embeddings.** We use Word2Vec trained on a Google News corpus to compute cosine similarities between textual concepts.

**ILP solver.** Similar to [4], we use the Gurobi Optimizer to solve the ILP for image selection and placement.

#### 3.2 Interactive Exploration

Let us consider a blogger writing about their last vacation trip. They took over 100 pictures during the trip, but would only like to include 5 representative ones in their trip report. With our system, they can save the time and effort of going through all the images and selecting and placing them within appropriate paragraphs of the textual narrative. Our system automatically performs Image Selection and Image Placement in one seamless optimization step.

In the home page (Figure 1 (a)), users are able to upload a few images or a big pool of images, enter/upload text, specify the number of images to be selected for the story, and specify their choice to automatically caption the images. Once the user uploads their files and their selections, a session is created in the back-end with a session-ID in case any further user interaction is required. This will arise when our system fails to automatically detect visual concepts from the images, and asks for user-specified image tags. Figure 1 (b) shows the output page, where the generated multi-modal story is displayed to the user. Additionally, a Story Analysis page provides the justifications for individual image–paragraph alignments. This is in the form of image–paragraph cosine similarities (Figure 1 c), and similar concepts from the two modalities (Figure 1 d).

The initial requests from the front-end, session handling, as well as automatic image tagging are all performed in a Flask server written in Python. The optimization is handled in a Java Servlets server to take advantage of its speed. We use the Bootstrap CSS framework to create an uncluttered and modern looking user interface.

#### 3.3 Caption Prediction

Most images from the real-life datasets used in [4] are captioned. Following this characteristic of multimodal narratives, we add an image captioning component in our application. While automatic caption generation is a well-studied problem [3], we resort to quote-based caption prediction to obtain more inspiring results. Using a visual-semantic embedding framework [5] to infer text–image similarities, we predict the top-10 related quotes for an image from the Quotes-500K dataset [7]. Among these, the quote with the highest cosine similarity to the aligned paragraph is then displayed as the image caption. Hence, our image captions are not only attractive, but also contextually meaningful. Figure 2 shows some examples.

### 4 USER STUDY

The story-specific selection of images from an image repository and their placement within the story is to a certain extent a matter of subjective preferences. While the problem has been formally characterized and evaluated using automated means [4], we study the inherent subjectivity that it entails through a user evaluation of

<sup>‡</sup><http://www.google.com/searchbyimage?>



Image and Caption	Aligned Paragraph
 <p>“What a rebellious act it is to love yourself naturally in a world of fake appearances.”</p>	<p>Watch out for the Korean wave! The catchy beats, colorful soap operas and gripping dramas are invading countries around the world with a massive force, and are here to stay. And while before it targeted the younger crowds, it’s now getting more and more popular among grown-up folks too!</p>
 <p>“All the world’s a stage and all men and women are merely players.”</p>	<p>No genre of media is excluded: Film, literature, graphic novels, language, food, fashion...you name it. But arguably, the genre with the biggest global impact is the new wave of Korean pop music, commonly referred to as K-pop, with its addicting melodies and innovative choreography.</p>

Figure 2: SANDI predicts contextually meaningful captions.

image–paragraph alignments. We use a random selection of images from both datasets from [4] – Lonely Planet and Asia Exchange.

The user study is designed to collect feedback, on a per-image basis, of which aligned paragraph – from the ground truth (GT) or from our system (SANDI) – is a better semantic fit for an image. We conduct the user study via Figure Eight (formerly, CrowdFlower).

**Design.** Each question consists of an image and two paragraphs (A and B). Contributors are asked to choose one of the following options – A More Relevant Than B, A Equally Relevant to B, A Less Relevant Than B – as answer to the question “Which text is better fitting with the given image?”. We collect 5 judgments per question for a total of 250 questions.

**Avoiding Bias.** The source of the paragraph (GT or SANDI) is not revealed to the contributor. The assignment of the paragraphs (GT to A, SANDI to B etc.) is randomized to eliminate bias. Moreover, data points are chosen such that GT and SANDI paragraphs are of similar length – a difference of maximum 20 words is allowed. This is done to avoid possible bias towards longer or shorter paragraphs.

**Quality Assurance.** Test questions are modeled such that one of the paragraphs belong to an unrelated story. This allows us to eliminate responses from inattentive contributors. The confidence score (pink horizontal bars in Figure 3) for each aggregated result depicts the level of agreement between multiple contributors.

**Result Aggregation.** For each question, the option selected by the majority of the contributors is reported. As shown in Figure 3, SANDI alignments were chosen to be more relevant 46.4% of times, whereas GT was chosen 40.6% of times. This shows that while GT alignments are generally chosen with care, SANDI alignments are even more semantically relevant. Both paragraphs were deemed equally relevant in 13.4% of the questions.

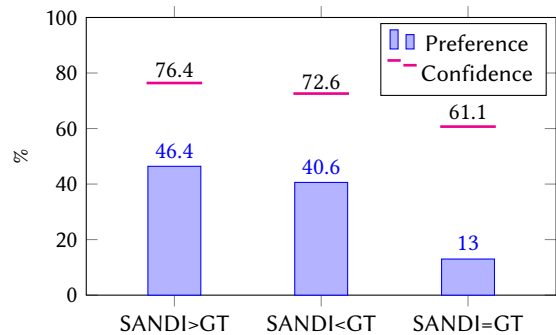


Figure 3: User choice of ground truth (GT) and SANDI alignments.

The observations from the user study support our hypothesis that the problem has a subjective component – the alignments in the ground truth are not absolute, and there exist other suitable alignments – which our application enables exploring.

## 5 CONCLUSION

In this paper, we have presented a web application that automatically selects relevant images from an image repository, and places them within contextual paragraphs of a given body of text, thus generating a multimodal story. In addition to ensuring a high level of semantic fit of images and aligned paragraphs, our application also guarantees uniform spacing of images for better visual appeal. Our user evaluation corroborates that the quality of the obtained image–paragraph alignments are comparable to human judgments. We believe that such an application will be of assistance to online content creators such as bloggers, journalists, commercial content writers, as well as for creation of personal social media posts.

Our system is available at <https://sandi.mpi-inf.mpg.de>, while a video demo is shown at <https://youtu.be/k5gu2pNxdNU>.

## REFERENCES

- [1] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A Corpus Of Text–Image Discourse Relations. *Proc. of NAACL-HLT*.
- [2] Ann Marie Barry. 1997. *Visual intelligence: Perception, image, and manipulation in visual communication*. SUNY Press.
- [3] R. Bernardi, R. Çakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *J. Artif. Intell. Res.* (2016).
- [4] Sreyasi Nag Chowdhury, Simon Razniewski, and Gerhard Weikum. 2019. Story-oriented Image Selection and Placement. *CoRR* (2019).
- [5] Fartash Faghri, David J. Fleet, Jamie Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *BMVC*.
- [6] 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. *NIPS*.
- [7] Shivali Goel, Rishi Madhok, and Shweta Garg. 2018. Proposing Contextually Relevant Quotes for Images. *ECIR*.
- [8] Dhiraj Joshi, James Ze Wang, and Jia Li. 2006. The Story Picturing Engine - a system for automatic text illustration. *TOMCCAP* 2, 1 (2006), 68–89.
- [9] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Relationship Detection with Language Priors. *ECCV*.
- [10] Paul Messaris and Linus Abraham. 2001. The role of images in framing news stories. *Framing public life*. Routledge, 231–242.
- [11] Hareesh Ravi, Lezi Wang, Carlos Muñiz, Leonid Sigal, Dimitris N. Metaxas, and Mubbasir Kapadia. 2018. Show Me a Story: Towards Coherent Neural Story Illustration. *CVPR*.
- [12] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. *CVPR*.
- [13] Bolei Zhou, Àgata Lapedriza, Jianxiang Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. *NIPS*.