

Image Position Prediction in Multimodal Documents

Masayasu Muraoka[†], Ryosuke Kohita[†], Etsuko Ishii[‡]

[†]IBM Research – Tokyo, Japan

[‡]Hong Kong University of Science and Technology, Hong Kong
mmuraoka@jp.ibm.com, kohi@ibm.com, eishii@connect.ust.hk

Abstract

Conventional multimodal tasks, such as caption generation and visual question answering, have allowed machines to understand an image by describing or being asked about it in natural language, often via a sentence. Datasets for these tasks contain a large number of pairs of an image and the corresponding sentence as an instance. However, a real multimodal document such as a news article or Wikipedia page consists of multiple sentences with multiple images. Such documents require an advanced skill of jointly considering the multiple texts and multiple images, beyond a single sentence and image, for the interpretation. Therefore, aiming at building a system that can understand multimodal documents, we propose a task called image position prediction (IPP). In this task, a system learns plausible positions of images in a given document. To study this task, we automatically constructed a dataset of 66K multimodal documents with 320K images from Wikipedia articles. We conducted a preliminary experiment to evaluate the performance of a current multimodal system on our task. The experimental results show that the system outperformed simple baselines while the performance is still far from human performance, which thus poses new challenges in multimodal research.

Keywords: resource creation, multimodal document understanding, vision and language

1. Introduction

Connecting the modalities of vision and language is one of the most ambitious goals in computer vision and natural language processing. A number of multimodal tasks has been proposed and made a great progress to this end. For example, caption generation (Lin et al., 2014) and visual question answering (VQA) (Antol et al., 2015; Agrawal et al., 2017) are well-known conventional multimodal tasks that have received much attention in recent years (Hossain et al., 2019; Kafle and Kanan, 2017; Wu et al., 2017). Caption generation aims to generate a caption to describe the contents of an input image with a natural language sentence while the goal of VQA is to answer a question written in natural language about a given image. These tasks can be solved by interpreting an image and the corresponding short text such as a caption or question.

However, real multimodal documents consist of multiple sentences and multiple images as seen anywhere. For example, newswire articles contain photographs of events, cooking recipes include pictures of intermediate stages of cooking, and Wikipedia articles contain various types of images showing people, buildings, scenery, and various objects. Being placed at appropriate positions in such a document, images help people understand the whole document by naturally considering the correspondence of the images and document. Unfortunately, existing multimodal tasks have not dealt with this high-level correspondence, because an instance in a typical dataset often consists of a single image and a single sentence. As a result, existing methods fail to understand real multimodal documents.

To solve this problem, we propose a task called image position prediction (IPP). Figure 1 shows an overview of our task. Given a document and a set of images, the goal of this task is to find an appropriate position of each image in the document that maximizes readers’ understanding of the document. The task requires considering multiple texts as well as multiple images. More specifically, the task involves three key challenges, all of which contribute to un-

derstanding multimodal documents. (i) It requires considering longer contexts of texts and the relations between them, in other words, document structures. (ii) A system to solve the task also has to relate multiple images. For example, both of the first two images on the right in Figure 1 complement the corresponding text by highlighting the difference of the devices separately shown in the images. (iii) Because documents describe various types of topics, broad coverage of vocabulary is needed, including common nouns as well as proper nouns. Other wide-ranging skills such as keyphrase extraction and inference on scenes may also be required. We expect that these challenges will pave the way for future applications on multimodal documents, such as caption generation for newswire articles, automatic picture book generation from texts, and album creation based on an event description.

To address this task, we automatically construct a dataset of 66K multimodal documents with 320K images from Wikipedia articles. We focus on Wikipedia because it has natural correspondences between the texts and images in its HTML sources. Also Wikipedia articles describe various topics and have structures similar to paragraphs, which we refer to as “sections” in this paper. Thus, Wikipedia offers all the challenges that we impose on our task. We believe that the dataset we construct is the largest among existing multimodal datasets in terms of the numbers of images, documents, and vocabularies.

We conduct a preliminary experiment to evaluate the performance of a current multimodal model on our task. We use the Pythia model (Jiang et al., 2018), the winning model in the VQA Challenge 2018. While the original model adopts an attention mechanism to jointly account for textual and visual features from an input of a question and image, it cannot consider sets of texts and images, which are the input to our task. Therefore, we extend it to accept sets of texts and images, so that the extended model can compute the interactions among them through the attention mechanism. The experimental results show that although

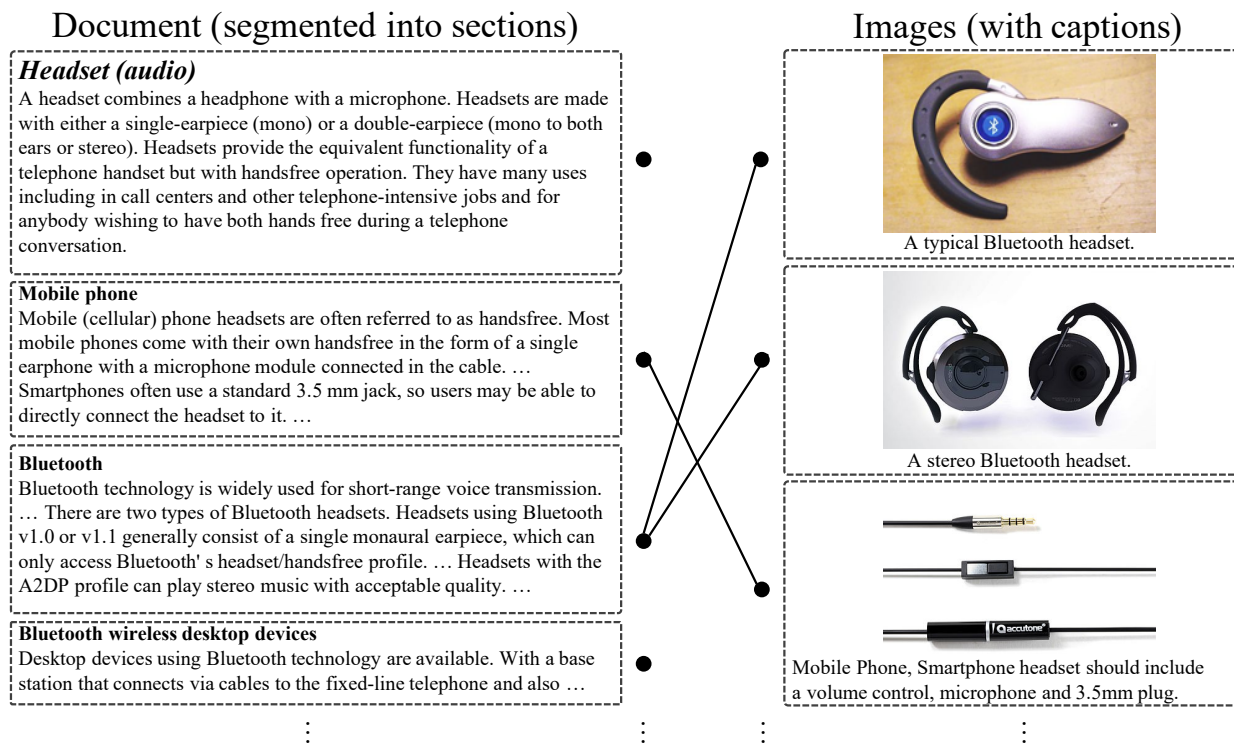


Figure 1: Overview of our IPP task. The example is taken from the English Wikipedia article “Headset (audio)”.

the model achieved promising results, the performance is still limited, which means that our task is reasonably difficult to solve and poses new challenges in multimodal research.

2. Related Work

Many works have proposed large multimodal datasets with different characteristics. MS COCO was proposed by Lin et al. (2014) as a benchmark dataset for image captioning. It contains more than 12K images, and each image has at least five ground-truth captions. The images were fetched from Flickr, an image-hosting service while the captions were separately annotated by using Amazon Mechanical Turk. Goyal et al. (2017) introduced a VQA v2 dataset with reduced language bias in the answers, making it more balanced than the first VQA dataset (Agrawal et al., 2017). Visual Genome (Krishna et al., 2017) is another multimodal dataset that has dense annotations of image regions, their region phrases, object attributes, relationships between objects, scene graphs, and QAs. Each image has an average of 50 regions marked, and each region phrase describes the contents of a marked region in an image with natural language. The descriptions of these datasets were independently annotated by crowdworkers who were not familiar with the background of the provided images, which made the words in the descriptions general in terms that common nouns are frequently used. In contrast, KVQA (Shah et al., 2019) was created with specialization in proper nouns to enable reasoning over world knowledge. In particular, knowledge on people in Wikipedia articles was included and used in the questions.

However, none of above datasets consists of instances with longer texts or multiple images, preventing systems from dealing with such instances. Krause et al. (2017) constructed a paragraph-captioning dataset containing paragraphs as captions instead of sentences. Therefore, this dataset requires paragraph-level reasoning to generate paragraphs describing given images. TQA (Kembhavi et al., 2017) is another challenging dataset that offers lessons obtained from middle-school textbooks. Each lesson consists of multiple topical paragraphs along with instructive images such as diagrams or illustrations. To answer questions in TQA, a system must understand the whole lessons by associating multiple paragraphs and images in the lessons.

More recently, researchers have proposed several multimodal tasks requiring advanced skills beyond those learned from longstanding multimodal tasks such as caption generation or VQA. Agrawal et al. (2016) proposed a task for sequencing jumbled image-caption pairs belonging to the same story in chronological order. Bosselut et al. (2016) tried to learn typical sequences of events from photo albums. Iyyer et al. (2017) compiled a new dataset from comic book panels which consist of pairs of stylized artwork and dialogue. The goal of the work was to examine the capabilities of multimodal models to understand comic books, by requiring them to predict the character utterances or images in the following panels in a comic book. Biten et al. (2019) addressed caption generation for news articles whose captions differ from crowdsourced captions because news articles involve various named entities. Zhu et al. (2015) aimed to align movie shots with the corresponding passages in the books. Hessel et al. (2019) trained a

model that aligns sentences with images in various types of multimodal documents, such as Wikipedia articles, cooking recipes, and photo albums, in which the ground-truth alignments of a sentence and image are not available during training. These multimodal tasks pose new challenges, and the derived datasets have unique characteristics of the tasks. Nevertheless, our task poses other new challenges. Furthermore, to the best of our knowledge, our dataset is the largest in terms of the number of images associated with longer texts and the vocabulary size, as well as other statistics, as we describe later in Section 4.2.

3. Task Formulation

Here, we formulate our IPP task and give the notations used throughout the paper.¹ We use a large dataset of multimodal documents, each consisting of a document and multiple images. Figure 1 shows an example. A document can include tens to hundreds of sentences and have structures among them. Images complement the document to help readers understand it. Some images are associated with captions. We assume that documents in the dataset have already been segmented into meaningful textual units, such as sections, paragraphs, or sentences, according to the structures in the documents. Given these pairs of images and textual units, the task is to find plausible assignments of the images with the textual units so that the assigned images complement the corresponding textual units.

We formalize this task as follows. For a multimodal document d , the set of images in d is denoted as V and the set of textual units in d is denoted as S . The numbers of elements in these sets are denoted as $|V|$ and $|S|$, respectively. Each $v \in V$ may consist of a visual part as an actual image, and a textual part as the caption. Each caption and textual unit $s \in S$ consist of a sequence of words. We regard $d = \langle V, S \rangle$ as a bipartite graph of $|V|$ image nodes and $|S|$ textual unit nodes, where the edges represent the correspondence between the images and the corresponding textual units in d .

We define IPP as a graph completion task of the bipartite graph by predicting all the plausible edges between V and S . Completing the graph requires reasoning over the whole graph, considering multiple texts as well as multiple images. Let $\mathbf{A} \in [0, 1]^{|V| \times |S|}$ represent an assignment matrix of edges, where a_{ij} , the (i, j) -element of \mathbf{A} , represents the assignment probability of an edge connecting the i -th image node and j -th textual unit node. While every image node $v \in V$ has a single (undirected) edge to a textual unit node $s \in S$, a textual unit can have no or multiple edges from different image nodes. Given this nature, we impose a restriction on \mathbf{A} such that each row \mathbf{a}_i in \mathbf{A} sums up to one, that is, for each i ,

$$\sum_j a_{ij} = 1. \quad (1)$$

This means that each row $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{i|S|})$ is a probability distribution over S ; we represent the probability distribution for each image v_i as $\phi(v_i)$ to directly indicate the

input image. Thus, we can obtain assignments of images to the corresponding textual units by taking the highest probability for each row \mathbf{a}_i :

$$\mathbf{y} = [y_1, \dots, y_{|V|}]^T, \quad (2)$$

$$y_i = \operatorname{argmax}_{j \in \{1, \dots, |S|\}} a_{ij}, \quad (3)$$

where y_i denotes the resulting plausible textual unit for an image v_i in V . Note that we can loosen this criterion by taking the k highest probabilities to consider more candidate textual units.

We evaluate a system solving the IPP task by measuring accuracy@ k , the averaged accuracy over a given dataset when we consider the k highest-confidence candidates for each instance. Specifically, if a system predicts the correct textual unit within the k candidates for a given image, it earns a ‘‘point.’’ Then, accuracy@ k is the fraction of the total obtained points to the total number of images in the dataset. In our experiment, we used $k \in \{1, 3, 5\}$.

4. Dataset

4.1. Dataset Curation

In this section, we explain how we automatically create our dataset for the IPP task from English Wikipedia.² In Wikipedia, an article can have sections delineated by the HTML tag $\langle h \rangle$, and images (with captions) denoted by the HTML tag $\langle img \rangle$. We consider any text between two different and consecutive $\langle h \rangle$ tags as the text of the section. Likewise, the image positions are determined by two consecutive $\langle h \rangle$ tags nesting an $\langle img \rangle$ tag. Thus, Wikipedia inherently suits the IPP task formulation, and we can automatically build a large dataset via these explicit markers. Here, we describe the data creation procedure in detail. First, we extract sections, images and captions for all articles (i.e., documents) in a Wikipedia dump³. The extracted sections do not contain texts in a table or an infoboxes. Consequently, we obtained 5,870,656 documents, not including ones failed to be parsed. Then, we picked up documents which have 10 to 50 sections and 2 to 30 images. This left 161,763 documents, and we then tried to collect all images in these documents. We restricted images by their original file extensions, allowing any of jpeg, JPEG, jpg, JPG, png, or PNG. In addition, we converted all the extracted images to an RGB format and compressed the image file size by reducing the image quality to some extent to make the subsequent processing easier. Eventually, we obtained a total of 66,947 documents with 320,200 images as our IPP dataset.

4.2. Analysis

We summarize the statistics of our dataset in Table 1 and 2 comparing with previously proposed multimodal datasets explained in Section 2. Our dataset offers notable characteristics from both quantitative and qualitative perspectives. **How many sentences and words are in our dataset?**

¹We borrow some notations and statements from the work by Hessel et al. (2019), because a data structure we use is similar to it.

²<https://github.com/muraoka7/tool4ipp>

³We used the latest English Wikipedia dump as of June 1st, 2019.

Dataset	#docs	#images	#t_units	Vocab size	Multi sents / t_unit	Multi images / t_unit	Included noun type	
MS COCO	–	123,287	616,767	captions	22,382	✗	✗	common
VQA v2	–	204,721	1,105,904	questions	13,634	✗	✗	common
Visual Genome	–	108,077	5,408,689	phrases	56,505	✗	✗	common
KVQA	–	24,602	183,007	questions	8,338	✗	✗	proper
Krause et al.	–	19,561	19,561	descriptions	9,719	✓	✗	common
TQA	1,076	3,181	9,343	topics	15,791	✓	✓	common
Ours	66,947	320,200	1,129,321	sections	2,139,704	✓	✓	common & proper

Table 1: Comparison of our dataset with previously proposed multimodal datasets. In the table, “t_unit” indicates a textual unit, which differs among the datasets.

Dataset	#sents		#words			#images	
	/ doc	/ t_unit	/ doc	/ t_unit	/ sent	/ doc	/ t_unit
MS COCO	5.0*	1.0	56.7*	11.3	11.3	–	1.0
VQA v2	5.4*	1.0	38.7*	7.2	7.2	–	1.0
Visual Genome	50.0*	1.0	263.9*	5.3	5.3	–	1.0
KVQA	7.4*	1.0	84.8*	11.4	11.4	–	1.0
Krause et al.	5.7	5.7	68.5	68.5	11.9	–	1.0
TQA	75.7	8.7	920.8	106.0	12.2	3.0	1.2
Ours	150.4	8.9	3,346.6	198.4	22.3	4.8	1.3

Table 2: Statistics in terms of a document (doc), textual unit (t_unit), and sentence (sent). Numbers with * are obtained by treating all the textual units associated with an image as a document, despite lacking document structures, unlike TQA and our dataset.

For a fair comparison, we applied the `spacy` package (Honnibal and Montani, 2017)⁴ to all the datasets listed in Table 1 to split sentences and tokenize words. Our dataset contains more than 10M sentences in 1M sections and each section has 8.9 sentences on average. The average sentence length is 22.3 words, and the average section length is 198.4 words, which is the largest among the compared datasets. The number of sentences per document is almost double that of TQA (150.4 vs. 75.7 from Table 2), and the number of words per document is more than 3.5 times greater. This confirms that understanding of longer texts is required for the IPP task with our dataset.

How many images belong to a document or textual unit?

The number of images in our dataset is 320K, which is the largest listed in Table 1. As previously explained, one section in our dataset can have multiple images (1.3 images on average), while one textual unit (e.g., a caption or question) in the other datasets, except for TQA, exactly corresponds to one image. In addition, our dataset has 4.8 images per document. It thus requires reasoning over multiple images for a system to complete the task.

How large is the vocabulary size?

Table 2 also lists the vocabulary size of our dataset and the others. We make sure that our dataset has the largest vocabulary size (>2M) by two orders of magnitude as compared to the others. Because of the nature of Wikipedia, which is one of the largest online encyclopedias, our dataset obviously contains various topics of both common and world

knowledge, expressed by common nouns and proper nouns, respectively. Hence, with our dataset, we can test the capabilities of a system to manage a huge variety of vocabulary to capture the semantics of the documents.

What skills are required to achieve the task?

We manually checked a subset of our dataset and categorized the challenging skills required to solve the IPP task. In addition to the key challenges explained in Section 1, we found that our task requires the following: *keyphrase extraction*, *object-keyphrase matching*, *inference on scenes*, *optical character recognition (OCR)*, *diagram understanding*, and *information integration/selection from whole documents*. Although this list is not exhaustive, we frequently observed instances requiring at least one of those skills to solve the task. Figure 2 shows examples of those instances.

5. Preliminary Experiment

To measure the complexity of our task, we study how well a current multimodal system suits our IPP task.

5.1. Extended Pythia Model

We extend Pythia (Jiang et al., 2018), the model that won the VQA Challenge 2018. The original model is based on a common architecture used in VQA (Kafle and Kanan, 2017; Wu et al., 2017). Given a question sentence and an image as an input, that model first encodes them separately with respective encoders to obtain feature vectors. It then computes an attention layer to attend specific parts in the image by considering what the question asks, and it generates the answer from the attended features.

⁴<https://spacy.io/>




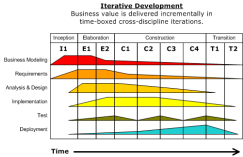
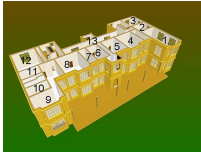

Required skill	keyphrase extraction, object-keyphrase matching	inference on scenes	OCR
Article name	Tricycle	Wine	Railway platform
Image (Caption)	 Spidertrike	 Oak wine barrels	 A common marking at curved platforms on the London Underground.
Section	Tricycle rickshaw Most cycle rickshaws, used for carrying passengers for hire, are tricycles (<i>followed by 100 words to describe cycle rickshaws</i>) <u>Spidertrike</u> is a recumbent cycle rickshaw that is used in central London and operated by Eco Chariots. (<i>followed by 59 words to describe spidertrike</i>) (a)	Storage Wine cellars, or wine rooms, if they are above-ground, are places designed specifically for the storage and aging of wine. Fine restaurants and some private homes have wine cellars. (<i>followed by 261 words but neither the words "oak" nor "barrels" appeared</i>) (b)	Curvature (73 words precede) Usually such platforms will have warning signs, possibly auditory, such as <u>London Underground's famous phrase "Mind the gap"</u> . (<i>followed by 77 words</i>) (c)
Required skill	OCR, diagram understanding	information integration/selection from whole documents, OCR, keyphrase extraction, object-keyphrase matching	
Article name	Unified Process	Montacute House	
Image (Caption)	 Diagram illustrating how the relative emphasis of different disciplines changes over the course of the project	 First floor: 1: Library (formerly known as the Great Chamber); (<i>followed by names of the remaining numbered rooms</i>)	 Second-floor plan. Key: 1: Long Gallery. (<i>followed by names of the remaining numbered rooms</i>)
Section	Iterative and incremental (10 words precedes) <u>The Elaboration, Construction and Transition phases are divided into a series of timeboxed iterations.</u> (<i>followed by 54 words</i>) <u>the relative effort and emphasis will change over the course of the project.</u> (d)	First floor The first floor contains one of the grandest rooms in the house, the <u>Library</u> . The room was formerly known as <u>the Great Chamber</u> ; (<i>followed by 444 words to describe the rooms in the floor</i>) (e)	Second floor A notable feature of the house is the 172-foot (52 m) second-floor <u>Long Gallery</u> , spanning the entire top floor of the house; (<i>followed by 192 words to describe the rooms in the floor</i>)

Figure 2: Selected samples of images and the corresponding textual units from our dataset, showing different skills required to accomplish our task. The shown textual units are ground-truth. The underlined textual parts are relevant to the required skills. (a) The instance requires extracting the word “Spidertrike” in the long textual unit (around 200 words) and matching it to the object in the given image. (b) It is required to understand the scene of the image because the correct textual unit describes a general topic (i.e., “Storage”), but does not have concrete words such as “oak” or “barrels” as given in the caption. (c) This instance can be easily solved if OCR is available because almost the same words “Mind the gap” appears in the associated textual unit and the image. (d) We can test whether a system can understand diagrams with this instance, where the system has to associate what the diagram depicts with the corresponding texts (e.g., the “Elaboration” column in the image is divided into two sub-columns, “E1” and “E2”, as explained in the textual unit). (e) Similar images appear in different textual units, in which the images complement the respective textual units, and thus, a system must relate multiple textual units and multiple images in the document.

We modify the Pythia model to receive a set of textual units as well as images as an input. This allows the model to relate multiple textual units and multiple images to consider the interactions among them. Figure 3 shows an overview of our extended Pythia model. We give a detailed explanation of it below.

For textual parts, we first tokenize a textual unit s and caption c with a WordPiece tokenizer (Wu et al., 2016)⁵,

⁵In practice, we use the tokenizer implemented in the follow-

ing PyTorch’s Transformers, mentioned below. We then obtain BERT embeddings through a pretrained BERT model (Devlin et al., 2019)⁶ for the first 512 subwords. Instead of using GloVe embeddings (Pennington et al., 2014) as in the original Pythia model, we adopt BERT embeddings because they have shown improvement in numerous natural

ing PyTorch’s Transformers, mentioned below.

⁶We used PyTorch’s Transformers (Wolf et al., 2019) as the pretrained model.

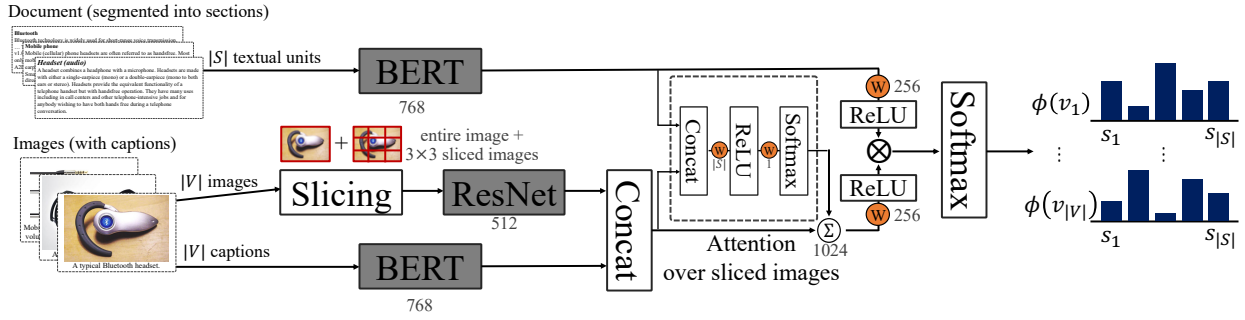


Figure 3: Overview of our extended Pythia model. The gray and white boxes indicate pretrained models with fixed parameters and operations, respectively. The W 's in orange circles indicate learnable parameters. The symbol \otimes denotes matrix multiplication. The gray numbers are the dimensionalities of the resultant vector representations for each feature.

language tasks (Wang et al., 2019) such as sentiment analysis, sentence similarity, question answering, and textual entailment. Although we discard the remaining subwords if we obtain more than 512 subwords for a given s or c , we consider that number large enough because the average number of words per textual unit is much lower (198.4) as seen in Table 2. If the number of subwords is less than 512, then we pad with zero vectors to make the length even (i.e., 512). Each BERT embedding is taken from the last hidden layer, corresponding to the last token in the subwords, and is represented by a d_{txt} -dimensional real-valued vector, for which we set $d_{\text{txt}} = 768$. Although the original model further encodes the obtained embeddings with recurrent neural networks (RNNs) such as GRU (Cho et al., 2014) or LSTM (Hochreiter and Schmidhuber, 1997), we did not use them because we can successfully obtain interactions among textual units in the subsequent layers, thereby decreasing the number of model parameters.⁷

As for visual parts, we use pretrained convolutional neural networks (CNNs) to encode images. Specifically, we use an 18-layer Residual Network (ResNet) (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015). Because our task requires focusing on specific parts of images in addition to whole-image understanding, we extract image features from an entire image as well as sliced sub-images of it. By slicing into 3×3 sub-images, we obtain $K = 10$ image features for one image, where the last feature is from the whole image. Each image feature is represented as a d_{img} -dimensional real-valued vector ($d_{\text{img}} = 512$) taken from the final average pooling layer in ResNet, and thus, we have a $K \times d_{\text{img}}$ feature matrix for each image. When we choose to use the captions of given images, we simply copy and concatenate the caption embeddings obtained through BERT with each feature vector, resulting in a matrix with a dimensionality of $K \times (d_{\text{img}} + d_{\text{txt}})$.

Once we obtain the textual embeddings and image features for a given document, we compute an attention layer to focus on a set of sub-images with different importance according to a given set of textual units. Specifically, we

can obtain the attention layer by applying a feature concatenation, a nonlinear and linear transformation, and softmax function as follows. Let $\mathbf{S} \in \mathbb{R}^{|S| \times d_{\text{txt}}}$ be the matrix of textual features, and $\mathbf{V}(i) \in \mathbb{R}^{K \times (d_{\text{img}} + d_{\text{txt}})}$ be the matrix of visual features for the i -th image ($1 \leq i \leq |V|$). We first concatenate \mathbf{S} with $\mathbf{V}(i)$ to create a matrix $\mathbf{M} = [\mathbf{M}(1), \dots, \mathbf{M}(|V|)]^T$ by concatenating all the rows of \mathbf{S} with each image feature vector, resulting in $\mathbf{M}(i) \in \mathbb{R}^{K \times (d_{\text{img}} + d_{\text{txt}} + |S| \times d_{\text{txt}})}$. We then obtain our attention layer α by the following equation:

$$\alpha = \text{softmax}((g_2 \circ f \circ g_1)(\mathbf{M})). \quad (4)$$

f and g_i ($i \in \{1, 2\}$) denote a nonlinear and linear transformation, respectively defined by $f(\mathbf{X}) = \text{ReLU}(\mathbf{X})$ and $g_i(\mathbf{X}; \mathbf{W}_i, \mathbf{B}_i) = \mathbf{X}\mathbf{W}_i^T + \mathbf{B}_i$. Note that \mathbf{W}_i and \mathbf{B}_i are learnable model parameters, and the dimensionalities are $\mathbf{W}_1 \in \mathbb{R}^{|S| \times (d_{\text{img}} + d_{\text{txt}} + |S| \times d_{\text{txt}})}$, $\mathbf{B}_1 \in \mathbb{R}^{(|V| \times K) \times |S|}$ where each row stores the same value, $\mathbf{W}_2 \in \mathbb{R}^{1 \times |S|}$, and $\mathbf{B}_2 \in \mathbb{R}^{(|V| \times K) \times 1}$. To compute the attended image features, we take a weighted sum over each image feature matrix $\mathbf{V}(i)$ with the attention weights given by α :

$$\widehat{\mathbf{V}}(i) = \sum_{k=1}^K \alpha_{i(K-1)+k} \mathbf{V}(i, k), \quad (5)$$

where $\mathbf{V}(i, k)$ denotes the k -th sub-image feature vector for the i -th image.

We then obtain the assignment probabilities $\phi(v_i)$ for each image $v_i \in V$ from $\widehat{\mathbf{V}}$ and \mathbf{S} , especially by

$$\phi(v_i) = \text{softmax}((g_3 \circ f)(\widehat{\mathbf{V}}(i)) (g_4 \circ f)(\mathbf{S})^T).$$

We set the resulting dimensionality of the linear transformation with \mathbf{W}_3 and \mathbf{W}_4 to 256. By stacking all the $\phi(v_i)$ for all images, we obtain the assignment matrix $\mathbf{A} = [\phi(v_1), \dots, \phi(v_{|V|})]^T$, from which we can also obtain the prediction of the image positions \mathbf{y} , where $y_i = \text{argmax} \phi(v_i)$ is the resulting textual unit. In the training phase, we adopt the cross-entropy loss over image positions and backpropagate it to optimize the model parameters.

5.2. Experimental Setup

We randomly split our dataset into train/dev/test sets with 53,557/6,695/6,695 documents, respectively. We iteratively

⁷In our experiment, we found that our model could achieve promising performance even without such RNNs. In fact, we found the training of such a larger model quite unstable because the training frequently caused the exploding gradient problem.

	acc@1	acc@3	acc@5
Random	6.09%	17.46%	27.24%
Top frequency	8.61%	17.13%	41.54%
Our model	21.28%	46.11%	62.55%
w/o caption	20.36%	45.36%	61.90%
w/o image	21.81%	46.76%	63.13%
Human	60.00%	–	–

Table 3: Values of accuracy@k for our extended Pythia model with different input features, as well as other naive baselines, on the test set.

trained our model with mini-batches of size 16 by using stochastic gradient descent. For each instance in a mini-batch, we randomly swapped the orders of the input images to prevent learning from just the orders. In contrast, we fixed the orders of the input textual units because they were linguistically structured and the order was meaningful. We set the number of epochs to 10,000 and preserved the model parameters giving the best performance on the dev set. As in the original work on Pythia (Jiang et al., 2018), we used AdaMax (Kingma and Ba, 2014) as an optimization algorithm and a warm-up strategy for effectively updating the learning rates with the initial learning rate of 0.001.⁸

5.3. Result

Here, we report the results of our model on the test set with the best parameters obtained on the dev set in terms of acc@1. Table 3 summarizes the ablative performance of our model with different input features, as well as the performance of other naive baselines. “Random” chooses image positions at random and “Top frequency” always predicts the most frequent positions in our dataset. Performance by humans is also shown in the bottom row, which is the average accuracy by the three authors who manually solved 100 images in randomly-selected 21 articles.

Our model using both image and caption features achieved accuracy@k of 21.28%, 46.11%, and 62.55% with $k \in \{1, 3, 5\}$, respectively. It thus outperformed the naive baselines by a large margin up to ~ 29 percentage points (accuracy@3). These results indicate that our model successfully captured the meanings of given sets of texts and images. It is also implied that a system adopting a typical architecture for VQA could solve the task to some extent.

However, our model is still far from human performance, and contribution of each modality (caption or image) is still limited. Without captions, the performance slightly decreased, by 0.65 to 0.92 points (fourth row). In contrast, when we used only captions (the “w/o image” row), the performance was slightly superior to that of the model using both images and captions. This can be reasonable because captions sometimes share similar or same words with textual units. Another reason could be that our model may not handle features from different modalities, i.e., texts and images, which is consistent with the results by Kembhavi et al. (2017) and should be investigated more in the future. We manually analyzed instances that our model predicted

correctly and incorrectly. Figure 4 shows the examples of the predictions by our model. We found that our model could learn the typical correspondences between texts and images. For example, our model learned that monochrome images tend to be associated with sections about past times such as “History” or “Early age” (first example in Figure 4(a)). Our model also learned that images of maps and landscapes are more likely to be placed in “geographic” sections (second one in Figure 4(a)), while building images were put in “architecture” sections (third one in Figure 4(a)). This implies that our model captured the general meanings of images and texts, even though OCR and diagram understanding were not used in our model. More interestingly, we observed some encouraging examples, where our model could assign similar images to different sections in a document (Figure 4(b)). We conjectured that our model considered the contexts (e.g., backgrounds or types of images) of the images together with the captions. On the other hand, our model could not cope well with the localization of the contexts, and thus, the prediction was affected by the general meanings of the images. For example, as shown in the first and second images in Figure 4(c), when images depict typical objects such as roads, our model wrongly associated them with the typical sections (the “Road” section), even though the captions share the same words with the correct sections. Our model also did not consider the structure of the document. Although the last example in Figure 4(c) required to relate the correct section and its subsections to describe the types of platforms, all of which are summarized in the input image, our model failed to capture it. These would be solved if a system could successfully account for the structured texts and the relative differences of the meanings of images in a given document. We believe that our dataset enables learning such semantics that lie in multimodal documents, which is not obtained via the existing multimodal datasets.

6. Conclusion




We paved the way for learning multimodal documents by proposing a novel task called image position prediction (IPP). The IPP task offered three key challenges and several challenging skills that allow machines to go beyond conventional multimodal tasks. To study our task, we also automatically created the largest multimodal dataset from Wikipedia articles. From a preliminary experiment, we showed that the proposed task is moderately difficult and the further research is needed.

Since our dataset contains correspondence between longer texts and images, it can be used to train multimodal encoders that effectively encode interactions in textual and visual information by enriching the encoded features with the other modality. It would be interesting to apply such encoders to other tasks including single-modal tasks to evaluate the effectiveness.

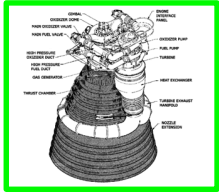

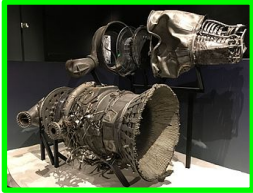
7. Acknowledgements

We thank the anonymous reviewers for their thoughtful comments. We also thank our colleagues for many discussions and their constructive comments that greatly improved the manuscript.



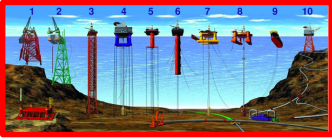
⁸See the original paper (Jiang et al., 2018) for more detail.

Article name	Rheinfelden	Izena Island	Dithmarschen
Image (Caption)	 Bentley Priory (c.1800)	 Map of the Okinawa Islands, showing the location of Izena ...	 Marne church and city hall
Section	History ... The Marquess of Abercorn acquired the estate, along with Bentley Priory, in 1839. ...	Geography ... to the northwest of Okinawa Island, and southeast of Iheya Island. ...	Architecture The Dithmarschen landscape was long dominated by churches. ...

(a) Examples that our model could capture typical correspondences.

Article name	Rocketdyne F-1		
Image (Caption)	 F-1 rocket engine components	 F-1 engine on display at INFINITY Science Center.	 Recovered F-1 engine parts on display at the Museum of Flight in Seattle.
Section	Design ... The heart of the engine was the thrust chamber, ... A gas generator was used to drive a turbine ... Below the thrust chamber was the nozzle extension ...	Locations of F-1 engines ... and the first stage from SA-515 is on display at the INFINITY Science Center at John C. Stennis Space Center in Mississippi. ...	Recovery ... On May 20, 2017 the Apollo permanent exhibit opened at the Museum of Flight in Seattle, WA and displays engine artifacts ...

(b) Examples that our model successfully distinguished the contexts of the images.

Article name	Transport in Delhi	Transport in Delhi	Oil platform
Image (Caption)	 Radio Taxi near airport	 The DND Flyway	 1, 2) conventional fixed platforms; 3) compliant tower; ...
Section (predicted)	Road Two upcoming bridges over Yamuna will connect Faridabad to Noida and Greater Noida. ...	Road Two upcoming bridges over Yamuna will connect Faridabad to Noida and Greater Noida. ...	Challenges Offshore oil and gas production is more challenging than land-based installations...
Section (gold)	Taxis ... Recently, Radio Taxis have started to gain ground in Delhi. ...	Expressways and National Highways ... DND Flyway connects Delhi with its other financial hub, ...	Types Larger lake- and sea-based offshore platforms and drilling rig for oil. (<i>the platforms in the image are described in the following subsections</i>)

(c) Examples that our model failed to predict the correct positions.

Figure 4: Example predictions of our model. The first and second rows show positive examples while the last indicates failure cases.

8. Bibliographical References

- Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., and Bansal, M. (2016). Sort story: Sorting jumbled images and captions into stories. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 925–931.
- Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Biten, A. F., Gomez, L., Rusinol, M., and Karatzas, D. (2019). Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.
- Bosselut, A., Chen, J., Warren, D., Hajishirzi, H., and Choi, Y. (2016). Learning prototypical event structure from photo albums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1769–1779.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hessel, J., Lee, L., and Mimno, D. (2019). Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):118:1–118:36, February.
- Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J., Daume, H., and Davis, L. S. (2017). The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7186–7195.
- Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., and Parikh, D. (2018). Pythia v0.1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.
- Kafle, K. and Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.
- Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., and Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Shah, S., Mishra, A., Yadati, N., and Talukdar, P. P. (2019). KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

- tations.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., and van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.
- 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 *Proceedings of the IEEE international conference on computer vision*, pages 19–27.