The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

News Content Completion with Location-Aware Image Selection

Zhengkun Zhang[1], **Jun Wang**[2], **Adam Jatowt**[3], **Zhe Sun**[4*], **Shao-Ping Lu**[1], **Zhenglu Yang**[1*]
[1]TKLNDST, CS, Nankai University, China,
[2]Ludong University, China,
[3]Kyoto University, Japan,
[4]Computational Engineering Applications Unit, RIKEN, Japan
{zhangzk2017, junwang}@mail.nankai.edu.cn, jatowt@gmail.com, zhe.sun.vk@riken.jp, {slu, yangzl}@nankai.edu.cn

Abstract

News, as one of the fundamental social media types, typically contains both texts and images. Image selection, which involves choosing appropriate images according to some specified contexts, is crucial for formulating good news. However, it presents two challenges: *where* to place images and *which* images to use. The difficulties associated with this *where-which* problem lie in the fact that news typically contains linguistically rich text that delivers complex information and more than one image. In this paper, we propose a novel end-to-end two-stage framework to address these issues comprehensively. In the first stage, we identify key information in news by using location embeddings, which represent the local contextual information of each candidate location for image insertion. Then, in the second stage, we thoroughly examine the candidate images and select the most context-related ones to insert into each location identified in the first stage. We also introduce three insertion strategies to formulate different scenarios influencing the image selection procedure. Extensive experiments demonstrate the consistent superiority of the proposed framework in image selection.

Introduction

Given recent developments in social media, large amounts of content are being on the Web every day. One of the key procedures in content creation is the selection of appropriate images. Current image selection systems (Wang et al. 2019; Liu et al. 2020) usually deliver only one image for each article, however, this practice does not match the requirements of many social media platforms, especially online news sites that often contain more than one image per article. Moerover, the selected images must be closely related to the article they accompany but sufficiently differ from each other according to the surrounding textual contexts that they are supposed to visually represent or enrich. To the best of our knowledge, automatically selecting multiple images for a given input text has rarely been explored in the literature. In this work, we focus on a novel news image selection task, which assumes selecting a series of suitable images to complete input news articles.

A sample article shown in Figure 1 depicts a news article taken from the $DailyMail$ website; this article is composed of several distinct events that cannot be sufficiently illustrated by a single image. In the present example, multiple images are required to complete the news content. This task, however, faces two key challenges: *where* to place images and *which* images to show. The first challenge is to identify the appropriate image insertion locations. In practice, when editors decide to insert an image into a certain location within the text of a news article, the context of this location tends to be salient with respect to the news content. The previous relevant works (Lin et al. 2014; Young et al. 2014) employ the image description, which is usually a single sentence containing limited information, for retrieving the corresponding image. Unlike these works, we face the crucial problem in the first challenge of this task of how to extract the important information from the linguistically richer news content, which also tend to be of longer size. The second challenge involves the selection of a series of images corresponding to the news content. Since the images are to be incorporated in the same article, their contents should be consistent according to the theme of the article, but, at the

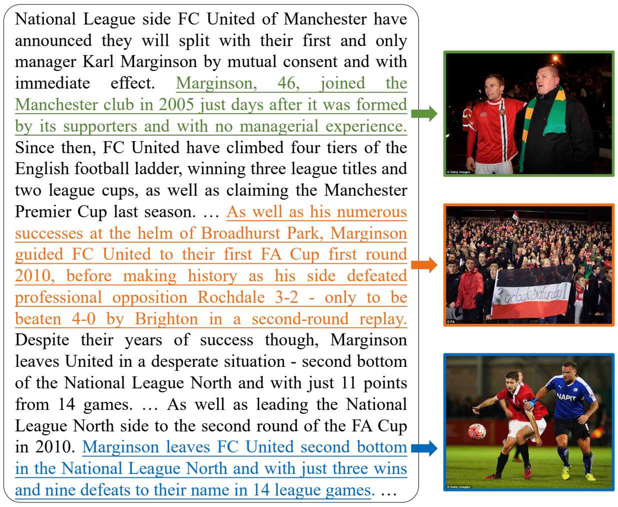

National League side FC United of Manchester have announced they will split with their first and only manager Karl Marginson by mutual consent and with immediate effect. Marginson, 46, joined the Manchester club in 2005 just days after it was formed by its supporters and with no managerial experience. Since then, FC United have climbed four tiers of the English football ladder, winning three league titles and two league cups, as well as claiming the Manchester Premier Cup last season. … As well as his numerous successes at the helm of Broadhurst Park, Marginson guided FC United to their first FA Cup first round 2010, before making history as his side defeated professional opposition Rochdale 3-2 - only to be beaten 4-0 by Brighton in a second-round replay. Despite their years of success though, Marginson leaves United in a desperate situation - second bottom of the National League North and with just 11 points from 14 games. … As well as leading the National League North side to the second round of the FA Cup in 2010. Marginson leaves FC United second bottom in the National League North and with just three wins and nine defeats to their name in 14 league games. …

Figure 1: Illustration of multiple news image selection task. Given a news article (left), our model extracts the key information (with lighter colors) and selects context-related images (indicated by the arrows).

[*] Corresponding authors.

same time, be sufficiently different to enrich the story from multiple aspects.

To address the aforementioned challenges, we propose a novel end-to-end two-stage image selection framework that innovatively integrates the processes of identifying salient contexts and selecting images. In the first stage of this framework, our model takes articles as inputs and processes them to extract key information. A state-of-the-art transformer language model with byte-pair-encoding is utilized to generate local embeddings of each possible image insertion location and address the importance of the image-related information. Considering the significance of a given location at some point in the article with respect to its entire article content, we use attention as the pointer to select the suitable image insertion locations. Regarding the second stage of our framework, we jointly consider the prior local embeddings of image locations and the global representation of the whole article as the context embedding of the image to simultaneously and smoothly reflect image discrepancy and commonness. Finally, we utilize the obtained context embeddings as the text part to calculate the similarity to the visual modality.

Three insertion strategies are designed to bridge the aforementioned stages. These strategies represent different scenarios influencing the image selection procedure. First, we introduce the coverage mechanism (Xu et al. 2015) to formalize the influence of a previously extracted image location on the present location. Next, we explicitly indicate the image discrepancy in the image selection stage and reformulate this stage as a two-stream architecture by measuring the difference between the present and previously selected image embeddings. Finally, we leverage the information of the previously selected image as input of the present image selection step to enhance the relation of the first stage to the output of the second stage.

The contributions of this work are as follows:

- We propose a novel news image selection task to jointly select multiple images for a given news article, which is challenging as it requires to determine *where* to place images and *which* images to use.
- We present an end-to-end two-stage architecture, that can effectively tackle the *where-which* problem, that is, locating the appropriate image insertion positions and selecting the corresponding images. We also introduce three insertion strategies to enable the seamless connection of these two stages and achieve better multiple image selection procedure.
- We conduct extensive comparative studies to evaluate the importance of extracting key information from news articles for multiple image selection.

The rest of this paper is organized as follows. First, the related work is reviewed in Section ; then, the proposed end-to-end two-stage multiple news image selection framework with three insertion strategies is presented in detail in Section . Next, comparative studies between the proposed framework and previously reported state-of-the-art image selection models are conducted in Section . Finally the conclusions are drawn in Section .

Related Work

Cross-Modal Retrieval

Automatic cross-modal retrieval has recently received increasing attention as a result of the advances in both computer vision and natural language processing. Modeling the similarity between image and text can be regarded as a classification problem, which directly answers whether two input samples match, represented as sm-LSTM (Huang, Wang, and Wang 2017), CMPM (Zhang and Lu 2018), MCB (Fukui et al. 2016), just to name a few. These approaches typically achieve rapid convergence in the training process, but performs poorly in exploiting the identity information of cross-modal features on account of its use of simple match/mismatch classifiers. On the other hand, embedding-based methods (Ben-Younes et al. 2017; Lee et al. 2018; Niu et al. 2017) project multimodal features into a common embedding space in which instance similarity is measured by conventional cosine or Euclidean distance. The latest state-of-the-art work, i.e., MTFH (Wang et al. 2019), follows a fully convolutional strategy with two-branch tensor fusion, in which visual and textual information are extracted separately by deep CNNs and RNNs. Despite excellent performance of current state-of-the-art models, however, cross-modal retrieval remains a challenging problem.

In fact, the current methods tend to deal with repetitive and relatively simple sentences (Lin et al. 2014; Young et al. 2014) written in a fairly consistent style. Hence, they are generally limited in terms of describing visual contents and offering deeper semantic interpretations. By contrast, our model leverages the linguistically rich texts and abundant information of news contents. Specifically, it utilizes an end-to-end two-stage framework designed to select multiple images and place them into appropriate locations jointly.

Context-Based Multimodal Research

Multimodal research has gained a certain level of popularity among scholars and consistently aims to incorporate contextual information. The latest attempts to process rich human sentences center on gathering new datasets that may be representative of different writing styles (Mathews, Xie, and He 2016; Gan et al. 2017; Mathews, Xie, and He 2018). In (Biten et al. 2019) and (Tran, Mathews, and Xie 2020), they collect the news articles from New York Times website and leverage contextual information to produce captions. In (Zhu et al. 2018), multimodal summarization with multimodal output is conducted on the dataset collected from Daily Mail website. For the image selection task, Upgrading the Newsroom(Liu et al. 2020) jointly consider the components of the articles for assisting selecting a single image. In addition, (Hessel, Lee, and Mimno 2019) discover relationships between images and sentences without relying on any explicit multimodal annotation.

In contrast to these methods, our approach analyzes text recognizing suitable image locations, which makes our model capable of considering different contexts of images. Thus, the images obtained by our framework are both text-related and context-independent.

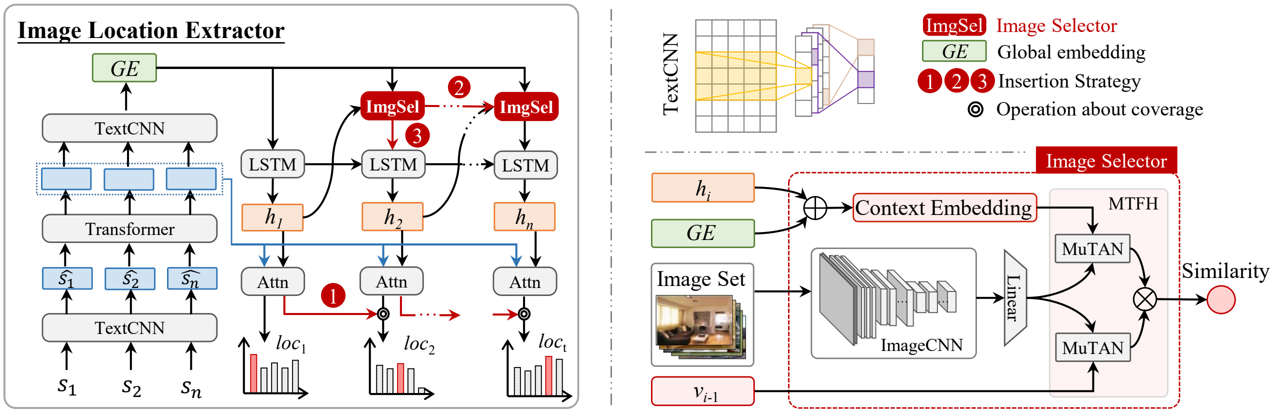


Figure 2: Illustration of the proposed approach for news image selection. In the image location extractor, we first encode articles at the sentence level via CNN. The transformer is then utilized to generate the global embedding of the text, and the Pointer Network is used to extract the appropriate image insertion locations with the local embeddings. Subsequently, the image selector (abbreviated to ImgSel) is used to select images from the gallery according to the context embedding, which is composed by the global and local embeddings obtained from the image location extractor. The insertion strategies are respectively indicated by the red lines and numbers, which represent the coverage mechanism, two-stream similarity and image pointer.

Proposed Method

As illustrated in Figure 2, our model for news image selection consists of two consecutive stages. In the first stage, the model looks through the complete text of a given article, to identify key information and indicate where the images should be inserted. The subsequent stage selects appropriate images to insert into these locations with the help of a multimodal tensor fusion network.

The first stage of our model takes news articles as inputs and encodes them completely. Due to the hierarchical structure of the text, unlike (Liu et al. 2020), which uses word-level encoding, we encode articles at the sentence level because semantic similarity is easier to be established at this granularity (Biten et al. 2019). Then, our model employs the transformer (Vaswani et al. 2017) to produce the embeddings of each sentence location. These embeddings are used to extract important information in the articles by recognizing whether an image should be inserted into a particular location. In addition, a global text embedding is generated by combining all of the location embeddings via a CNN. The Pointer Network (Vinyals, Fortunato, and Jaitly 2015) is utilized in this stage to select image locations step by step and ensure the coherence of images in the same article. The hidden state of each time step is treated as the local context embedding of each possible location for image insertion.

Subsequently, in the second stage, considering the text-related and context-independent characteristics of images in articles, global text embeddings are used along with the local embeddings of the extracted image insertion locations to select images. We obtain the composed context embedding by merging these two embeddings as the text input. A multimodal tensor fusion network (Ben-Younes et al. 2017) with a fully convolutional layer is then used to calculate the similarity between the inputs of different modalities.

Additionally, three insertion strategies are designed to reflect different scenarios influencing the procedure of image selection. In the first strategy, we introduce the coverage mechanism (Xu et al. 2015) to consider the influence of a previously extracted image location on the current location in the first stage. We define the image similarity score calculated between the current and previous image embeddings in the second strategy. The image selection stage then becomes a two-stream architecture that can explicitly represent difference between images. We enhance the relation between these two stages by utilizing the third strategy that employs the selected image information from the previous in the current time step of the location decoder in the first stage.

Problem Definition

Let $A = \{s_i\}_{i=1}^{N}$ denote an article and $I = (v, loc)$ be the corresponding image, where v and loc are raw visual and location information describing the image. Given a query consisting of a single article A, the goal of news image selection is to find a series of images $\{I\}$ with the highest relevance. In the first stage of our model, we design the Image Location Extractor to gradually extract the appropriate image insertion locations. In the subsequent stage, we define a similarity function $S(I, T)$ to, ideally, assign high similarity scores between the images and their corresponding contexts.

Image Location Extractor

In this stage, our model looks through the whole article and recognizes appropriate image insertion locations. In the encoder part of this stage, inspired by the state-of-the-art for semantic textual similarity tasks (Arora, Liang, and Ma 2017), we use sentence-level encoding to represent the news articles in our model because domain, purpose and context are better preserved at the sentence level (Biten et al. 2019). Formally, $\{w_0, w_1, ..., w_n\}$ is the i^{th} sentence of the article and w_k is the word vector obtained from the pre-trained RoBERTa (Liu et al. 2019) model, a recent improvement over the popular BERT (Devlin et al. 2019) model. To obtain the sentence-level features for the article, we apply a convolutional model (Kim 2014) to generate the sentence representation s_i:

$$w_{1:n} = w_1 \otimes w_2 \otimes ... \otimes w_n, \tag{1}$$

$$s_i^k = f(W_s \cdot w_{k:k+d-1} + b_s), \tag{2}$$

$$s_i \triangleq [s_i^1, s_i^2, ..., s_i^{n-d+1}], \tag{3}$$

where $\otimes$ is the concatenation operator, and s_i^k is the k-th word of the sentence with length n. The convolution filter W_s and the bias term b_s are applied to a window of d words to produce feature s_i^k. The feature map of the sentence s_i is composed of the features produced from each possible window of words. We then apply a max-overtime pooling operation (Collobert et al. 2011) over the feature map and take the maximum value $\hat{s_i} = max\{s_i\}$ as the feature corresponding to this particular filter.

A bidirectional transformer (Vaswani et al. 2017) is used to generate the local embeddings of each location loc_i between sentences. In addition, a global text embedding GE is generated by combining all location embeddings via the same CNN layer described above with a different convolution filter W_{GE} and bias term b_{GE}.

In the decoder part of this stage, we employ the Pointer Network (Vinyals, Fortunato, and Jaitly 2015) to recognize appropriate image insertion locations:

$$u_{loc}^t = W_p^T tanh(W_{p1} e_{loc} + W_{p2} h_t), \tag{4}$$

$$P(h_t | h_1, ..., h_{t-1}, loc) = \text{softmax}(u_{loc}^t), \tag{5}$$

where W_p, W_{p1}, and W_{p2} are learnable parameters, and e_{loc} is the output of the glimpse operation (Vinyals, Bengio, and Kudlur 2015). h_t is the hidden state of time step t, which is regarded as the local context embedding of the image insertion location extracted in this time step.

Image Selector

In this stage, our model selects appropriate images to fill the extracted locations with the help of a Multimodal Tucker Fusion network (Ben-Younes et al. 2017). For each image insertion location, we produce the context embedding C of the image by merging the global text embedding GE and the local embedding h_t of the location extracted in time step t:

$$C_t = (1 - \lambda_h) GE + \lambda_h h_t = (1 - \lambda_h) F_{\text{CNN}}(loc_{1:N}) + \lambda_h h_t, \tag{6}$$

where λ_h is a hyperparameter reflecting the balance between the global and local embeddings. The similarity score $S_{IC}(I, C)$ is then calculated by the context embedding and the image embedding through a fusion network:

$$m = \sum_{r=1}^{R} (W_{\tilde{I}}^r \tilde{I}) \odot (W_{\tilde{C}}^r \tilde{C}), \tag{7}$$

$$S_{IC}(I, C) = \sigma(W_m m), \tag{8}$$

where $W_{\tilde{I}}$, $W_{\tilde{C}}$, and W_m are learnable parameters, $\odot$ denotes the element-wise product in matrices, R is the number of fusion spaces, and σ is the sigmoid function.

Insertion Strategy

After inserting an image to represent part of the article content, the next image and its insertion location may be influenced by the previous choice. Thus, the order of image insertion is of key importance. We introduce three insertion strategies to realize different approaches to handle the influence of the previous image selection step on the next one:

Coverage Mechanism. In the first strategy, we consider the location-level situation, in this case, when an image is inserted into the article, the next image is usually not expected to be placed in another location with a similar context (especially those previously extracted image locations). In other words, we expect the previous image insertion location can influence the decision of choosing next image location. We modify the Pointer Network by introducing the coverage mechanism (Xu et al. 2015), so that each time step of the decoder could be aware of the previous time step's attention distribution of locations:

$$a_{loc}^t = \text{softmax}(u_{loc}^t), c_t = \sum_{t'=0}^{t-1} a_{loc}^{t'}, \tag{9}$$

$$u_{loc}^t = W_p^T \tanh(W_{p1} e_{loc} + W_{p2} h_t + W_c c_t), \tag{10}$$

where W_c is a learnable parameter.

Two-stream Similarity. When it comes to the second strategy, we consider the requirement that the currently selected image should be different from the previous one. We calculate the image similarity score $S(I, I)$ from the current and previous image embeddings. The image selection stage then becomes a two-stream architecture.

$$\begin{aligned} S &= S_{IC}(I, C) - \lambda_S S_{II}(I, I_{t-1}) \\ &= S_{IC}(I, C) - \lambda_S \sigma(W_{m'} \sum_{r=1}^{R} (W_{\tilde{I}}^r \tilde{I}) \odot (W_{\tilde{I'}}^r \tilde{I}_{t-1})), \end{aligned} \tag{11}$$

where $W_{m'}$, $W_{\tilde{I}}$, and $W_{\tilde{I'}}$ are learnable parameters, σ is the sigmoid function, and λ_S is a hyperparameter that indicates the influence of the image similarity score.

Image Pointer. The third strategy is that we consider that the information of the previously selected image may influence the next image insertion location. Therefore, besides the local embedding of the previous location, we utilize the selected image information from the previous to the present time step of the decoder:

$$u_{loc}^t = W_p^T tanh(W_{p1} e_{loc} + W_{p2} h_t + W_c c_t + W_I I_{t-1}), \tag{12}$$

where W_I is a learnable parameter.

Optimization

As our model is composed of two stages, we define below the loss functions for each of them. For the image location extractor, the loss function is calculated as follows:

$$L_{ILE} = -\sum \text{log} P(h, loc), \tag{13}$$

where $P(h, loc)$ is the probability distribution of multimodal blocks in the document, which is calculated by Equation (5).

In the image selector, for each positive pair of an image and a context (I_p, C_p), we additionally sample their hardest negatives which are given by $I_h = \mathrm{argmax}_{h \neq p} S_{IC}(I_h, C_p)$. Then, we calculate the loss function as follows:

$$L(I_p, C_p) = [\alpha - S_{IC}(I_p, C_p) + S_{IC}(I_h, C_p)]_+, \quad (14)$$

where α is a constant value of the margin, and the operator $[z]_+ = \max(0, z)$ compares the tolerance value with zero.

Furthermore, considering that the selected images should be both coherent yet diverse, we introduce a novel loss function for news image selection:

$$L_{IS} = L(I_p, C_p) - \lambda_S \sum [\alpha - S_{II}(I_p, I_q) + S_{IC}(I_p, C_p)]_+, \quad (15)$$

where $S_{II}(I_p, I_q)$ denotes the image similarity score between the selected images for the same news contents and λ_S is the hyperparameter used in Equation (11).

Finally, we utilize the linear combination of these loss functions as the final function, which can be optimized with Adam optimizer (Kingma and Ba 2014):

$$L = L_{ILE} + L_{IS}. \quad (16)$$

Experiments

Datasets and Implement Details

We conduct experiments on two real-world datasets: (1) NYTimes800k (Tran, Mathews, and Xie 2020) consists of 444,914 articles collected from the New York Times public API[1]. We use the same data split setting as in (Tran, Mathews, and Xie 2020), that is, the training and validation splits contain 433,561 and 2,978 articles, respectively. We report results of the test set with 8,375 articles. (2) MSMO (Zhu et al. 2018) contains 307,993 articles collected from the Daily Mail website[2]. However, it lacks any information about how the documents are organized, such as image locations. Therefore, we extend this dataset by collecting the image locations of each document according to the URLs that the authors provided. The number of documents in our training and validation splits are 287,467 and 10,265, respectively[3]. Similar to (Zhu et al. 2018), we compute the results on the test set with 10,261 articles.

For the visual feature representation, we employ VGG19 (Simonyan and Zisserman 2015) to extract the CNN features for 49 regions. Through the global average pooling on the feature map, an image can be represented by a 2048-dimensional global feature vector. As to the textual feature representation, we employ the pre-trained RoBERTa model to generate 768-dimensional word embeddings. The size and the number of the transformer encoder are 512 and 1, respectively, with a dropout of 0.1. In the decoder of the image location extractor, the dimension and number of the pointer network layers are both equal to those of the transformer encoder. We train our model using Adam optimizer with a mini-batch size of 16 for 50 epochs on each dataset. The initial learning rate is 0.0001, decayed by 2 every 10 epochs.

	NYTimes800k	MSMO
Article	444,914	307,993
Image	792,971	2,004,848
AvgArtSent	31	30
AvgArtToken	974	957
AvgImage	2	6

Table 1: Characteristics of the experimental datasets: AvgArtSent, AvgArtToken, and AvgImage respectively represent the average number of the sentences, tokens and images per news article.

Baselines and Evaluation Metrics

We compare our model with several state-of-the-art models, including classification-based methods: sm-LSTM (Huang, Wang, and Wang 2017), which utilizes a selective multimodal Long Short-Term Memory network (sm-LSTM) for instance-aware image selection; CMPM (Zhang and Lu 2018), which learns image-text embeddings via the cross-modal projection matching and cross-modal projection classification for image selection; and embedding-based methods: MTFH (Wang et al. 2019), which explicitly learns an image-text similarity with rank-based tensor fusion instead of seeking a common image-text embedding space for image selection. Newsroom (Liu et al. 2020), which is equipped with char-level word embeddings and adopts a hierarchical self-attention mechanism to attend to both key words within a piece of text and informative components of a news article.

We conduct three kinds of news image selection tasks: (1) *single image selection*, i.e., retrieving a ground truth image when given a query that consists of a single text; (2) *location prediction*, i.e., predicting the ground truth image locations when given a query comprising a single news content; (3) *multiple image selection*, i.e., selecting multiple ground truth images when given a query with a single news content.

The commonly used evaluation metric for image retrieval tasks is R@{1, 5, 10}, which is defined as the recall rate at top 1, 5, and 10. We use these metrics in the single image selection task. On the other hand, since the multiple image selection is proposed in this work, we reformulate the recall rate for multiple image evaluation. In particular, for the tasks of location prediction and multiple image selection, we respectively count the pieces of news that are accurately predicted with 1, 2, or all image locations to calculate R@1 for location prediction and R@10 for multiple image selection. We represent these recall rate as R/{1, 2, N} in Table 2. Similar to (Wang et al. 2019), we average the results by folding test sets into news samples with 1,000 test images.

Image Selection

Table 2 shows the overall image selection results of the compared approaches on the MSMO and NYTimes800k datasets. We make the following observations:

Our model achieves competitive performance for both tasks on two datasets. It indicates that our proposed frame-

[1] https://developer.nytimes.com/apis

[2] http://www.dailymail.co.uk

[3] Several websites were missing when we crawled the data, and thus, our training and validation datasets were slightly different from the original MSMO dataset.

		Single Image			Location			Multiple Image		
NYTimes800k dataset	Approaches	R@1	R@5	R@10	R/1	R/2	R/N	R/1	R/2	R/N
	sm-LSTM (Huang, Wang, and Wang 2017)	5.8	16.0	29.6	-	-	-	58.0	24.2	8.0
	CMPM (Zhang and Lu 2018)	6.3	18.9	31.0	-	-	-	59.9	26.3	8.6
	MTFH (Wang et al. 2019)	8.8	20.8	35.6	-	-	-	61.3	30.4	8.9
	Newsroom (Liu et al. 2020)	9.2	22.7	36.9	-	-	-	65.6	33.1	9.5
	Ours: base model	11.8	23.9	38.7	95.9	64.8	45.5	68.5	36.1	11.3
	Ours: base + coverage	11.9	24.1	39.3	97.6	68.2	51.4	70.1	37.8	11.7
	Ours: base + coverage + image pointer	**12.6**	**24.7**	**39.6**	**98.8**	**70.7**	**52.4**	**70.9**	37.4	11.8
	Ours: base + cov. + img. ptr. + similarity	11.5	24.2	39.0	96.9	69.4	50.9	69.2	**37.9**	**11.9**

		Single Image			Location			Multiple Image		
MSMO dataset	Approaches	R@1	R@5	R@10	R/1	R/2	R/N	R/1	R/2	R/N
	sm-LSTM (Huang, Wang, and Wang 2017)	3.9	15.1	28.6	-	-	-	54.7	21.2	5.4
	CMPM (Zhang and Lu 2018)	3.2	19.0	30.1	-	-	-	55.0	22.3	6.5
	MTFH (Wang et al. 2019)	4.6	19.9	32.5	-	-	-	58.7	25.8	6.9
	Newsroom (Liu et al. 2020)	5.1	20.8	33.0	-	-	-	60.5	27.7	8.7
	Ours: base model	8.6	21.3	36.8	95.5	43.2	30.6	66.4	31.7	9.3
	Ours: base + coverage	9.1	**21.6**	36.9	96.9	45.7	35.5	67.8	31.9	9.6
	Ours: base + coverage + image pointer	**9.8**	21.5	**37.4**	**97.0**	**46.0**	**36.1**	**67.9**	31.8	9.4
	Ours: base + cov. + img. ptr. + similarity	8.9	21.4	36.5	96.6	45.4	35.7	67.5	**32.2**	**9.9**

Table 2: Image selection results on NYTimes800k and MSMO: three image selection tasks are tested (i.e., single image selection, location prediction, and multiple image selection), and the best results under different metrics in each task are marked in bold (pairwise t-test at 5% significance level). Our base model contains the image location extractor (assembling the Transformer with the Pointer Network) and the image selector. Based on the base model, we gradually add three insertion strategies.

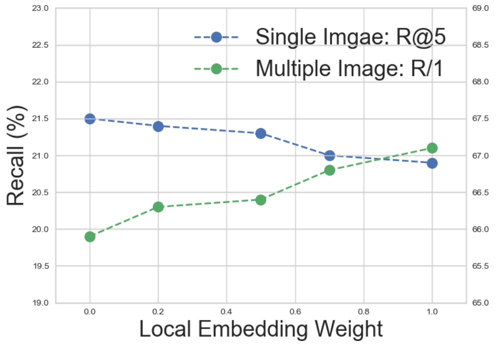


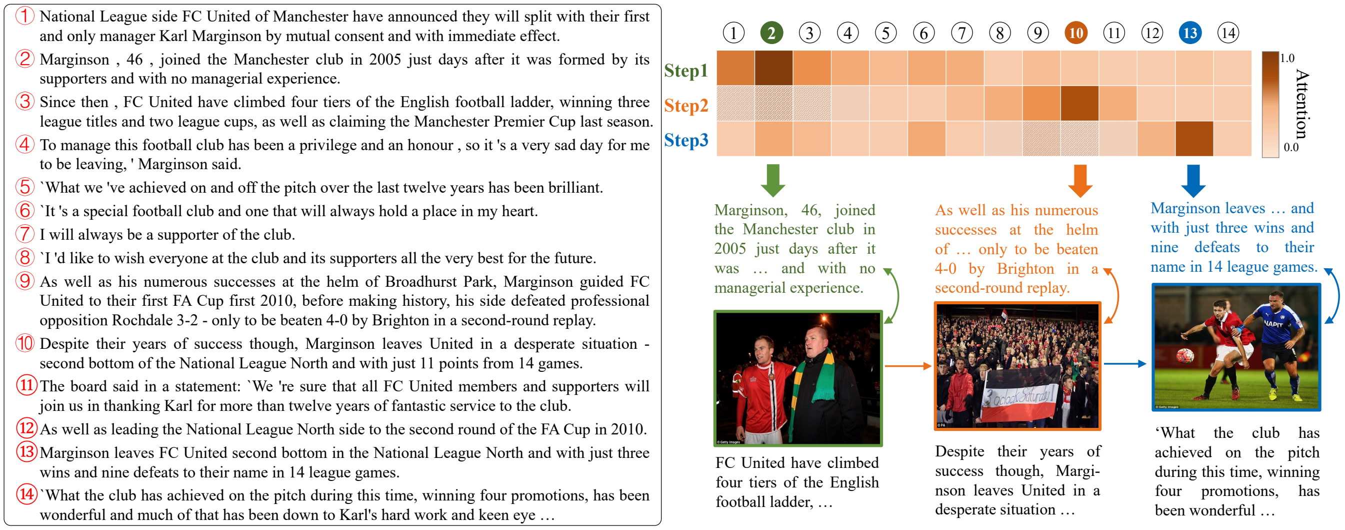
Figure 3: Performance of our model in single and multiple image selection tasks when varying local embedding weight.

work possesses the superior image selection ability to fully encode the interactions between images and news articles.

The improvements on the news image selection tasks (i.e., location prediction and multiple image selection) by our model is more remarkable than that in the single image retrieval task, indicating the advanced ability of extracting key information from the image location extractor in our proposed two-stage architecture, hence improving the accuracy of the multiple image selection. We notice that the overall performance on NYTimes800k (including ours and the compared approaches) is better than that on MSMO. The reason for this may be due to the fact that there are more images per news articles in the MSMO dataset (as revealed in Table 1). The proportion of news articles which contain more than one image in the NYTimes800k dataset is 41%, while this proportion on the MSMO dataset is 83%. More content-related images may confuse the model, which results in a lower performance. Besides, compared to the R/1 results in the multiple image selection task, R/2 and R/N results are not good enough, which indicates that selecting multiple images for a given news article still remains a challenging task.

Besides, we investigate the effects of different proportion of global and local context embeddings. We gradually increase the balance parameter λ_h in Equation (6), which indicates the proportion of global and local embeddings in our image selector, and implement the tasks of single and multiple image selection. As shown in Figure 3, the multiple image selection performance is positively correlated with the local embedding weight, that is, a larger weight will yield better multiple image selection. However, the case of single image selection is just opposite. Intuitively, when we increase the proportion of local embedding, the difference between context embeddings for a certain article also increases along with it, which can guide the model to find more different images for each context. As a result, the model is more likely to find suitable images, while suffers from an increasing risk of unsatisfying precise selection requirements.

We provide in Figure 4 an example of result for the best performing model. In this example, we take a news article with 14 sentences as the input, and we gradually determine where to place images. We extract the image insertion location with the highest attention score at each step.



1 National League side FC United of Manchester have announced they will split with their first and only manager Karl Marginson by mutual consent and with immediate effect.
2 Marginson , 46 , joined the Manchester club in 2005 just days after it was formed by its supporters and with no managerial experience.
3 Since then , FC United have climbed four tiers of the English football ladder, winning three league titles and two league cups, as well as claiming the Manchester Premier Cup last season.
4 To manage this football club has been a privilege and an honour , so it 's a very sad day for me to be leaving, ' Marginson said.
5 `What we 've achieved on and off the pitch over the last twelve years has been brilliant.
6 `It 's a special football club and one that will always hold a place in my heart.
7 I will always be a supporter of the club.
8 `I 'd like to wish everyone at the club and its supporters all the very best for the future.
9 As well as his numerous successes at the helm of Broadhurst Park, Marginson guided FC United to their first FA Cup first 2010, before making history, his side defeated professional opposition Rochdale 3-2 - only to be beaten 4-0 by Brighton in a second-round replay.
10 Despite their years of success though, Marginson leaves United in a desperate situation - second bottom of the National League North and with just 11 points from 14 games.
11 The board said in a statement: `We 're sure that all FC United members and supporters will join us in thanking Karl for more than twelve years of fantastic service to the club.
12 As well as leading the National League North side to the second round of the FA Cup in 2010.
13 Marginson leaves FC United second bottom in the National League North and with just three wins and nine defeats to their name in 14 league games.
14 `What the club has achieved on the pitch during this time, winning four promotions, has been wonderful and much of that has been down to Karl's hard work and keen eye …

Marginson, 46, joined the Manchester club in 2005 just days after it was … and with no managerial experience.

As well as his numerous successes at the helm of … only to be beaten 4-0 by Brighton in a second-round replay.

Marginson leaves … and with just three wins and nine defeats to their name in 14 league games.

FC United have climbed four tiers of the English football ladder, …

Despite their years of success though, Marginson leaves United in a desperate situation …

'What the club has achieved on the pitch during this time, winning four promotions, has been wonderful …

Figure 4: An example news article (left) and the corresponding images with locations (right) from the MSMO test set: in the location prediction stage (steps 1-3), we extract the locations with the highest attention scores measured by our location extractor; after that, we pick and insert the most context-relevant images into these locations in the image selection stage.

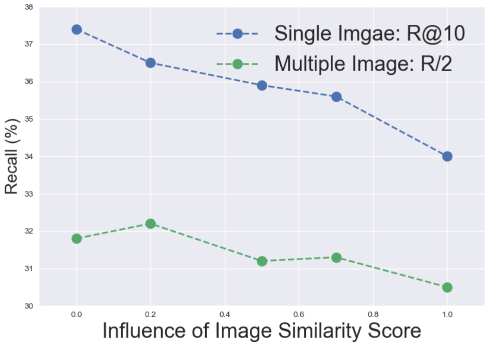


Figure 5: Performance of our model in single and multiple image selection tasks when varying the image similarity score.

Subsequently, we select the most relevant images from the gallery according to the context surrounding these locations. Finally, we insert these images into their locations to complete news content.

Analysis of Insertion Strategies

When combined with the proposed three insertion strategies (i.e., coverage mechanism, image pointer, and two-stream similarity score), our model gains remarkable improvements compared with the basic two-stage model and achieves the state-of-the-art performance in most cases, as illustrated in Table 2. The main reason is that these strategies exploit the sequence information in the image insertion process by utilizing the part of results from the previous time steps.

Compared to other strategies, when composing the coverage mechanism, the performance improvement is evident. Considering its ability of eliminating repetition on the time steps of the decoder, we use it to guide a better image location extraction, aiming to lower the attention scores of the previously extracted image locations and their context-similar locations. Its performance has been validated on almost all of the evaluation metrics in Table 2. In terms of the image pointer, it enables the image location extractor to obtain the previous selected image, which can also improve the performance of location extraction.

As to the two-stream similarity score, its effects in single image selection and location prediction are not as significant as in multiple image selection, as shown in Table 2. To further investigate its effects, we take the MSMO dataset as testbed and tune the balance parameter λ_S in Equation (11) to assess the performance of our model with this strategy. From the results in Figure 5, we notice that considering the image discrepancy can greatly promote multiple image selection, despite its limited contribution to single image selection. We assume this is attributed to the intrinsic property of multiple image selection, that is, the selected images should be consistent in content while distinctive to enrich the story (as described in Section).

Conclusion

In this paper, we propose a novel news image selection task, which consists of selecting multiple images for the news content with linguistically rich text and abundant information. We propose an end-to-end two-stage architecture, that allows to identify key information of the articles by extracting important image insertion locations and to select appropriate images. We further introduce three insertion strategies for better multiple image selection. Experimental results demonstrate that the proposed method achieves the state-of-the-art performance, and they also indicate the importance of extracting key information from the articles and the benefits of the insertion strategies.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Shandong Province under grant ZR2019MA049 and the Ministry of Education of Humanities and Social Science project under grant 16YJC790123.

References

Arora, S.; Liang, Y.; and Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR 2017*.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV 2017*, 2612–2620.

Biten, A. F.; Gomez, L.; Rusinol, M.; and Karatzas, D. 2019. Good News, Everyone! Context Driven Entity-aware Captioning for News Images. In *CVPR 2019*, 12466–12475.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural Language Processing (almost) from Scratch. *JMLR* 12: 2493–2537.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*, 4171–4186.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *EMNLP 2016*, 457–468.

Gan, C.; Gan, Z.; He, X.; Gao, J.; and Deng, L. 2017. Stylenet: Generating Attractive Visual Captions with Styles. In *CVPR 2017*, 3137–3146.

Hessel, J.; Lee, L.; and Mimno, D. 2019. Unsupervised Discovery of Multimodal Links in Multi-image, Multi-sentence Documents. In *EMNLP 2019*, 2034–2045.

Huang, Y.; Wang, W.; and Wang, L. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *CVPR 2017*, 2310–2318.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014*, 1746–1751.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* .

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked Cross Attention for Image-Text Matching. In *ECCV 2018*, 201–216.

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 *ECCV 2014*, 740–755.

Liu, F.; Lebret, R.; Orel, D.; Ordet, P.; and Aberer, K. 2020. Upgrading the Newsroom: An Automated Image Selection System for News Articles. *TOMM 2020* .

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:1907.11692* .

Mathews, A.; Xie, L.; and He, X. 2016. SentiCap: Generating Image Descriptions with Sentiments. In *AAAI 2016*, 3574–3580.

Mathews, A.; Xie, L.; and He, X. 2018. Semstyle: Learning to Generate Stylised Image Captions using Unaligned Text. In *CVPR 2018*, 8591–8600.

Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; and Hua, G. 2017. Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding. In *ICCV 2017*, 1881–1889.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR 2015*.

Tran, A.; Mathews, A.; and Xie, L. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *CVPR 2020*, 13035–13045.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *NIPS 2017*, 5998–6008.

Vinyals, O.; Bengio, S.; and Kudlur, M. 2015. Order Matters: Sequence to Sequence for Sets. In *ICLR 2015*.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer Networks. In *NIPS 2015*, 2692–2700.

Wang, T.; Xu, X.; Yang, Y.; Hanjalic, A.; Shen, H. T.; and Song, J. 2019. Matching Images and Text with Multi-modal Tensor Fusion and Re-ranking. In *ACM MM 2019*, 12–20.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML 2015*, 2048–2057.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL 2014* (2): 67–78.

Zhang, Y.; and Lu, H. 2018. Deep Cross-modal Projection Learning for Image-Text Matching. In *ECCV 2018*, 686–701.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *EMNLP 2018*, 4154–4164.