PFAN++: Bi-Directional Image-Text Retrieval with Position Focused Attention Network

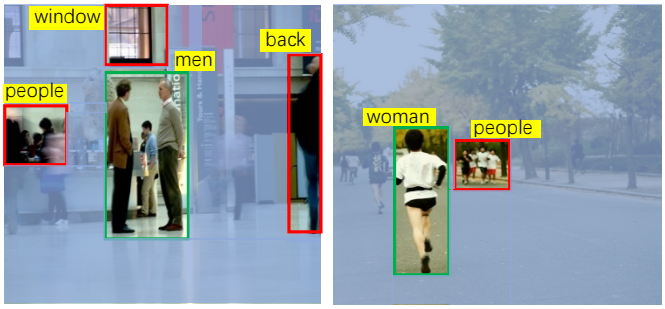
Yaxiong Wang*, Hao Yang*, Xiuxiu Bai, Xueming Qian, *Member* IEEE, Lin Ma, Jing Lu, Biao Li and Xin Fan

***Abstract*—Bi-directional image-text retrieval and matching attract much attention recently. This cross-domain task demands a fine understanding of both modalities for learning a measure of different modality data. In this paper, we propose a novel position focused attention network to investigate the relation between the visual and the textual views. This work integrates the prior object position to enhance the visual-text joint-embedding learning. The image is first split into blocks, which are treated as the basic position cells, and the position of an image region is inferred. Then, we propose a position attention to model the relations between the image region and position cells. Finally, we generate a valuable position feature to further enhance the region expression and model a more reliable relationship between the visual image and the textual sentence. Experiments on the popular datasets Flickr30K and MS-COCO show the effectiveness of the proposed method. Besides the public datasets, we also conduct experiments on our collected practical large-scale news dataset (Tencent-News) to validate the practical application value of the proposed method. As far as we know, this is the first attempt to test the performance on the practical application. Our method achieves the competitive performance on all of these three datasets.**

***Index Terms*—Image-Text Matching. Attention Mechanism. Cross-Domain. Position Embedding Learning.**

I. Introduction

With the constantly springing up of multimedia data like text, image, video on the Internet, cross-modal retrieval has attracted much attention in both computer vision and multimedia communities. Bidirectional image-text retrieval is one of the important branches for various multimedia related applications like image-text matching [5, 10], natural language object retrieval [3], image captioning [7, 9], and visual question answering (VQA) [11, 15]. Therefore, many researchers have dedicated extensive efforts to study the relationship between the visual and the textual contents [2, 4, 6, 13, 16-19, 21-23, 40, 52-57].



(a) Two men in formal wear talking next to people moving

(b) A woman in a white top runs along a tree lined path

Fig.1: Position can indicate the importance of the regions

Image and text are two most commonly used multimedia data in daily life, they both contain rich information but reside in heterogeneous modalities. Comparing to information retrieval within the same modality, the designed model for cross-modal retrieval need not only learn the features for image and text to express their respective content but a measure for cross-modal similarity calculation. Therefore, cross-modal retrieval poses extra critical challenges. Relevance estimation based on subspace learning is a popular strategy, and a classic structure for image-text matching is the two-branch network. One branch projects the image and another models the text, the shared subspace is learned by the popular triplet loss [4, 13, 17, 19, 21]. For example, Faghri et al. [21] design a two-branch network trained by their hard triplet sampling strategy. To preserve the locality of each modality, Zhang et al. [40] propose to learn a matrix based measure for cross-modal retrieval. Besides the network structure studying, more and more scholars recently design their embedding networks based on attention mechanism, which attempts to capture the correspondences between the detected visual objects and the textual items (words or phrases). Many studies have validated that the attention is helpful to model a more reliable relationship between image and text. Lee et al. [2] first detect the image objects, then they calculate the attention weights based on object features and word vectors from the visual view and the textual view respectively. Huang et al. [22] think conventional attention only considers the local information. Therefore, they propose

This work was supported in part by the NSFC under Grant 61772407 and 61732008, and National Key Research and Development Project with No: 2019YFB2102500. The conference version of this work has been accepted by IJCAI 2019.

Yaxiong Wang, is with School of Software Engineering, Xi'an Jiaotong University, Xi'an China. He is now an intern in the Department of PCG, Tencent (e-mail: wangyx15@stu.xjtu.edu.cn).

Hao Yang, is with Department of PCG, Tencent (e-mail: applehyang@tencent.com).

Xiuxiu Bai is with School of Software Engineering, Xi'an Jiaotong University, Xi'an China (corresponding author, e-mail: xiubai@xjtu.edu.cn).

Xueming Qian, is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, and SMILES LAB, Xi'an Jiaotong University, Xi'an China (co-corresponding author, e-mail: qianxm@mail.xjtu.edu.cn).

Lin Ma is with Tencent AI Lab, Shenzhen China (e-mail: forest.linma@gmail.com).

Jing Lu is with Department of PCG Tencent, Beijing China (e-mail: luckielu@tencent.com).

Biao Li is with Department of PCG Tencent, Beijing China (e-mail: biotli@tencent.com).

Xin Fan is with Department of PCG, Tencent, Beijing China (e-mail: hsinfan@tencent.com).

* Equal contributions.

an attention mechanism that takes the entire features of image and sentence into account.

However, existing corresponding learning methods only focus on the visual feature of the image regions while ignore the relative position information in the images, which is an important and helpful prior knowledge. In general, if an object region is closer to the center, it may express the main semantics of the image with higher probability, while the marginal ones may not be that important. Just as shown in Fig. 1(a), the main semantic part corresponding to the word "men" locates at the center of the image, while the peddling objects lie on the brink. From this observation, an intuitive idea is to simply pay more attention to the regions closer to the center. However, not all regions near the center are important. As shown in Fig. 1(b) which exhibits that a woman (the most important object) lies in the lower left part. Furthermore, simply assigning attention to the region based on the fixed position (the center for example) cause a bad extendibility. From above observations and considerations, we design a position feature for the region to integrate position information and propose an attention mechanism to adaptively construct the position feature for each region.

In this paper, a novel position focused attention network is developed to study the fine-grained interplay between the image regions and the words. We first generate the basic position cells and select some valuable cells to infer the position of the regions. A position attention mechanism is further proposed to distinguish the different importance of the associated position cells. Then the final position feature and the visual feature are integrated to form the final feature representation for the region. Besides the local features, the whole features of image and sentence are also introduced to compensate the semantic order information. Finally, a visual-text attention algorithm SCAN [2] is employed to calculate the local and global relevance between the image and the sentence, the overall network parameters are trained by the popular triplet loss. The contributions of this paper can be summarized as follows:

a. We design a novel position feature for image region representation, by which we integrate the position prior information to form a more reliable and complete expression of the image region.

b. A position focused attention mechanism is proposed to determine the fine-grained relationship between the image region and the position cells. Position attention can help us build a more valuable position feature for the image region.

c. Apart from the local information, such as image regions, fragments of sentence, our developed network also fuses the global characteristics of the image and the text into our relevance estimation to capture the semantic order information in the image.

d. Besides two public datasets, we make the first attempt to evaluate the application value on a practical news dataset and our method achieves the competitive performance on all of these three datasets.

Comparing to our previous work [48], this paper integrates the global characteristics to enhance the sharing subspace learning and make the performance step further. In the experiment section, much more detailed results and systematic discussions are presented to clarify the contribution of each part in our framework.

The remainder of this paper is organized as follows: In section II, we review the related works of the existing image-text matching methods. Section III elaborates the details of each process in our system. Experiments and related discussions are shown in sections IV and V respectively. Finally, conclusions and future works are given in section VI.

II. Related work

The key of the image-text matching task is to learn the similarity function between two different modalities, deep learning based method classically model the function as a two-branch network. Researchers have designed various network structures to investigate the relationship between the image and the text. Directly global similarity learning and local correspondences learning based on attention are two popular strategies. Hereinafter, each aspect is presented.

A. Global Similarity Learning

Recently, a rich line of studies has explored mapping the visual information and the textual content into a common semantic subspace to investigate the relationship between image and text [4, 6, 13, 14, 16-20, 21-26, 28, 42, 43, 45, 46]. Kiros et al. [8] make the first attempt to learn cross-view representations with a hinge-based triplet ranking loss, where the image is encoded by Convolutional Neural Networks (CNN) and sentences are encoded by the Recurrent Neural Networks (RNN). Faghri et al. [21] think that the hard training samples can make the network converge faster and learn a more reliable embedding, therefore they pay much attention to the hard triplets to learn the joint embedding of image and text. Gu et al. [4] introduce the generative loss into the cross-modality task to learn the visual-semantic subspace, which yields a significant performance improvement. In [13], Wang et al. attempt to preserve the locality of the data within the same modality, and sample the triplets from each modality to learn the image-text sharing subspace. Niu et al. [23] take the part of speech into consideration, they think the noun for a sentence is the most important component and give the priority consideration to the noun, they develop a hierarchical multimodal LSTM to encode the sentence by the aid of tree LSTM. Similar to [23], Huang et al. [18] also take the part of speech into consideration, they think that noun, verb, adjective and numeral form the main force of a sentence, which motivates them to reconstruct the sentence for more accuracy semantics learning. Zheng et al. [19] treat each image-descriptions as a category, the triplet loss incorporating the cross-entropy loss is used to update the embedding network. Zhang et al. design a 2-way network [6], which attempts to approximate the correspondences between the image and the text by the distribution of cross-domain data within the mini-batch. Plummer et al. [1] propose a strategy to simplify the representation requirements for individual embedding, and the underrepresented concepts take advantage of the shared representations to learn the joint-embedding. Different from the hinge-based triplet ranking loss that shows solicitude for the distance of the positive sample and the negative sample with respect to the anchor, CCA (Canonical Correlation Analysis) based methods aim at learning nonlinear transformations of cross-modality data by the deep networks such that the learned new representations are highly linearly correlated. Deep CCA is also a popular baseline in the cross-modality field [14, 20, 25, 26, 28]. Wang et al. [23] follow the classic CCA idea and extend a more

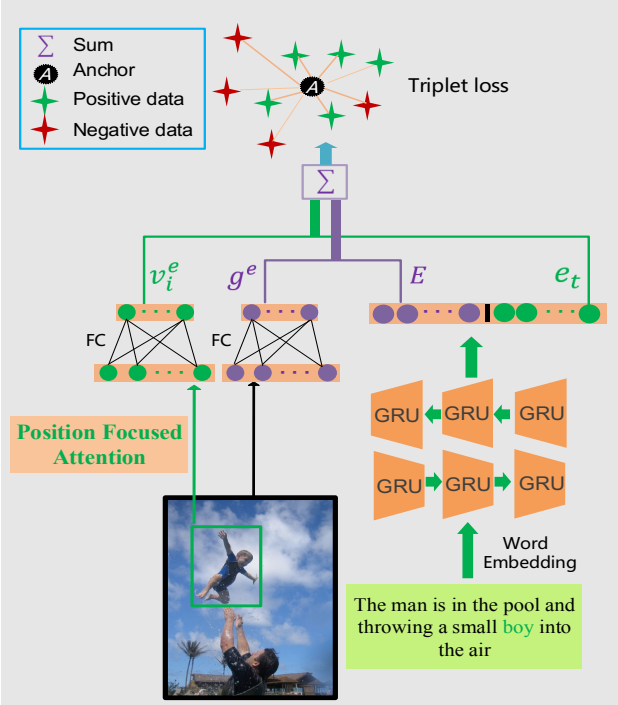


Fig.2: The flowchart of proposed PFAN++ model

stable and accurate deep CCA model. Chang et al. [12] propose a soft CCA to search an efficient solution for deep CCA optimization.

B. Local Correspondences Learning

Apart from the efforts to study the global similarity between image and text, many of the recent research works attempt to maximize the alignments between detected objects in the image and the items in sentence [2, 17, 22, 27, 29, 31-35]. As a consequence, the attention mechanism is proposed, which aims at focusing on the most valuable part of data with respect to a task-specific context. In computer vision, attention mechanism is usually designed for more accuracy correspondences between the image regions and the fragments of sentence. Lee et al. [2] design an attention mechanism from visual and textual views. It first attends to words in the sentence with respect to each image region, an attention weight is calculated for each word to indicate the importance of the current image region. Attention from text to image is designed analogously. Karpathy et al. [27] encode the image at object level with R-CNN [29], and the image-text similarity is inferred by accumulating the similarity scores of all possible region-word pairs. Instead of investigating the correspondences between the image regions and the sentence fragments, Nam et al. [17] think that the different regions of image can make responses for different convolutional kernels, and design attention based on feature maps rather than the detected image regions. In [29], Anderson et al. design a combined bottom-up and top-down attention, the bottom-up part extracts the image regions, while the top-down mechanism determines the weights for detected image objects. Li et al. [56] attempt to learn a robust feature capturing the key semantics for the image to enhance the image-text matching. Huang et al. [22] think that only considering the image regions and sentence is unilateral and introduce the whole image and the whole sentence feature into the alignment procedure to fuse the global and local information. Andreas et al. [31] employ a series of net module together with a language parser to indicate which neural net module to use. In [57], Huang et al. pay attention to the few-shot image-sentence matching and propose a gate visual-semantic embedding model. Yang et al. [34] propose a multiple-layer Stacked Attention Networks (SAN) to infer the answer for the query image. However, existing works all ignore the position clues of image regions

III. OUR APPROACH

This section will elaborate the details of our proposed framework. Fig. 2 shows the flowchart of this paper, We first extract the image and word features. Then, the designed position cells together with the position attention construct a position feature for each image region. And the visual feature together with the generated position feature form the final region's representation. Finally, the alignments between the regions and the words are studied by a visual-textual attention [2]. The similarity of the image and the sentence is estimated by the local and global similarities. We employ the triplet ranking loss to train the overall network.

Next, we describe the input representation in subsection *A*. In subsection *B*, the position information integration is presented, subsection *C* and *D* elaborate the image-sentence relevance calculation and global-local joint embedding learning respectively.

A. Input Feature Representation

Image Feature. In this paper, an image I is represented by a set of local features $\{v_1, v_2, \dots, v_n\}$ and a global feature g, where n is the number of image regions and g, v_i are both D-dimensional feature**.** Since our attention mechanism is focused on the image regions, especially the objects in image. Therefore, we detect the objects in image utilizing the Faster R-CNN model [37]. In order to get a better feature representation, we feed the detected object into the ResNet-101 [36] pre-trained on Visual Genomes [38] by Anderson et al. [29] to extract the visual feature. Finally, the input image is represented by n D-dimensional feature vectors, which are the local features. The global representation is also extracted by the pre-trained model [29], i.e. g is a feature with D-dimensional as well. D is 2048 in our experiment.

Text Feature: On the subject of corresponding image description, the basic item is the word in the sentence. Each word is represented with a one-hot vector, which indicates the index in the vocabulary. Then the one-hot representation is embedded into d-dimensional vector by a linear mapping layer, $x_t = W \times w_t$, where w_t is a one-hot of word in a sentence with T words $\{w_1, w_2, \dots, w_T\}$, $W \in R^{d \times N}$ is the embedding matrix, N is the vocabulary size.

B. Position Information Integration

The relative position of the object in the whole image is an important and useful clue, which is helpful to infer the significance of the object region, just as shown in Fig. 1. Motivated by this observation, we fuse the position information into the learning procedure to capture more reliable and credible fine-grained interplay between the image and text elements. In order to generate a valuable position feature for the region, a position attention is proposed. In this subsection, we present our position attention mechanism. We first introduce the initial positional representation in part 1) and elaborate the block embedding in part 2), part 3) presents our position focused attention.

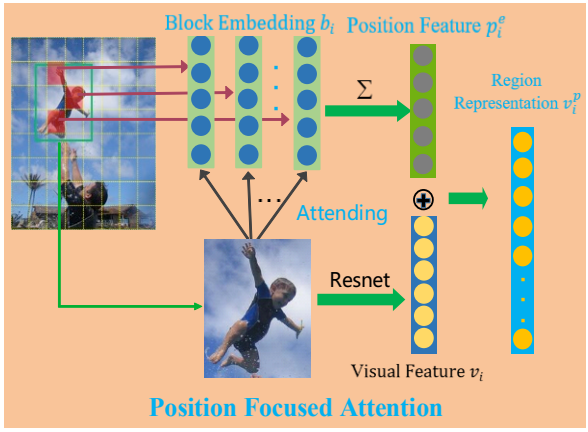


Fig.3: The proposed position focused attention mechanism

1) Initial Position Representation

Given an image $I = \{v_1, v_2, \dots, v_n\}$, in order to reveal the relative position for a region v_i in the whole image I, we first equally split the image I into $K \times K$ blocks B and treat each block as a basic position cell. The position of each block is initially represented by an index $k \in [1, K^2]$, we locate the region v_i according to its overlap with the fixed blocks. Let $p_i \in R^L$ denote the position index vector of region v_i, which is defined as the indexes of the top L overlapping blocks with the region v_i, i.e. the indexes in p_i meet:

$$OV\left(v_i, b_{p_{ij}}\right) \geq OV\left(v_i, b_q\right),\ j = 1,2, \dots, L \quad (1)$$

where $p_{ij} \in [1, K^2]$ is the block index of the j-th maximum overlapping with the region v_i, $q \in [1, K^2] \backslash p_i$, the operator "\" means removing, and $OV\left(v_i, b_q\right)$ is the intersecting pixel number between region v_i with the q-th block:

$$OV\left(v_i, b_q\right) = |v_i \cap b_q| \quad (2)$$

where $b_q \in B$ is the q-th block. We also define an additional vector $a_i \in R^L$ for region v_i to record the corresponding overlapping to distinguish the importance of different positions:

$$a_{ij} = OV(v_i, b_{p_{ij}}) \in R \quad (3)$$

and a_{ij} is normalized for further processing.

2) Block Embedding

L indexes of blocks are introduced to indicate the relative position of the region. To get a more accurate description for the position, we embed the block index into a dense representation. The split blocks B are regarded as the position vocabulary, and each block $b_i \in B$ is represented by the one-hot vector, which indicates the index in the position vocabulary. We next apply an embedding layer to project the one-hot representation into ι-dimensional vector, we still denote the new embedding vector as b_i for the sake of simplicity.

We can then simply generate the position representation of region v_i based on the embedding block vector:

$$p_i^e = \sum_{j=1}^{L} b_{p_{ij}} \times a_{ij} \quad (4)$$

3) Position Focused Attention

After obtaining the block embedding, we can represent the position of region v_i according to the Eq. (4). However, it's insufficient to directly use the rate of the overlapping area. Since there are many blocks completely covered by the region and the contributions of these blocks will be equal accordingly. From this consideration, an adaptive weight is assigned to each block with respect to each region. As shown in Fig. 3, the proposed attention aims at deciding how much weight should pay to the position cell for the current region:

$$\beta_{ij} = \tanh(f(v_i, b_{p_{ij}})), i \in [1, n], j \in [1, L] \quad (5)$$

where β_{ij} is the attention that the region v_i should pay to the position cell $b_{p_{ij}}$ and f is the bilinear function:

$$f(v_i, b_{p_{ij}}) = v_i^T M b_{p_{ij}} \quad (6)$$

where $M \in R^{D \times \iota}$ is the mapping matrix.

Besides the completely covered blocks should be different, another intuition is that the more of the block is covered by the region, the more important it should be. According to above considerations, we improve the Eq. (4) accordingly as following:

$$p_i^e = \sum_{j=1}^{L} b_{p_{ij}} \times \gamma_{ij} \quad (7)$$

and

$$\gamma_{ij} = \frac{\gamma'_{ij}}{\sum_j \gamma'_{ij}}, \text{where } \gamma'_{ij} = \frac{\exp(\beta_{ij})}{\sum_j \exp(\beta_{ij})} \times a_{ij} \quad (8)$$

The final position representation of region p_i^e is then concatenated with the visual feature v_i to allow the region feature to carry position information, i.e. $v_i^p = [v_i, p_i^e] \in R^{D+\iota}$.

C. Image-Sentence Relevance Calculation

Given an image I with n regions and a sentence S with T words, we utilize a fully-connected layer to project the final region representation v_i^p into a h-dimensional feature $v_i^e \in R^h$. As for the words, the final feature is obtained by feeding the embedding vector into a bi-directional GRU [44], whose dimension of the hidden state is also set as h. The final representation e_t of the word is the average of forward and backward feature:

$$e_t = \frac{e_t^f + e_t^b}{2} \in R^h \quad (9)$$

where e_t^f and e_t^b are the forward and backward features, respectively.

Following the work in [2], an attention weight α_{it} for region v_i with respect to the word w_t is calculated, which decides how much attention to pay the region for current word. The visual vector of the current word is then defined as the weighted combination of region representation:

$$v_t = \sum_{i=1}^{n} \alpha_{it} v_i^e \quad (10)$$

The semantic relevance between the image I and the sentence S is taking the average of the relevances between all the semantic features in S and the attending visual vectors:

$$S(I, S) = \frac{\sum_t r(e_t, v_t)}{T} \quad (11)$$

where T is the number of words in the sentence, $r(\cdot,\cdot)$ is the cosine similarity function.

On the other hand, the procedure of attending image to the text is analogous to the above. The attention weights are assigned to the words with respect to each image region, and semantic vectors for regions are generated. The visual relevance of sentence and image is estimated according to Eq. (11) analogously.

D. Global-Local Joint Embedding Learning

Only region features for image representation will miss the semantic order of the whole visual information. Inspired by the work [22] that the global information can benefit the joint-embedding learning. In this paper, besides the local regions and the words, the global features are also introduced to enhance the embedding learning.

For the image *I*, we gather all the global and the local features as the complete representation of the image: $\{v_1^p, v_2^p, \dots, v_k^p\}$ and g. As for the global representation of the sentence *S*, we first embed the original one-hot vector of word, then the word sequence is fed to the bi-GRU, the last hidden state E encodes the whole semantics of the sentence and is treated as the global feature.

We record the embedding of the global visual feature as g^e and the overall similarity between image and sentence is integrated:

$$\mathcal{S}' = \lambda \mathcal{S}(I,S) + (1-\lambda)\frac{g^e \cdot E}{||g^e|| \times ||E||} \tag{12}$$

where λ is the weight scalar.

The triplet loss is employed for network training, which is a common ranking objective for image-text matching. As reported in [2] [21], the hardest negative samples can make more contribution to the convergence and the reliability of the network. Therefore, in this work, we only pay close attention to the hardest sample for a positive pair in a mini-batch by following Lee et al. [2] and Fagphri et al. [21]. Given a matching triplet pair (I,S), the hardest negative visual sample is the most similar unmatched image to sentence *S*:

$$\tilde{I} = \arg\max_{\dot{I} \in C \setminus I} \mathcal{S}'(\dot{I}, S) \tag{13}$$

where *C* is the collection of the data in a mini-batch, and the hardest negative semantic sample can be picked up in a similar way:

$$\tilde{S} = \arg\max_{\dot{S} \in C \setminus S} \mathcal{S}'(I, \dot{S}) \tag{14}$$

The loss for a mini-batch is defined as followings:

$$Loss = \frac{1}{|C|}\sum_{(I,S)\in C}\Big(\big[\eta - \mathcal{S}'(I,S) + \mathcal{S}'(I,\tilde{S})\big]_+ + [\eta - \mathcal{S}'(I,S) + \mathcal{S}'(\tilde{I},S)]_+\Big) \tag{15}$$

where $\eta \in R$ is the margin between matched and unmatched pair and $[\cdot]_+$ ensures that the output is nonnegative.

IV. EXPERIMENTS

To demonstrate the effectiveness of our proposed methods, we conduct our Position Focused Attention Network on two public datasets: Flickr30K and MS-COCO, and a practical Chinese news dataset: **Tencent-News**[1]. We denote this work as PFAN++ to discriminate with our conference version. We systematically make comparisons with several latest start-of-the-art methods and thoroughly investigate the performance of the proposed PFAN++. As for the performance measure criterion for sentence retrieval or image retrieval, we apply the commonly used recall on top H (R@H), which is defined as the percentage of correct items in the top H retrieved results.

A. Implements Details

1) Dataset

We evaluate PFAN++ on the widely used and authoritative dataset Flickr30K and MS-COCO, the data splits for these two datasets follow the work [27] and [2]. Besides the public datasets, a practical news dataset, Tencent-News, is also collected to evaluate the value of the proposed method in the practical application. We construct this dataset from the crawled Tencent News data, which can be used for training image-text models to further support Chinese corpus.

Tencent-News. For a piece of news, the title and one perfect matching image in this news make up a basic data item, and the title is regarded as the description of this image. In this way, we collect 143,317 training pairs and 1,000 pairs for validating. There are 141,736 different images, 130,230 different titles in total. In the test procedure, we manually label 510 news and several corresponding images (≥ 5) for performance evaluation. There are 510 titles and 2,794 labeled images in total. Each title has 5.5 candidate images, 2.3 irrelevant images and 3.2 relevant images on average. In this practical dataset, we focus on the news auto-image recommendation task, i.e. the news editor inputs the news title, and the model can automatically output several related candidate images for this news, which can remarkably alleviate the effort of editors and speed up the news publish. In this application scene, we only focus on the task of the image retrieval.

2) Training Details

All of our experiments are conducted on a workstation with NVIDIA Tesla GPU. Adam optimization algorithm is used to train the overall network. The mini-batch size is 128. The image region is extracted by the Faster R-CNN model [37], and we retrain 36 detected regions for the image representation, i.e. $n = 36$. The *K* is set as 16, i.e. each image is split into 16×16 blocks. We select the first 15 blocks with the maximum overlapping with a region to infer the position, i.e. $L = 15$. The dimension of joint embedding is fixed as 1024. The weight parameter λ is set as 0.5.

For the image region, the block index is first embedded into 200-dimensional space, and the original 2048-dimensional visual vector together with 200-dimensional position feature is mapped into the 1024-dimensional feature by a linear projection layer. The global image feature is feed into a fully connected layer with shape (2048, 1024) to obtain its embedding. In our experiment, directly optimizing the global fully connected layer together with the other two branches can't get satisfied performance. We guess this is because it is too hard to optimize three network branches simultaneously by only one type of loss. Therefore, we first pre-train an auto-encoder network for global feature with structure ($2048 \Rightarrow 1024 \Rightarrow 1024 \Rightarrow 2048$) by the respective training data, and utilize the fixed encoder to project the global image feature.

On the subject of the word, the one-hot vector is first embedded into 300-dimensional dense representation, then the dense representation is fed into the bi-GRU whose hidden dimension is set as 1024 as well. For Flickr30K dataset, the training procedure begins with a learning rate of 0.0002, which is discounted by 10 for every 15 epochs. For MS-COCO dataset, we begin with a learning rate 0.0005, which is discounted by 10 for every 15 epochs. On the Tencent-News dataset, the parameter settings except the embedding size are the same as the Flickr30K, we set the embedding size as 512 to get better performance.

B. Performance Evaluation

1) Comparison with the Competing Method

In this subsection, we make comparisons with several state-

[1]The Tencent News data download link and our code can be found at: https://github.com/HaoYang0123/Position-Focused-Attention-Network/

Table 1: Comparisons of cross-modal retrieval on Flickr30K dataset with the competing methods

methods	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN [2]	67.4	90.3	95.8	48.6	77.7	85.2	77.5
PFAN [48]	70.0	91.8	95.0	50.4	78.7	86.1	78.7
GVSE [57]	68.5	90.9	95.5	50.6	79.8	87.6	78.8
VSRN [56]	71.3	90.6	96.0	54.7	81.8	88.2	80.4
ACMM [58]	80.0	95.5	98.2	50.2	76.8	84.7	80.9
UNITER [59]	-	-	-	-	-	-	88.5
MMCA [61]	74.2	92.8	96.4	54.8	81.4	87.8	81.2
Unicoder-VL [60]	**86.2**	**96.3**	**99.0**	**71.5**	**90.9**	**94.9**	**89.8**
PFAN++-P	66.1	89.4	95.2	50.9	78.1	86.2	77.6
PFAN++ t-i	67.2	91.2	96.1	50.8	77.8	85.3	78.1
PFAN++ i-t	67.3	88.6	93.7	45.7	75.4	83.8	75.7
PFAN++ t-i+i-t	70.1	91.8	96.1	52.7	79.9	87.0	79.6

Table 2: Comparisons of cross-modal retrieval on MS-COCO dataset with the competing methods

methods	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
1K Test Images							
SCAN [2]	72.7	94.8	98.4	58.8	88.4	94.8	84.7
PFAN [48]	76.5	96.3	99.0	61.6	89.6	95.2	86.4
GVSE [57]	72.2	94.1	98.1	60.5	89.7	95.8	85.0
VSRN [56]	76.2	94.8	98.2	62.8	89.7	95.1	86.1
ACMM [58]	81.9	98.0	99.3	58.2	87.3	93.9	86.4
MMCA [61]	74.8	95.6	97.7	61.6	89.8	95.2	85.8
Unicoder-VL [60]	**84.3**	**97.3**	**99.3**	**69.7**	**93.5**	**97.2**	**90.2**
PFAN++-P	75.3	95.1	97.8	60.9	88.6	94.8	85.4
PFAN++ t-i	75.4	95.5	98.2	60.9	88.9	94.7	85.6
PFAN++ i-t	72.0	94.6	98.5	56.4	86.1	92.6	83.4
PFAN++ t-i+i-t	77.1	96.5	98.3	62.5	89.9	95.4	86.7
5K Test Images							
SCAN [2]	50.4	82.2	90.0	38.6	69.3	80.4	68.5
PFAN [48]	50.8	83.9	89.1	39.5	69.5	80.8	68.9
GVSE [57]	47.2	76.6	88.4	31.2	61.2	70.5	62.5
VSRN [56]	53.0	81.1	89.4	40.5	70.6	81.1	69.3
ACMM [58]	**63.5**	**88.0**	**93.6**	36.7	65.1	76.7	70.6
MMCA [61]	54.0	82.5	90.7	38.7	69.7	80.8	69.4
Unicoder-VL [60]	62.3	87.1	92.8	**46.7**	**76.0**	**85.3**	**75.0**
PFAN++-P	50.9	83.2	88.6	40.3	70.1	78.2	68.6
PFAN++ t-i	49.7	83.1	89.8	39.4	70.0	78.7	68.5
PFAN++ i-t	48.3	81.6	87.3	37.6	66.7	77.2	66.5
PFAN++ t-i+i-t	51.2	84.3	89.2	41.4	70.9	79.0	69.3

Fig. 4: The visualization figures of attending image region to each word

of-the-art methods and verify the performance of our proposed model. Table 1 and 2 show the performances of all methods on Flickr30K and MS-COCO dataset, where the PFAN++ t-i means only employing the loss of attending text to image to train the network and PFAN++ t-i + i-t fuses the models from the PFAN++ t-i and the PFAN++ i-t, the PFAN++-P indicates the PFAN++ w/o position attention (i-t + t-i fused). From Table 1, the method Unicoder-VL [60] and UNITER [59] surpass the other methods by a large margin, this is because these two approaches use extra millions of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3024822, IEEE Transactions on Multimedia

PFAN++:
1. Six people ride mountain bikes through a jungle environment .
2. Men , surrounded by nature , are riding mountain bikes .
3. There are six men mountain biking in a forest terrain .

PFAN:
1. Six people ride mountain bikes through a jungle environment .
2. Six People riding bikes on a trail in the forest .
3. A group of people is bike riding in the woods .

SCAN:
1. Six People riding bikes on a trail in the forest .
2. Six people ride mountain bikes through a jungle environment .
3. Five cyclists , all wearing the same uniforms , are riding one behind the other in a bicycle race .

(a)

PFAN++:
1. Two young boys sitting on a sunlit floor smiling and holding a black lab puppy .
2. Two little boys smiling and holding a tiny , black puppy .
3. Two young boys pose with a puppy for a family picture .

PFAN:
1. Two young boys sitting on a sunlit floor smiling and holding a black lab puppy .
2. Two young boys pose with a puppy for a family picture.
3. Two children sitting with a black puppy .

SCAN:
1. Three girls are smiling for a picture .
2. Three girls smiling for the camera .
3. Two little boys smiling and holding a tiny , black puppy .

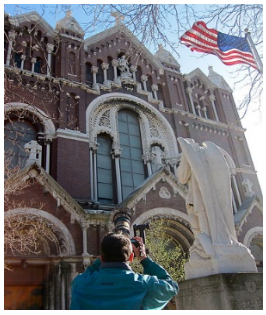
(b)

PFAN++:
1. A blond-haired baby is sitting on the floor playing with toys while looking at a black and white cat .
2. The infant has plenty of toys , but attention is drawn to the nearby cat .
3. A baby girl looking at a black and white cat while holding a toy .

PFAN:
1. A baby playing with her toys looking at a black and white cat .
2. a black and white cat looking at a baby .
3. A blond-haired baby is sitting on the floor playing with toys while looking at a black and white cat .

SCAN:
1. A blond-haired baby is sitting on the floor playing with toys while looking at a black and white cat .
2. A baby laughing on the floor .
3. a black and white cat looking at a baby .

(c)



PFAN++:
1. Man taking picture of church while the american flag blows in the wind.
2. A man wearing a turquoise jacket is taking a picture of a church .
3. A man in a jacket is taking a photograph of a large building .

PFAN:
1. A man in a blue jacket is taking a picture of a church .
2. A man with a blue jacket photographing a large building .
3. Man taking picture of church while the american flag blows in the wind

SCAN:
1. A man in a blue jacket is taking a picture of a church .
2. A man wearing a turquoise jacket is taking a picture of a church .
3. Two men are having a conversation in a cathedral or mosque .

(d)

Fig. 6: Four sentence retrieval results of PFAN++, PFAN and SCAN

Table 3: Performances on Tencent-News dataset

	MAP@1	MAP@2	MAP@3	A@1	A@2	A@3
SCAN	67.2	70.6	75.7	67.2	69.1	73.6
PFAN	76.0	79.0	**82.0**	76.0	76.3	79.7
PFAN++	**77.2**	**80.3**	81.9	**77.2**	**77.1**	**80.9**

image-text pairs to pre-train the model. Except for these two pre-training models, we find that our PFAN++ is competitive with the other methods. For example, the MMCA [61] outperforms our method on Flickr30K, while the PFAN++ is more outstanding on MS-COCO 1K and comparable with the MMCA [61] on MS-COCO 5K test. Comparing with our PFAN++, the ACMM [58] is powerful on text retrieval but relatively weak on image retrieval. Besides, from Table 1 and 2, we can observe that our PFAN++ could win once at least when competing with the latest mothods except for the pre-training approaches. Furthermore, if we only consider the methods using the standard training data, the PFAN++ achieves the best average performance on MS-COCO 1K, which reveals our model is still an effective methodology.

Comparing with PFAN, the sentence retrieval of PFAN++ outperforms a few, but the image retrieval performs much better than PFAN. For example, the R@1 of image retrieval task on Flickr 30K can be improved from 50.4 to 52.7, which validates the global feature is helpful. To validate the effectiveness of the designed position attention, we also conduct experiments for PFAN++ without position attention (PFAN++-P). The results are also reported in Table 1 and 2. From these two tables, we can clearly see that the position attention is important for our system. The performances of

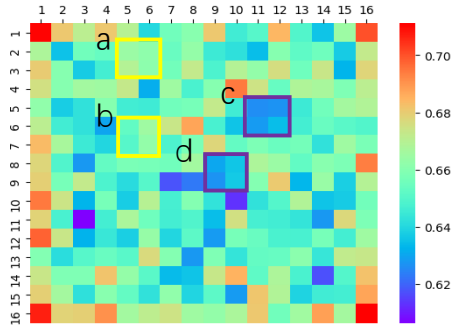


Fig. 5: The visualization of position embedding similarity

PFAN++ on both text retrieval and image retrieval all outperform the PFAN++-P when the depth varies from 1 to 10. This reveals the proposed position attention can help capture a more reliable and credible relationship between the image and the sentence, which validate the contribution of the proposed position attention.

Table 3 shows the performances on Tencent-News dataset. For a returned image list of a news title query, we not only care about the number of relevant images but the ranking order of the relevant items, therefore, we use the Mean Average Precision (MAP) [41, 47] and the Accuracy (A) to evaluate the performance. Accuracy with depth M (A@M) is defined as the number of the correct items divided by M, while the AP with depth M is defined as follows:

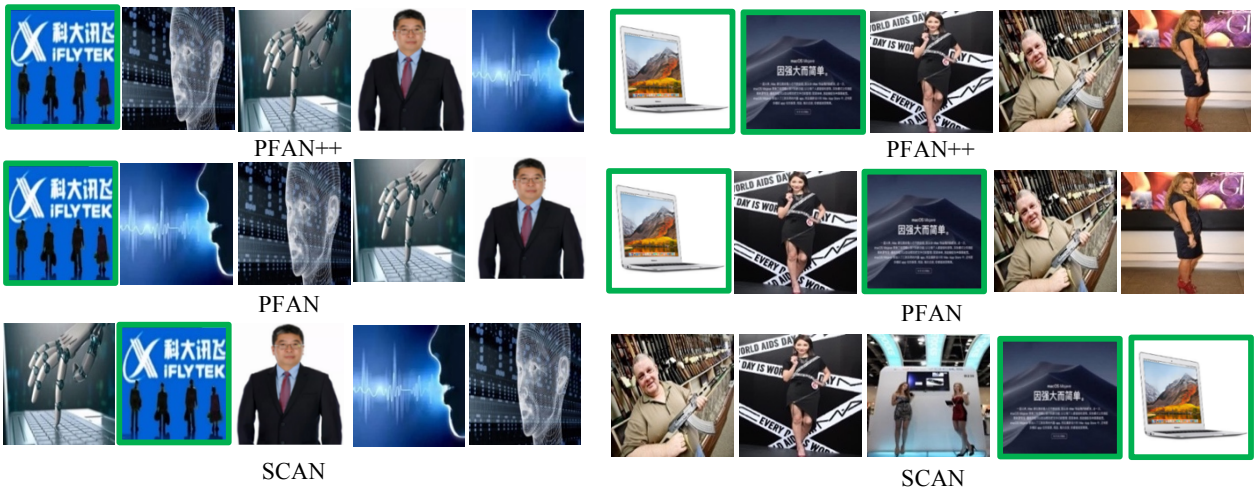
$$AP@M = \frac{1}{M}\sum_{i=1}^{M}\left(\sum_{j=1}^{i}\frac{r_j}{i}\right) \quad (16)$$

where r_j indicates the j-th candidate image is relevant

(a) Results of query: "A girl is in a field surrounded by trees and pushing a pink scooter on the grass."

(b) Results of query: "Five people standing in front of a body of water."

Fig. 7: Two image retrieval results of PFAN++, PFAN and the SCAN on Flickr30K



(a) Results of query: "第三季度扣非净利下滑九成 科大讯飞虚胖症缘何难解"(non-net profit declined by 90% in the third quarter, why is it difficult to solve the puffiness of IFLYTEK)

(b) Results of query: "一文打尽所有爆料，关于最新款 Mac、iPad 的信息都在这里了" (catch all stuff in one article, all the information about the latest Mac、iPad is here)

Fig. 8: Two exemplary queries on Tencent-News dataset

(labeled as 1) or irrelevant (labeled as 0) to the query title. The mean AP@M on all the test data is the MAP@M.

It is clear from Table 3 that our PFAN++ is still more outstanding than method SCAN and PFAN, the PFAN++ can outperform the SCAN by seven points on average under both MAP and Accuracy. Tables 1-3 show the effectiveness and the practical application value of the proposed method.

C. Result Visualization

In this subsection, we would visualize some results to intuitively see the ability of the proposed network on relationship capture aspect. The parameters are the same as the settings in subsection IV A) 2).

1) Position Embedding Visualization

Given an image, we split the image into blocks to infer the relative position of the image region. Instead of applying the simple one-hot representation, we utilize the embedding layer to adaptively learn the position embedding for better position representation. More details can be found in Section III.

The learned block index embedding should preserve the locality, i.e. the neighbor position embeddings should be close to each other. In order to see if the learned position embedding preserve the locality, we first compute the similarity matrix $SM = R^{16\times16}$ of the block position embeddings, and the component $SM(i,j)$ is defined as the average similarity between the block embedding in i- th row and j-th column and its adjacent embeddings:

$$SM(i,j) = \frac{1}{|\mathcal{L}(b_m)|}\sum_{d\in\mathcal{L}(b_m)} \exp\left(-\frac{||b_m-d||^2}{2\sigma^2}\right) \quad (17)$$

the SM is calculated based on Gaussian kernel function, where the $m = i\times16 + j$, $\mathcal{L}(b_m)$ is the collection of position embedding that is directly adjacent to the b_m, $\sigma \in R$ is a scalar, we simply set it as the average distance between all the position embeddings.

Fig. 5 shows the visualization result of matrix SM, we can find that the closer block position embeddings share the higher similarity in most cases. For example, the four blocks in box (a)-(d) are very similar to each other, and many analogy situations can be found in Fig. 5, which means the learned position embeddings preserve the locality.

We can also observe that the similarities of some position embeddings close to the center are with lower value ((c), (d)), while many marginal blocks are highly similar to each other. We guess that this is because the regions in the corners are relatively similar, just as shown in the lower right region in Fig. 1(a) and (b). Therefore, the position features in corner

a）图片中的孩子是一名唐氏综合症患儿，|孩子自己一个人靠墙坐着，|而拉布拉多注意到孩子低落的情绪，|想要逗孩子开心起来，|但是显然，孩子并不想理拉布拉多(The child in the picture is with Down syndrome, |he sits alone against the wall, |and Labrador notices the child's low mood, |it wants to make the child happy, |but obviously, the child does not want to play with Labrador)

b）倪妮，大家都很熟悉了，|尤其是给人印象最深刻的是她极高的时尚品味，|不管是私服穿搭还是出席活动，|穿搭都十分在线，这当然离不开她的高颜值，|姣好的身材，以及对时尚的独特见解 (Ni Ni, everyone is very familiar with her, |especially the one that is most impressive is her high fashion taste, |whether it is wearing a private service or attending an event, |she is very radiant, this is of course inseparable from her beautiful face, |great figure, and special insights into fashion)

Fig. 9: Two exemplar image-sentence fragments matching instances, the blue vertical line indicates the sentence fragments.

frequently meet similar visual content and fit the similar visual feature in each iteration. The final learned position embedding should be similar. However, the content close to the center in different images are violently changed, the corresponding block position needs to fit various visual contents, therefore, there will be obvious differences accordingly.

2) Attention Visualization

We design a position attention mechanism to adaptively determine the importance of the block position to the region, the region positional and the visual feature are then concatenated and fed into the image-text attention mechanism to investigate the interplay between regions and words. More details can be found in Subsection III-B. In this subsection, we visualize the attention results in this paper.

An exemplary visualization result is shown in Fig. 4, where the green box indicates the image region, the word with the maximum attention weight to the region is exhibited in each figure. The red frames indicate the blocks of the region attending, we exhibit the blocks of the first 6 maximum weights for each region and the brighter ones are with higher weights. There are two observations from Fig. 4:

a. The correspondings between the image regions and the sentence words is satisfied, most of the words can attend to their related semantic regions, like the words “wearing”, “shirt”, “shorts”, and so on.

b. The brighter blocks indeed reveal the more important part of the regions. For example, the third image in the first row, the brightest block locates in the center of the region, which is one of the most semantic related parts. From the sixth image, we can get the similar observation.

D. Retrieval Experiments

In this section, we exhibit some sentence and image retrieval results to intuitively display the performances. We only make comparisons with the current best method SCAN [2] and PFAN to show the superior of our method PFAN++.

1) Sentence retrieval

This subsection presents some sentence retrieval results of proposed PFAN++, PFAN and SCAN. Fig. 6 exhibits the top 3 retrieved results of four image queries for three methods respectively, where the red color indicates the irrelevant

results. From Fig. 6, we can see that the SCAN suffers from the lack of relevance. For example, in Fig. 6. (b), the results of SCAN introduce two irrelevant results in the top 3 retrieved results, and the results in Fig. 6 (a), (c) and (d) all introduce one irrelevant results. While our method can recall the relevant sentences for these four image queries, which reveals the superior of our method.

2) Image Retrieval

Fig.7 shows two exemplar sentence queries and the corresponding top-5 retrieval images. The green frame indicates the ground truth image of the query in the test dataset. From Fig. 7, we can find that the top-5 retrieved results of the PFAN++, PFAN, and SCAN are similar. From Fig. 7 (a)-(b), we can obtain two important observations:

a. The proposed PFAN++, PFAN and the SCAN can both recall the ground truth image in top-5 retrieval results for these two queries.

b. Our PFAN++ and the PFAN can pick up the truly relevant image for the query sentence, although there are many images that have similar semantics with the ground truth, while the SCAN is not that good.

For example, there are four common images of our PFAN++, PFAN and the SCAN in Fig. 7 (a). Fig. 7 (a) in fact

shows a hard query, the image (1) and the image (2) are very similar to each other, and we can’t discriminate them only according to the first half of the query sentence: “a girl is in a field surrounded by trees”. To pick up the truly relevant image, the model must accurately capture the relation between the image content and the second half of the sentence: “pushing a pink scooter on the grass.” From Fig. 7(a), our PFAN++ correctly models the relation and pick up the truly relevant image, while the SCAN doesn’t properly capture the relation and ranks the correct image in the second place. The

Table 4: The performances of PFAN++ with different split size on Flickr30K

methods	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
PFAN++-2×2 t-i	64.9	86.1	90.0	47.3	75.9	83.9
PFAN++-4×4 t-i	64.2	87.1	91.2	48.2	76.7	84.1
PFAN++-8×8 t-i	**67.4**	88.8	93.7	49.3	**78.2**	**85.8**
PFAN++-16×16 t-i	67.2	**91.2**	**96.1**	**50.8**	77.8	85.3
PFAN++-32×32 t-i	65.3	88.3	93.8	49.7	77.5	84.3
PFAN++-2×2 i-t	65.1	88.3	91.4	43.1	72.6	82.3
PFAN++-4×4 i-t	66.7	88.6	91.9	43.9	73.6	83.3
PFAN++-8×8 i-t	**67.3**	89.7	93.8	44.1	74.0	82.8
PFAN++-16×16 i-t	**67.3**	88.6	93.7	**45.7**	75.4	83.8
PFAN++-32×32 i-t	**67.3**	**90.3**	**94.5**	45.4	**76.0**	**84.0**
PFAN++-2×2 t-i+i-t	66.8	88.7	94.1	48.3	75.4	84.9
PFAN++-4×4 t-i+i-t	67.5	90.2	94.9	48.7	76.8	85.3
PFAN++-8×8 t-i+i-t	68.9	89.9	94.7	49.6	77.4	85.6
PFAN++-16×16 t-i+i-t	**70.1**	**91.8**	**96.1**	**52.7**	**79.9**	**87.0**
PFAN++-32×32 t-i+i-t	68.9	90.4	94.7	49.8	78.4	85.9

Table 5: The performances of PFAN++ with different numbers of blocks for position inferring on Flickr30K

methods	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
PFAN++-1 t-i	55.2	86.1	91.2	43.4	72.1	80.0
PFAN++ 5 t-i	57.1	87.1	92.1	44.9	73.2	82.4
PFAN++-10 t-i	56.4	87.0	93.2	45.4	72.9	81.3
PFAN++-15 t-i	**67.2**	**91.2**	96.1	**50.8**	**77.8**	85.3
PFAN++-25 t-i	65.6	88.7	**96.6**	49.4	76.9	**85.8**
PFAN++-1 i-t	66.2	88.0	92.7	44.1	73.6	82.1
PFAN++-5 i-t	67.2	89.2	93.3	45.9	74.9	83.8
PFAN++-10 i-t	**68.0**	**90.2**	93.9	46.2	**75.6**	**84.1**
PFAN++-15 i-t	67.3	88.6	93.7	45.7	75.4	83.8
PFAN++-25 i-t	67.8	89.8	**94.2**	**46.3**	75.0	83.9
PFAN++-1 t-i+i-t	64.9	88.2	93.3	47.3	76.7	85.5
PFAN++-5 t-i+i-t	65.3	88.9	95.7	51.2	77.0	85.9
PFAN++-10 t-i+i-t	66.1	90.7	**96.2**	**53.1**	78.4	86.0
PFAN++-15 t-i+i-t	**70.1**	**91.8**	96.1	52.7	**79.9**	**87.0**
PFAN++-25 t-i+i-t	69.4	90.9	94.8	50.5	78.0	85.7

PFAN++ is also more outstanding on the query shown in Fig. 7(b).

Fig. 8 shows two exemplary queries on Tencent-News dataset, from which we can find that the PFAN++ and PFAN achieve more satisfactory results than SCAN. For example, there is only one appropriate image for the query title in Fig. 8 (a), the PFAN++ picks up the most relevant images, while the SCAN puts the correct image in the second place. From Fig. 8 (b), the superior of our PFAN++ is more obvious.

From the retrieval results shown in Figs. 6-8, it is clear that our proposed model PFAN++ can capture the more accurate and reliable relation between the sentence and the image, which validates the effectiveness of proposed method.

E. An Interesting Exploration Experiment: Short Dynamic Video Generation for Tencent News

When users browse a piece of news, they usually want to get more information from the news with a lower time cost. This can be achieved by providing a very short video about the news. In order to generate a satisfied short video, the news main description need to be extracted and need to be actually matched with the images of the news, which is a key step in overall procedure. The steps of our short video generation are summarized in Appendix A.

Fig.9 shows the image-sentence fragments matching results of two news. The blue vertical line indicates the sentence fragments and the corresponding matched image is placed by the same order. From Fig. 9, we find that the most of the sentence fragments can be assigned to an appropriate image. For example, the example in Fig.9 a) shows a good match, the images and the sentence fragments are consistent and the final generated video will form a coherent semantics. We supply 466 dynamic examples, which can be found at https://drive.google.com/file/d/1XfVGJXzaBca67y1V_6c3NVoQYCylTFrs/view?usp=sharing.

V. DISCUSSION

In this subsection, we conduct experiments on Flickr30K to explore the impact on the final performance of some parameters in our proposed methods, including the split size K and the number of block L employed to infer the region position. Besides, the time cost is also analyzed in this section.

1) Discussion About the Split Size

To infer the position of the region, we equally split the image into $K \times K$ blocks, and the parameter K is set as 16 in our baseline method. In this subsection, we investigate the impacts of different K. Table. 4 shows the performances of PFAN++ with different split sizes on the Flickr30K dataset, where the PFAN++-#×# means PFAN++ with split size #. From Table. 4, we can see that the performance of PFAN++ with split 2 and 4 is not satisfactory, especially the PFAN++ with 2×2 split is much worse, whose R@1 of image retrieval is only 47.3. We guess the reason for the poor performance may stem from two aspects: first, limited position cells are too sparse to support the accuracy position inferring. Second,

Table 6: Time cost comparison on Flickr30K for four methods ("s" and "m" indicates the second and minute, respectively).

		VSE++[21]	SCAN[2]	PFAN[48]	PFAN++
Feature Extraction	Resnet152[36]	0.0330s	-	-	0.0330s
	Faster RCNN[37]	-	0.3084s	0.3084s	0.3084s
Inference/image		**0.0266s**	0.2916s	0.3571s	0.3790s
Total test cost/image		**0.0594s**	0.6000s	0.6655s	0.6874s
Training cost/epoch		**2.184m**	14.22m	14.99m	15.33m

the fewer split size means each position cell would cover much large region, consequently, the semantics of each position embedding is ambiguous, which also would cause a bad position inferring. With the split size increasing, we can get satisfactory results. For example, the R@1 of PFAN++-8×8 t-i can reach 67.4 for text retrieval, and the fused result of PFAN-16×16 achieves the best performance. Therefore, we set the split size as 16. Although the split size is set as 16 in our baseline method, the PFAN with other split sizes like 8 and 32 can also achieve satisfactory performance.

2) Discussion About the Parameter L

The first 15 block positions are applied to generate the position feature in our baseline method. In this subsection, the split size is fixed at 16, we vary the number of blocks employed for position inferring to investigate the performance effect of different L.

Table. 5 shows the performance of PFAN++ employing different numbers of blocks to generate the position feature, where the PFAN++-# means the PFAN++ with $L = \#$., i.e. $L = 1, 5, 10, 15, 25$. From Table 5, the case of L=1 performs worst, because the position attention could not be equipped with only one position candidates, in addition, using only one position blocks for position inferring is also not an excellent strategy. The R@1 of the PFAN++-5 and 10 t-i also perform much worse than other situations, this may be because the too few block positions are insufficient for the region's position inferring. The performances of the PFAN++ with $L = 15$ and 25 are very competitive, the PFAN++-15 shows the best performance on average. The final fused performance of PFAN++-25 performs a little worse than the PFAN++-15, we guess this because too many block positions introduce some redundant information and confuse the joint learning procedure, which causes the performance dropping.

3) Efficiency Analysis

In this subsection, we simply analyze the time efficiency of the proposed PFAN++. To get the ranking texts for an image query, our model need first extract the image features and feed the visual features and the text forward the model to obtain the similarity score, by which we could pick up the relevance text for the image (if text serves as query, the procedure is analogous). We conduct the experiments on Flickr30K and count the time of the similarity vector calculation for each test image (1000 in total) VS 5000 candidate sentences, the average time costs are reported in Table 6. As shown in Table 6, the VSE++ [21] is the most time-saving method due to its simple input, architecture, and loss calculation. Comparing to VSE++ [21], the other three methods all use the fine-grained object features, and the similarity/loss computation is more complex. Consequently, their time costs are much higher. PFAN [48] introduces the position embedding layer and the position attention mechanism, therefore, its inference speed is lower than SCAN [2]. Our PFAN++ further introduces an additional layer for the global feature based on PFAN, which also slows down the inference speed. Although our PFAN++ is the most time-consuming, it could boost the performance on all three datasets while only sacrifice 0.0219(s) in inference stage comparing to PFAN. When the text serves as query, most of the time would be spent on image feature extraction. For PFAN++, since there are 1000 images, and the number of candidates is 5 times fewer than the text retrieval, the total time cost for a text query would be around: (0.033+0.308) ×1000+0.379/5(s) (feature extraction + model inference). In practice, we could extract and save the regional and global features of images beforehand, which could save around 44.86% of the text-retrieval time, and 99.9% of the image-retrieval time.

As for the training time, because of the complexity of model architecture and similarity/loss computation, the conclusion is the same as in the test stage. The VSE++ [21] is still the fastest. Comparing to PFAN [48], our time cost of PFAN++ is 0.333 mins slower, which is acceptable considering the performance improvement.

VI. Conclusion and Future Work

In this paper, we develop a position focused attention network for the image-text bi-directional retrieval task. Instead of only paying attention to the regions themselves, the clue of the region position is taken into consideration. We first split the image and utilize the split blocks to infer the relative position of the region, an attention weight is then assigned to the block with respect to each region and position feature is then adaptively generated by the designed position attention. Positional feature and visual feature are concatenated to form the final representation of the region. Besides the local feature, the global information is also introduced to enhance the embedding learning, which makes the performance step further. The experiments on the popular Flickr30k and MS-COCO datasets reveal that integrating the position information can help model a more reliable relation between the image and the text. We further collect a practical dataset (Tencent-News) and make the first attempt to evaluate the application value of our image-text model. The results on these three datasets are all much better than the competing methods and achieve the competitive performance. In the future, we will fuse more semantic information to learn the cross-modality relations.

Appendix A: Short Dynamic Video Generation

Our short video generation consist of four steps:

Step 1: Key Sentence Extraction. As for the expression of the main news content, we simply extract one sentence (called key sentence) to represent the key content of the news. Since the news title has been exhibited for the user when user browses, we don't repeatedly show it and choose another sentence to allow the user to get more information by a simple glance. We first extract the news summary by method TextRank [49] and the key sentence is selected according to the similarity with the news title under the BERT [50] feature representations.

Step 2: Text Detection. Since the sentence fragment assigned to an image will serve as a caption in the bottom. The images with texts in the bottom need to be removed in advance. Therefore, we employ the text detection framework [51] and remove the images with text in their bottom.

Step 3: Image-Sentence Fragments Matching. Our extracted key sentence is long enough, therefore, we need first cut the key sentence into fragments with appropriate

length. Then the fragments and the images in this news are matched by our image-text matching model PFAN++.

Step 4: Short Video Generation. Since the occurrences of the images in news obey the semantic order of the overall content, the order of frames (images) in the generated video should be consistent with the order in the original news. That is the next frame (image) should be one of the subsequent images of current frame (image). With the image-sentence fragments similarity in step 3, we employ the Breadth-First-Search algorithm to find an optimal image sequence. For each frame, the top-k images with the highest matching scores are chosen as candidates. Assuming that there are n sentence fragments $\{f_i\}_{i=1}^{n}$ and m frames (images, $\{I_j\}_{j=1}^{m}$), for every fragment f_i, the k frames with the highest matching scores are retained and we set k as the total number of candidate images to guarantee finding at least one satisfied sequence. By this way, the semantic order of frames can be preserved.

The matched sentence fragment for each frame is treated as the caption at the bottom of the frame, finally, a short video is generated by splicing all the image-fragment pairs.

REFERENCES

[1] B. Plummer, P. Kordas, M. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik, "Conditional Image-Text Embedding Networks", ECCV, pp. 258-274, 2018.

[2] K. Lee, X. Chen, Hua, H. Hu, and X. He, "Stacked Cross Attention for Image-Text Matching", ECCV, 2018.

[3] Hu, R., Xu, H., Rohrbach, M., Feng, J, Saenko, K., and Darrell, T, Natural language object retrieval. In: CVPR. pp. 4555–4564, 2016.

[4] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models", CVPR, 2018.

[5] Yan, F., and Mikolajczyk, K, "Deep correlation for matching images and text". CVPR, 2015.

[6] A. Eisenschtat, and L. Wolf, "Linking Image and Text with 2-way Nets", CVPR, pp. 1855-1865, 2017.

[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., and Bengio, Y, "Show, attend and tell: Neural image caption generation with visual attention", ICML. pp. 2048 – 2057, 2015.

[8] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", arXiv/1141.2539, 2014.

[9] Vinyals, O., Toshev, A., Bengio, S., and Erhan. D, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge". TPAMI, vol.39, no. 4, pp. 652 – 663, 2017.

[10] Ma, L., Lu, Z., Shang, L., and Li, H. "Multimodal convolutional neural networks for matching image and sentence". ICCV. pp. 2623 – 2631 2015.

[11] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L, and Parikh, D, "VQA: visual question answering". ICCV. pp. 2425 – 2433, 2015.

[12] X. Chang, T. Xiang, T. Hospedals, "Scalable and Effective Deep CCA via Soft Decorrelation", CVPR, pp. 1488-1497, 2018.

[13] L. Wang, Y. Li, and J. Huang, S. Lazebninik, "Learning Two-Branch Neural Networks for Image-Text Matching Tasks". IEEE TPAMI, early access (1-1), 2018.

[14] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation", CVPR, 2015.

[15] Lin, X., and Parikh, D. "Leveraging visual question answering for image-caption rank", ECCV. 2016.

[16] Y. Zhang, and H. Lum "Deep Cross-Modal Projection Learning for Image-Text Matching", ECCV, pp. 707-723, 2018.

[17] H. Nam, J. Ha, and Jeonghee Kim, "Dual Attention Networks for Multimodal Reasoning and Matching", CVPR, pp. 2156-2164, 2017.

[18] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning Semantic Concepts and Order for Image and Sentence Matching", CVPR, 2018.

[19] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y. Shen, "Dual-Path Convolutional Image-Text Embedding with Instance Loss", CVPR, 2018.

[20] Y. Gong. L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections", ECCV, pp.529-545, 2014.

[21] F. Faghri, D. Fleet, J. Kiros, and S. Fidler, "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives", BMVC, 2018:12.

[22] Y. Huang, W. Wang, and L. Wang, "Instance-Aware Image and Sentence Matching with Selection Multimodal LSTM", CVPR, pp. 7254-7262, 2017.

[23] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, "Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding", ICCV, pp. 1899-1907, 2017.

[24] Y. Hu, L. Zheng, Y. Yang and Y. Huang, "Twitter100k: A Real-World Dataset for Weakly Supervised Cross-Media Retrieval", IEEE Trans on Multimedia, 20(9), 2018.

[25] G. Andrew, R. Arora, J. Abilmes, and K. Livescu, "Deep Canonical Correlation Analysis", ICML, 2013.

[26] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On Deep Multi-View Representation Learning", ICML, pp. 1083-1092, 2015.

[27] A. Karpathy, and F. Li, "Deep Visual-Semantic Alignments for Generating image descriptions", CVPR, pp. 3128-3138, 2015.

[28] W. Wang, R. Arora, K. Livescu, and N. Srebro, "Stochastic Optimization for Deep CCA via Nonlinear Orthogonal Iterations", Allerton, pp. 688-695, 2015.

[29] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Caption and VQA", CVPR, 2018.

[30] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning Deep Representations of Fine-Grained Visual Descriptions", CVPR, pp. 49-58, 2016.

[31] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering", arXiv:1601.01705, 2016.

[32] Jba, V. Mnih, and K. Kavukcuoglu, "Multiple Object Recognition with Visual Attention", ICLR, 2015.

[33] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions", TACL, 2: 67-78, 2014.

[34] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering", CVPR, pp. 21-29, 2016

[35] K. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering", CVPR, 2016

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", CVPR, 2016.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE TPAMI, vol. 39, no. 6, pp. 1137-1149, 2017.

[38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma,, M. Bernstein, and F. Li, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image", IJCV, vol. 123, no. 1, pp. 32-73, 2017.

[39] D. Kingma, and J. Ba, "Adam: A Method for Stochastic Optimization", ICLR, 2015.

[40] L. Zhang, B. Ma, G. Li, Q. Huang and Q. Tian, "Cross-Modal Retrieval Using Multiordered Discriminative Structured Subspace Learning", IEEE Trans on Multimedia, 18(2), 2016.

[41] Y. Wang, L. Zhu, X. Qian, and J. Han, "Joint Hypergraph Learning for Tag-Based Image Retrieval", IEEE Trans on Image Processing, 27(9), 4437-4451, 2018.

[42] L. Ma, W. Jiang, Z. Jie, Yugang Jiang, and Wei Liu, "Matching Image and Sentence with Multi-faceted Representation", early access, IEEE TCSVT, 2019.

[43] Lin Ma, Wenhao Jiang, Zequn Jie, and Xu Wang, "Bidirectional image-sentence retrieval by local and global deep matching", Neurocomputing 345:36-44, 2019.

[44] Linchao Zhu, Zhongwen Xu, and Yi Yang, "Bidirectional multirate reconstruction for temporal modeling in videos", CVPR, pp. 1339-1348, 2017.

[45] Y. He, S. Xiang, C. Kang, J. Wang and C. Pan, "Cross-Modal Retrieval via Deep and Bidirectional Representation Learning", IEEE Trans on Multimedia, 18(7), 2016.

[46] C. Kang, S. Xiang, S.Liao, C. Xu and C. Pan, "Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval", IEEE Trans on Multimedia, 17(6), 2015.

[47] Xueming. Qian, Dan. Lu, Yaxiong. Wang, Li. Zhu, Yuanyan Tang, and Meng. Wang, "Image Re-Ranking Based on Topic Diversity", IEEE TIP, vol. 26, no. 8, 2017.

[48] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li and Xin Fan. "Position Focused Attention Network for Image-Text Matching". IJCAI, 2019.

[49] Rada Mihalcea and Paul Tarau, "TextRank: Bringing Order into text". EMNLP, 2004:404-411.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL-HIT, 2019:4171-4186.

[51] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detection Text in Natural Image with Connectionist Text Proposal Network", ECCV, 2016.

[52] S. Karaoglu, R. Tao, T. Gevers, and A. Smeulders, "Words Matter: Scene Text for Image Classification and Retrieval", IEEE Trans on Multimedia, 19(5), 2017.

[53] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang and J. Xu, "COCO-CN for Cross-Lingual Image Tagging, Caption, and Retrieval", IEEE Trans on Multimedia, 21(5), 2019.

[54] E. Yu, J. Sun, J. Li, X. Chang, X. Han, and A. Hauptmann, "Adaptive Semi-Supervised Feature Selection for Cross-Modal Retrieval", IEEE Trans on Multimedia, 21(5), 2019.

[55] L. Zhang, B. Ma, G. Li, Q. Huang and Q. Tian, "Generalized Semi-supervised and Structure Subspace Learning for Cross-Modal Retrieval", IEEE Trans on Multimedia, 20(1), 2018.

[56] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual Semantic Reasoning for Image-Text Matching", ICCV:4653-4661, 2019.

[57] Y. Huang, Y. Long, and L. Wang, "Few-Shot Image and Sentence Matching via Gated Visual-Semantic Embedding", AAAI:8489-8496, 2019.

[58] Y Huang, and L. Wang, "ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching", ICCV: 5773-5782, 2019.

[59] Y. Chen, L. Lin, L. Yu, A. Kholy, F. Ahmed, Z. Gan, Y, Chen, and J. Liu, "UNITER: Universal Image-Text Representation Learning", ECCV 2020.

[60] G. Li, N. Dan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training", AAAI:11336-11344, 2020.

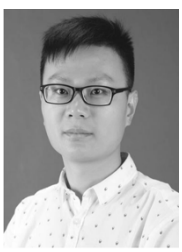
[61] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-Modality Cross Attention Network for Image and Sentence Matching", CVPR:10938-10947, 2020.

Yaxiong Wang received the B.S. degree from Lanzhou University, Lanzhou, China, in 2015. He is currently working towards the Ph.D degree at School of software Engineering, Xi'an Jiaotong University, Xi'an, China. He is now a Post-Graduate at SMILES Laboratory, Xi'an Jiaotong University. His current research interests include tag-based image retrieval and imge-text matching.

Hao Yang received B.E. degree from WuHan University in 2012 and Ph.D degree from Institute of Computing Technology, Chinese Academy of Sciences in 2018. He is now working in Department of PCG, Tencent, Beijing, China. His current research interests include cross media retrieval and video recommendation.

Xiuxiu Bai received the B.S. degree from Xi'an Jiaotong University in 2009, and the Ph.D. degree from Xi'an Jiaotong University in 2016. She visited Edinburgh University, from 2017 to 2018. She is currently an Assistant Professor with the School of Software Engineering, Xi'an Jiaotong University. Her research interests include computer vision and visual neuroscience.

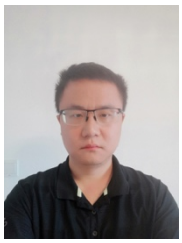
Xueming Qian (M'10) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, in 2008. He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Assistant Professor with Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory at Xi'an Jiaotong University. He received the Microsoft Fellowship in 2006. He received outstanding doctoral dissertations of Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively. His research interests include social media big data mining and search. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and Ministry of Science and Technology.



Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noah's Ark Laboratory, Hong Kong, from 2013 to 2016. He is currently a Principal Researcher with the Tencent AI Lab, Shenzhen, China. His current research interests lie in the ares of computer vision and multimodel deep learning. He was a recipient of the Microsoft Research Asia Fellowship in 2011 and the Best Paper Award from the Pacific-Rim Conference on Multimedia in

2008. He was the Finalist of the HKIS Young Scientist Award in engineering Science in 2012.

Biao Li received the B.S and M.S degrees from Beijing university of Posts and Telecomunications in 2008 and 2011, respectively. He is a senior researcher in News Algorithm Center, Department of PCG, Tencent, Beijing, China. His main research interests include cross-media retrieval and news recommendation & retrieval.



Jing Lu received the Ph.D degree from University of Science and Technology Beijing in 2012, and served as a postdoc researcher in Peking University in 2016. He is a senior researcher in News Algorithm Center, Department of PCG, Tencent, Beijing, China. He currently focuses on the research of image-text matching, nature language process and recommendation system.

Xin Fan received the B.S., M.S. and Ph.D degrees from the University of Science and Technology of China, in 2001, 2004 and 2007 respectively. He was a Scientist with Yahoo Labs Beijing and a senior architect with the Core Search Department, Baidu. He is currently a Director with the Algorithm Center of News Product and Technology Department, Tencent. His current research interests include machine learning, NLP and data mining in search and recommendation.