Context-Aware Multi-View Summarization Network for Image-Text Matching

Conference Paper · October 2020

DOI: 10.1145/3394171.3413961

5 authors, including:

Leigang Qu
Shandong University
4 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Meng Liu
Shandong Jianzhu University
26 PUBLICATIONS **888** CITATIONS
SEE PROFILE

Context-Aware Multi-View Summarization Network for Image-Text Matching

Leigang Qu
Shandong University
leigangqu@gmail.com

Meng Liu*
Shandong Jianzhu University
mengliu.sdu@gmail.com

Da Cao
Hunan University
caoda0721@gmail.com

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

Qi Tian
Huawei Cloud & AI
tian.qi1@huawei.com

ABSTRACT

Image-text matching is a vital yet challenging task in the field of multimedia analysis. Over the past decades, great efforts have been made to bridge the semantic gap between the visual and textual modalities. Despite the significance and value, most prior work is still confronted with a multi-view description challenge, *i.e.*, how to align an image to multiple textual descriptions with semantic diversity. Toward this end, we present a novel context-aware multi-view summarization network to summarize context-enhanced visual region information from multiple views. To be more specific, we design an adaptive gating self-attention module to extract representations of visual regions and words. By controlling the internal information flow, we are able to adaptively capture context information. Afterwards, we introduce a summarization module with a diversity regularization to aggregate region-level features into image-level ones from different perspectives. Ultimately, we devise a multi-view matching scheme to match multi-view image features with corresponding text ones. To justify our work, we have conducted extensive experiments on two benchmark datasets, *i.e.*, Flickr30K and MS-COCO, which demonstrates the superiority of our model as compared to several state-of-the-art baselines.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Novelty in information retrieval**.

KEYWORDS

Image-Text Matching; Cross-Modal Retrieval; Multi-View Summarization; Context Modeling

ACM Reference Format:
Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-Aware Multi-View Summarization Network for Image-Text Matching. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413961

* Meng Liu (mengliu.sdu@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are the corresponding authors.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413961

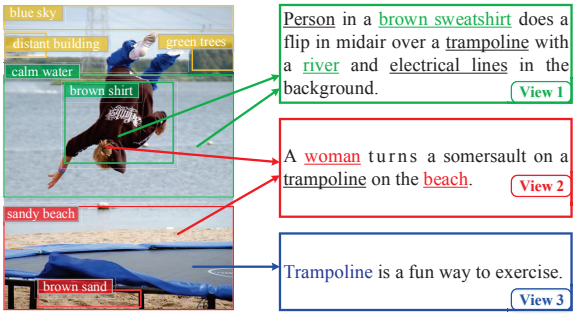


Figure 1: Multi-view descriptions for a given image. The underlined words indicate the objects, and the boxes in the image refer to the detected salient regions. Different colors correspond to different views.

1 INTRODUCTION

Recent years have witnessed the rapid growth of multimedia data, such as texts and images, inducing many researchers to work on the multimodal representation, understanding, and reasoning. As a fundamental task of multimodal interaction, image-text matching, focusing on measuring the semantic similarity between an image and a text, has attracted extensive research attention. It indeed facilitates various applications, such as cross-modal retrieval [3, 37], visual question answering (VQA) [2, 24] and multimedia understanding [10, 38, 39].

However, it is non-trivial to build an effective image-text matching model, due to the following reasons: 1) **Multi-view description.** As the old saying goes, "there are a thousand Hamlets in a thousand people's eyes". Thereby, the same image may have multi-view descriptions. For instance, the image in Figure 1 is depicted by three descriptions from different angles. To be more specific, the first concerns more details (*e.g.*, "brown sweatshirt", "river" and "electrical lines"). The second focuses on the gender and beach. Nevertheless, the third summarizes the image from a more abstract point, ignoring all entities. From this knowable, a single and general image representation is difficult to contain the complete information of all views, leading to situations where the textual descriptions from certain views fail to match the

corresponding image. In light of this, it is essential to explore multi-view understanding and representation of the image. And 2) **The heterogeneous gap**. Image-text matching requires comprehensive understanding of complex visual semantics and various textual information simultaneously. However, the heterogeneous gap induced by the inconsistent distributions of visual and textual modalities would greatly impede its implementation. Therefore, it is vital to bridge this gap for more robust image-text matching.

It is worth mentioning that some pioneer efforts [5, 33, 34, 44] have been devoted to tackling the issue of heterogeneous gap and achieved promising performance. In particular, they commonly embed images and texts as global vectors into a unified semantic space, whereby cross-modal similarities can be calculated for matching. However, it would be difficult for such a compact representation to capture fine-grained semantic details in texts and images. Afterwards, to further exploit the fine-grained relationships between images and texts, several studies [16, 18, 37] have adopted the attention mechanism to perform semantic alignments between visual regions and words. Particularly, they aggregate fragment features guided by local cross-modal affinities, *i.e.*, region-word similarities, to obtain global representations. Nevertheless, these methods ignore the effect of intra-modal context information and takes unavoidable computational cost for the pair-wise similarity estimation. They hence are unsuitable for large-scale applications. To alleviate such problems, some schemes [17, 40] leverage intra-modal interaction to perform semantic reasoning from local fragments (*e.g.*, visual regions and words) to global holistic representations (*e.g.*, the image-level feature and text embedding). Though some semantic relations and indwelling patterns have been captured, they cannot adaptively exploit the context-aware representation from the internal information flow. Besides, they still suffer from the multi-view description problem.

To address the above issues, we propose a novel **C**ontext-**A**ware **M**ulti-vi**E**w summa**R**iz**A**tion network, CAMERA for short, which exploits the context-aware representations and aggregates visual regions based on a multi-view summarization module. Specifically, as illustrated in Figure 2, we first apply the bottom-up attention to identify the salient regions at the object/stuff level. Thereafter, we introduce an adaptive gating self-attention (AGSA) module to exploit the intra-modal context for images and texts, which is capable of controlling the information flow more flexibly. Following that, we devise a multi-view summarization module to aggregate region-level features into multiple image-level embeddings from different views. We next present a diversity regularization to circumvent the information overlap among different views. Meanwhile, to learn the semantic alignment in the joint space, we adopt a bidirectional triplet loss with hard negative sample mining.

The contributions of this work are three-fold:

- We propose a novel image-text matching model CAMERA, which is able to summarize region-level representations from multiple views, and hence capture cross-modal semantic alignments more accurately.
- We present an AGSA module to acquire the intra-modal context-enhanced representations for images and texts concurrently by integrating the gate mechanism and the multi-head self-attention.
- We have conducted extensive experiments on two benchmark datasets: Flickr30K [42] and MS-COCO [19], to validate the effectiveness of our approach. As a side contribution, we have released our codes to benefit other researchers[1].

2 RELATED WORK

2.1 Image-Text Matching

According to the granularity of representation, studies on image-text matching can be categorized into two groups: 1) global embedding based methods [5, 6, 34, 44], and 2) local inference based methods [3, 16, 18, 21, 26, 37]. The former ones first embed the whole images and sentences into a joint embedding space, and then calculate the visual-semantic similarity. For instance, Frome *et al.* [6] designed the first visual-semantic embedding model. To be specific, it respectively employs the CNN and Skip-Gram [25] to obtain the features of images and texts, and adopts a ranking loss for metric learning. Wang *et al.* [34] proposed a two-branch network optimized via large-margin objectives including the cross-view ranking and within-view neighborhood preservation constraints. To learn cross-modal alignments more efficiently, some researchers have turned to design more novel loss functions. Faghri *et al.* [5] designed a ranking loss based on violations incurred by relatively hard negatives. Zhang *et al.* [44] presented a cross-modal projection matching loss and a classification loss to obtain discriminative features. Though some promising performance has been achieved, these approaches learn global representations directly, completely ignoring the discriminative local clues.

The latter ones first abstract more fine-grained representations of local fragments (*e.g.*, visual regions and words), and then aggregate local similarities to infer image-text alignments. For example, Li *et al.* [18] adopted a co-attention mechanism to model the relations between image regions and words for alignment. Differently, Nam *et al.* [26] built the Dual Attention Networks which attend to image regions and words iteratively. Specifically, the visual regions used in these two approaches are both obtained directly from feature maps, they hence share the same size and shape. However, it impedes the flexibility of visual representation. Inspired by the cognitive process of bottom-up attention mechanism in the human vision system, Anderson *et al.* [1] presented a bottom-up attention implemented with Faster R-CNN [28] to model the relevance between words and visual objects, and it has gained promising performance in image captioning and VQA. Motivated by this, recent research attention has been paid to designing cross-modal attention scheme, aiming to explore fine-grained matching [22, 23]. For instance, Lee *et al.* [16] presented a cross attention to align both image regions and words to infer image-text similarity. Wang *et al.* [37] proposed a cross-modal adaptive message passing method to perform fine-grained interaction and filter irrelevant information with a gating strategy. Liu *et al.* [21] designed a focal attention network comprising the pre-assigning and re-assigning attention, which focuses on eliminating irrelevant fragments. Although these approaches have considered the multi-view description problem to some extent on the ground of another modality, they suffer from extremely high computational complexity and overlook the context information.

[1] https://acmmmcamera.wixsite.com/camera.

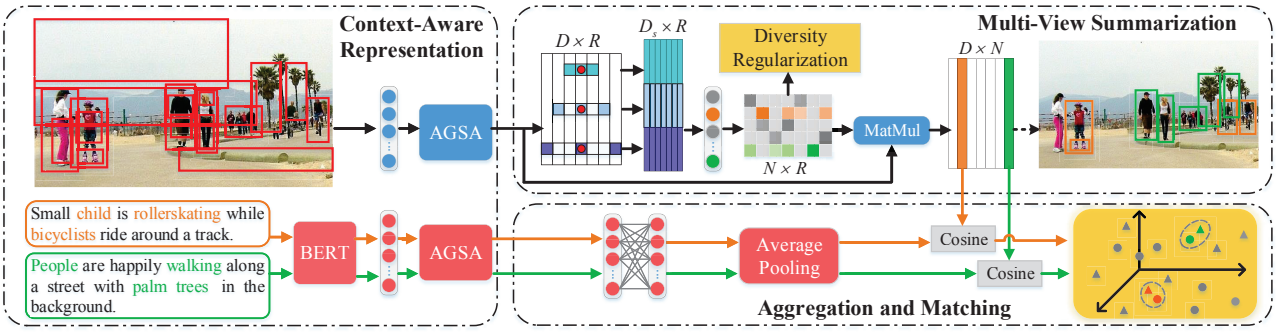


Figure 2: Schematic illustration of the CAMERA model. It comprises three components: 1) *Image Embedding*. The region-level features extracted from the bottom-up attention are enhanced by the visual AGSA module. And then the multi-view summarization module is applied to obtaining view-specific image features. 2) *Text Embedding*. The BERT followed by textual AGSA and average pooling operation is used to obtain context-enhanced sentence embeddings. And 3) *Loss Function*. Apart from using triplet loss for alignment learning, the diversity regularization is designed to restrict the information overlap.

2.2 Intra-modal Context Modeling

Context information plays a pivotal role in understanding a sentence for many natural language processing tasks, such as neural machine translation [4, 32], text summarization [29], and question answering [27]. Analogously, visual contextual relationships can contribute to obtaining fine-grained image region representations, which would benefit various tasks including image captioning [41], VQA [7], and image-text matching [17, 31, 35, 40]. To exploit visual and textual context and capture implicit relations among intra-modal fragments, researchers have presented some structured models for different multi-modal tasks. In particular, Yao *et al.* [41] built a hierarchical tree from the instance level, region level to the whole image level, and employed a Tree-LSTM to enhance these multi-level visual features for image captioning. Gao *et al.* [7] explored intra- and inter-modality relations with a dynamic fusion approach to boosting the performance of VQA. In the field of image-text matching, Li *et al.* [17] performed local-global semantic reasoning by using Graph Convolutional Network (GCN) and Gated Recurrent Unit. For learning comprehensive representations, Wang *et al.* [35] and Shi *et al.* [31] refined visual relationships by leveraging external scene graphs [13]. Wu *et al.* [40] considered fragment relations in images and texts to obtain self-attention embeddings, acquiring promising intra-modal context modeling. Although the aforementioned image-text matching methods have gained performance improvement by exploiting context, they lack effective control over the internal information flow.

3 OUR PROPOSED METHOD

As shown in Figure 2, our proposed CAMERA comprises three components: the image embedding branch (Section 3.2), the text embedding branch (Section 3.3), and the loss function (Section 3.4). Considering that both the visual and textual branches are based on the AGSA module, we hence first introduce this module (Section 3.1), and then successively elaborate the above three components.

3.1 Adaptive Gating Self-Attention

To take advantage of the intra-modal context information, we present the AGSA module to enhance the representations of images and texts. As illustrated in Figure 3, it mainly contains two parts, *i.e.*, the multi-head self-attention and the gate mechanism.

3.1.1 Multi-Head Self-Attention. Let $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ respectively denote the query, key, and value, where d_k and d_v represent dimensions of them, n refers to the sequence length. The self-attention mechanism can be formulated as follows,

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}, \tag{1}$$

where the $Softmax$ operation is performed for each row. To further strengthen the representation discrimination, the multi-head self-attention comprising H paralleled self-attention mechanisms is designed to capture context information from different subspaces,

$$\mathbf{h}_i = Attention(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i), \tag{2}$$

where $\mathbf{h}_i$ refers to the output of the i-th head, and

$$\begin{cases} \mathbf{Q}_i = \mathbf{X}\mathbf{W}_i^Q, \\ \mathbf{K}_i = \mathbf{X}\mathbf{W}_i^K, \\ \mathbf{V}_i = \mathbf{X}\mathbf{W}_i^V, \end{cases} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the input, n and d are respectively the sequence length and dimension of $\mathbf{X}$. $\mathbf{W}_i^Q \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}_i^V \in \mathbb{R}^{d \times d_v}$ are learnable projection matrices for query, key, and value, respectively. Note that in this work, we set $d_k = d_v = d/H$.

3.1.2 Gate Mechanism. The aforementioned self-attentions exploit the intra-modal interactions by calculating the dot-product similarities between queries and keys. However, the projected queries and keys may contain noisy or meaningless information. To adaptively pass the informative messages and restrain the useless ones, we design an adaptive gate mechanism with fusion strategy. Specifically, for the i-th head, we first map $\mathbf{Q}_i \in \mathbb{R}^{n \times d_k}$ and $\mathbf{K}_i \in \mathbb{R}^{n \times d_k}$ into a joint space, and then perform the fusion operation. This process is represented as,

$$\mathbf{G}_i = (\mathbf{Q}_i\mathbf{W}_G^Q + \mathbf{b}_G^Q) \odot (\mathbf{K}_i\mathbf{W}_G^K + \mathbf{b}_G^K), \tag{4}$$

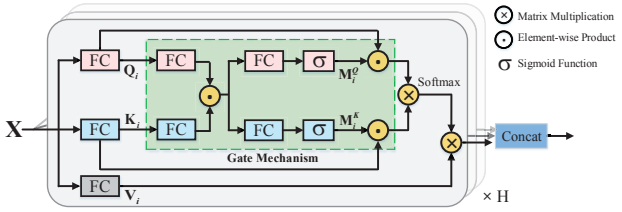


Figure 3: Illustration of our proposed AGSA module.

where $\mathbf{G}_i \in \mathbb{R}^{n\times d_k}$ is the fusion result, $\odot$ denotes element-wise product operation, $\mathbf{W}_G^Q$ and $\mathbf{W}_G^K \in \mathbb{R}^{d_k\times d_k}$ are the learnable projection matrices, while $\mathbf{b}_G^Q$ and $\mathbf{b}_G^K \in \mathbb{R}^{1\times d_k}$ are the bias vectors.

Thereafter, the gating masks $\mathbf{M}_i^Q$, $\mathbf{M}_i^K \in \mathbb{R}^{n\times d_k}$ corresponding to $\mathbf{Q}_i$ and $\mathbf{K}_i$ are generated by adopting two fully-connected layers followed by sigmoid function,

$$\begin{cases}\mathbf{M}_i^Q = \sigma(\mathbf{G}_i\mathbf{W}_M^Q + \mathbf{b}_M^Q),\\ \mathbf{M}_i^K = \sigma(\mathbf{G}_i\mathbf{W}_M^K + \mathbf{b}_M^K),\end{cases} \tag{5}$$

where $\sigma(\cdot)$ denotes sigmoid function, and $\mathbf{W}_M^Q$, $\mathbf{W}_M^K$, $\mathbf{b}_M^Q$, $\mathbf{b}_M^K$ are the parameters of fully-connected layers.

Finally, these obtained gating masks are applied to controlling the information flow of original $\mathbf{Q}_i$ and $\mathbf{K}_i$, and the dot-product attention is conducted instead of Eqn.(2), as,

$$\tilde{\mathbf{h}}_i = Attention(\mathbf{M}_i^Q \odot \mathbf{Q}_i, \mathbf{M}_i^K \odot \mathbf{K}_i, \mathbf{V}_i), \tag{6}$$

where $\tilde{\mathbf{h}}_i \in \mathbb{R}^{n\times d_v}$ is the gating self-attention result of the i-th head. By concatenating the results of multiple heads, we could obtain the context-enhanced representation of the input. Formally, the AGSA module is formulated as,

$$F_{AGSA}(\mathbf{X}) = Concat(\tilde{\mathbf{h}}_1, ..., \tilde{\mathbf{h}}_H) + \mathbf{X}. \tag{7}$$

3.2 Image Embedding

Considering that an image often conveys more information than a sentence, we exploit fine-grained multi-view representations of images to solve the multi-view description problem. Particularly, we first extract the region-level features of an image, and then refine the features with the visual AGSA to obtain the context-aware region representations. Finally, the summarization module is proposed to summarize the region representations into image-level features of different views.

3.2.1 Bottom-up Feature Extraction. Following [16, 17], we detect salient regions with the bottom-up attention[2] [1]. To be specific, given an image I, we select the top R (R=36) ROIs with the highest class confidence scores. Moreover, we adopt average pooling to obtain the feature of the i-th region $\mathbf{f}_i \in \mathbb{R}^{2048}$, and its position information can be represented by $\mathbf{p}_i = (x_i, y_i, w_i, h_i)$, where x_i and y_i are top left coordinates, and w_i and h_i denote width and height of the corresponding region, respectively. We then utilize a fully-connected layer to project each region feature $\mathbf{f}_i$ as follows,

$$\mathbf{v}_i = \mathbf{W}_v\mathbf{f}_i + \mathbf{b}_v, \tag{8}$$

where $\mathbf{W}_v \in \mathbb{R}^{D\times 2048}$ and $\mathbf{b}_v$ are the learnable parameters.

[2]This attention model is pre-trained on the Visual Genomes dataset [15] based on Faster R-CNN [28], where ResNet-101 [9] is applied as the backbone.

To capture the spatial characteristics from a global perspective, we select absolute position encoding instead of relative position encoding [36]. Specifically, for a raw position vector $\mathbf{p}_i$, we add two extra dimensions (*i.e.*, aspect ratio and area) and normalize them to obtain the new position vector $\hat{\mathbf{p}}_i \in \mathbb{R}^6$,

$$\hat{\mathbf{p}}_i = (\frac{x_i}{w}, \frac{y_i}{h}, \frac{w_i}{w}, \frac{h_i}{h}, \frac{w_i}{h_i}, \frac{w_i h_i}{wh}), \tag{9}$$

where w and h are respectively the width and height of the image I. Thereafter, we adopt a fully-connected layer followed by a sigmoid function for absolute position encoding as,

$$\tilde{\mathbf{p}}_i = \sigma(\mathbf{W}_p\hat{\mathbf{p}}_i + \mathbf{b}_p), \tag{10}$$

where $\mathbf{W}_p \in \mathbb{R}^{D\times 6}$ and $\mathbf{b}_p$ are the weights and bias, respectively.

3.2.2 Context-aware Region Representation. To exploit the semantic and spacial relations among visual regions, we instantiate the AGSA module to capture the informative context. More concretely, we pack the semantic embeddings of image regions into a matrix $\mathbf{V} \in \mathbb{R}^{R\times D}$, and their position embeddings into another matrix $\tilde{\mathbf{P}} \in \mathbb{R}^{R\times D}$. Thereafter, a fusion operation is conducted. Taking the fused region-level features as input, the instantiated visual AGSA performs the intra-modal reasoning as follows,

$$\bar{\mathbf{V}} = F_{AGSA}(\mathbf{V} \odot \tilde{\mathbf{P}}), \tag{11}$$

where $\bar{\mathbf{V}} \in \mathbb{R}^{R\times D}$ is the context-enhanced region feature matrix.

3.2.3 Multi-view Summarization. Understanding an image from different views means that each region acquires different attention from different views, *i.e.*, view-specific importance. In this paper, we adopt the pyramid dilated convolution [43] to model the regional correlations, based on which the view-specific importance scores are calculated.

Given the context-enhanced region feature $\bar{\mathbf{v}}_i \in \mathbb{R}^D$, we utilize the dilated convolution to calculate its importance vector. As the dilation rate increases, the receptive field of the kernel is enlarged without reducing the regional resolution[3]. For example, if the dilation rate increases from 1 to 2, the receptive field will increase from 3 to 5 for a kernel with the size 3. As the dilated convolution can aggregate multi-scale contextual information by using different factors, *i.e.*, dilation rate, we design a pyramid dilated convolution layer with K parallel kernels, as reported in Table 1. The outputs of these kernels are concatenated as,

$$\mathbf{s}_i = Concat(\mathbf{s}_i^1, ..., \mathbf{s}_i^K), \tag{12}$$

where $\mathbf{s}_i^k$ represents the output of the k-th kernel.

Afterwards, a fully-connected layer followed by softmax is adopted to compute the summarization matrix $\bar{\mathbf{S}} = [\bar{\mathbf{s}}_1; ...; \bar{\mathbf{s}}_R] \in \mathbb{R}^{R\times N}$, where $\bar{\mathbf{s}}_i$ is the i-th row vector, and N is the total number of views. Formally, we summarize the above process as follows,

$$\begin{cases}\tilde{\mathbf{s}}_i = \mathbf{W}_s\mathbf{s}_i + \mathbf{b}_s,\\ \bar{s}_{ij} = \frac{\exp(\tilde{s}_{ij})}{\sum_{i=1}^R \exp(\tilde{s}_{ij})},\end{cases} \tag{13}$$

where $\tilde{\mathbf{s}}_i \in \mathbb{R}^N$ represents the intermediate importance scores of the i-th region over N views, $\mathbf{W}_s \in \mathbb{R}^{N\times D_s}$ and $\mathbf{b}_s \in \mathbb{R}^{1\times D_s}$ are

[3]When the dilation rate equals to one, the dilation convolution is equivalent to the standard convolution.

Table 1: Pyramid dilated convolution configurations, where w^k, r^k, c^k denote kernel size, dilation rate and output channel of k-th convolution kernel, respectively.

k	1	2	3	4	5	6	7
w^k	1	3	3	3	5	5	5
r^k	1	1	2	3	1	2	3
c^k	256	128	128	128	128	128	128

respectively the weights and bias. Finally, the region-level features are summarized into image-level representations as follows,

$$\mathbf{V}^* = \bar{\mathbf{S}}^\top \bar{\mathbf{V}}, \tag{14}$$

where $\mathbf{V}^* \in \mathbb{R}^{N \times D}$ is the final multi-view image feature matrix.

3.3 Text Embedding

As for the textual branch, we use the pre-trained BERT [4] to extract the context-enhanced word-level embeddings. Concretly, the given text containing L words is first tokenized by WordPiece tokenizer, and then the corresponding word features are extracted by BERT. The corresponding output can be denoted by $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_L]$, where $\mathbf{e}_i \in \mathbb{R}^{768}$. Afterwards, we utilize a fully-connected layer to transform $\mathbf{e}_i$ to a D-dimensional vector as,

$$\mathbf{t}_i = \mathbf{W}_t \mathbf{e}_i + \mathbf{b}_t, \tag{15}$$

where $\mathbf{W}_t \in \mathbb{R}^{D \times 768}$ and $\mathbf{b}_t$ are the weights and bias, respectively.

In order to capture high-level contextual relations and improve the flexibility of our model, the textual AGSA is employed to $\mathbf{T} = [\mathbf{t}_1, , ..., \mathbf{t}_L]$ as follows,

$$\bar{\mathbf{T}} = F_{AGSA}(\mathbf{T}), \tag{16}$$

where $\bar{\mathbf{T}} \in \mathbb{R}^{L \times D}$ is the word feature enhanced by AGSA. Note that the two AGSA modules in the image embedding branch and text embedding branch are independent, namely the parameters are not shared. In order to further improve non-linear representation capability, we add a multi-layer perceptron (MLP) followed by a residual connection to obtain the final word embedding $\tilde{\mathbf{t}}_i$, *i.e.*,

$$\tilde{\mathbf{t}}_i = (\mathbf{W}_2[\mathbf{W}_1 \bar{\mathbf{t}}_i + \mathbf{b}_1]_+ + \mathbf{b}_2) + \bar{\mathbf{t}}_i, \tag{17}$$

where $[x]_+ = max(x, 0)$, $\mathbf{W}_1 \in \mathbb{R}^{D \times D}$ and $\mathbf{W}_2 \in \mathbb{R}^{D \times D}$ are the weights, while $\mathbf{b}_1$ and $\mathbf{b}_2$ are the biases. Finally, we use average pooling to aggregate the word features into a sentence feature,

$$\mathbf{z} = \frac{1}{L} \sum_{i=1}^{L} \tilde{\mathbf{t}}_i, \tag{18}$$

where $\mathbf{z} \in \mathbb{R}^D$ is the final context-enhanced sentence feature vector.

3.4 Loss Function

3.4.1 Bidirectional Triplet Ranking Loss. To achieve multi-view matching, we define the cross-modal scoring function to measure the similarity of the image-text pair (I, T) as,

$$s(I, T) = \max_{i=1,\ldots,N} \cos(\mathbf{v}_i^*, \mathbf{z}), \tag{19}$$

where $\cos(\cdot, \cdot)$ is the cosine similarity between two vectors.

Having obtained the matching score, we adopt the bidirectional triplet ranking loss with hard negatives to enforce the alignment between the image-text pair (I, T), which is defined as,

$$L_R = [\alpha - s(I, T) + s(I, \hat{T})]_+ + [\alpha - s(I, T) + s(\hat{I}, T)]_+, \tag{20}$$

where α is the margin parameter, $\hat{T} = argmax_{j \neq T} s(I, j)$ and $\hat{I} = argmax_{i \neq I} s(i, T)$ are the hardest negatives in a mini-batch.

3.4.2 Diversity Regularization. Although it is inevitable that there will be overlaps among different understanding views, *e.g.*, the word "*trampoline*" appears in all the three descriptions in Figure 1, too much redundancy is unexpected. To avoid the redundancy and ensure the diversity of multiple summarization vectors, we introduce a diversity regularization, which is formulated as,

$$L_D = ||\hat{\mathbf{S}}^\top \hat{\mathbf{S}} - \mathbf{U}||_F^2, \tag{21}$$

where $\mathbf{U} \in \mathbb{R}^{N \times N}$ is an unit matrix, $\hat{\mathbf{S}} \in \mathbb{R}^{R \times N}$ is the L2-normalization result of the multi-view intermediate importance matrix $\tilde{\mathbf{S}} \in \mathbb{R}^{R \times N}$, and $|| \cdot ||_F$ denotes the frobenius norm of matrix.

Finally, we merge the triplet loss and the diversity regularization to obtain our final loss as,

$$L = L_R + \lambda L_D, \tag{22}$$

where λ serves as the trade-off parameter.

4 EXPERIMENTS

To justify the effectiveness of our CAMERA model, we carried out experiments in terms of the cross-modal retrieval involving 1) Image-to-Text, *i.e.*, retrieving sentences that can well depict the content of a given image; and 2) Text-to-Image, *i.e.*, retrieving images that are semantically consistent with a given text query.

4.1 Datasets and Evaluation Metric

Flickr30K [42]: It consists of 31,783 images collected from the Flickr website[4], and each image is associated with 5 sentences. Following the settings in [5, 16, 40], we utilized 1,000 images for validation, 1,000 images for testing, and the rest for training.

MS-COCO [19]: This dataset contains 123,287 images, each of which corresponds to 5 manually annotated sentences. Similarly, we followed the splits of [5, 8, 16], namely 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Likewise, we adopted two kinds of evaluation settings: 1) **MS-COCO 1K**, averaging results from 5 folds of 1K test images; and 2) **MS-COCO 5K**, testing on the full 5K test images.

We measured the retrieval performance with the common evaluation metric in information retrieval, namely Recall@K (K=1, 5, 10). It is defined as the fraction of queries that correctly retrieve desired items in the top-K of ranking list.

4.2 Implementation Details

To train our model, we employed Adam [14] as the optimizer. The mini-batch size is set 128, and the learning rate is set as 0.0001. As for Flickr30K, the learning rate is divided by 10 every 10 epochs and we utilized 30 epochs. As for MS-COCO, it is divided by 10 every 20 epochs and we used 40 epochs. Besides, the dimension of the joint embedding space is 2,048. To obtain the original word embeddings, we used and freezed the basic version of the pre-trained BERT [4], which has 12 layers, 12 heads, 768 hidden units and 110M parameters in total. Besides, the visual and textual AGSA modules have the same settings, where 1 layer and 64 heads are considered.

[4] https://www.flickr.com/.

Table 2: Performance comparison between our proposed CAMERA and the state-of-the-art baselines on Flickr30K. Our results are highlighted in bold.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
(VGG)						
VQA [20]	33.9	62.5	74.5	24.9	52.6	64.8
sm-LSTM [11]	42.5	71.9	81.5	30.2	60.4	72.3
SAN [12]	67.0	88.0	94.6	51.4	77.2	85.2
SCG [31]	71.8	90.8	94.8	49.3	76.4	85.6
(ResNet)						
CMPM + CMPC [44]	49.6	76.8	86.1	37.3	65.7	75.5
VSE++ [5]	52.9	80.5	87.2	39.6	70.1	79.5
TIMAM [30]	53.1	78.8	87.6	42.6	71.6	81.9
SAN [12]	75.5	92.6	96.2	60.1	84.7	90.6
(Faster R-CNN)						
SCAN (ensemble) [16]	67.4	90.3	95.8	48.6	77.7	85.2
CAMP [37]	68.1	89.7	95.2	51.5	77.1	85.3
BFAN (ensemble) [21]	68.1	91.4	-	50.8	78.4	-
SAEM [40]	69.1	91.0	95.1	52.4	81.1	88.1
PFAN (ensemble) [36]	70.0	91.8	95.0	50.4	78.7	86.1
VSRN (ensemble) [17]	71.3	90.6	96.0	54.7	81.8	88.2
SGM [35]	71.8	91.7	95.5	53.5	79.6	86.5
CAMERA (single)	76.5	95.1	97.2	58.9	84.7	90.2
CAMERA (ensemble)	**78.0**	**95.1**	**97.9**	**60.3**	**85.9**	**91.7**

In the summarization module, the number of summarization views N is set to 12. Moreover, the margin α in Eqn.(20) is set as 0.2 and the trade-off parameter λ in Eqn.(22) is set as 0.01.

4.3 Performance Comparison

To justify the effectiveness of our proposal, we compared our proposed CAMERA with the following state-of-the-art baselines: VQA [20], sm-LSTM [11], VSE++ [5], CMPM + CMPC [44], SCAN [16], SCG [31], TIMAM [30], SAEM [40], CAMP [37], BFAN [21], PFAN [36], SAN [12], VSRN [17], and SGM [35]. Note that we directly quoted the results from their original papers. According to the used visual backbones, we divided these methods into three groups, *i.e.*, VGG, ResNet and Faster R-CNN. Since some of them are ensemble models, namely averaging similarity scores of two single models trained independently, we also reported our ensemble results by combining two same single models for fair comparisons.

Table 2 summarizes the performance comparison results on the Flickr30K dataset. We could see that our CAMERA model achieves the best performance, substantially surpassing all the baselines. It displays consistent improvements over the state-of-the-art method VSRN [17] with the same backbone. Particularly, it obtains 9.4% and 10.2% relative gains on R@1 for image-to-text and text-to-image retrieval, respectively. Besides, our method also gains an advantage over the recent SAN [12]. Therefore, it can be seen from the above comparisons that the CAMERA model for context modeling and multi-view visual understanding has great superiority.

Table 3 and Table 4 present retrieval results under the MS-COCO 1K and 5K testing settings, respectively. By jointly analyzing them, we found that our CAMERA is superior to the baselines with the same backbone setting (*i.e.*, Faster R-CNN with ResNet-101). Moreover, SAN [12] outperforms other baselines including ours on MS-COCO. We believed its major gain comes from the deeper backbone ResNet-152 and an extra saliency detector. Nevertheless, its guiding interaction strategy significantly increases the complexity of retrieval during the inference, confining its efficiency on large-scale data. In contrary, our CAMERA obtains representations from two independent branches, which enables our model to extract features of two modalities in parallel effectively.

Table 3: Performance comparison between our proposed CAMERA and the state-of-the-art baselines in terms of the MS-COCO 1K setting. Our results are highlighted in bold.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
(VGG)						
VQA [20]	50.5	80.1	89.7	37.0	70.9	82.9
sm-LSTM [11]	53.2	83.1	91.5	40.7	75.8	87.4
SAN [12]	74.9	94.9	98.2	60.8	90.3	95.7
SCG [31]	76.6	96.3	99.2	61.4	88.9	95.1
(ResNet)						
CMPM + CMPC [44]	56.1	86.3	92.9	44.6	78.8	89.0
VSE++ [5]	64.6	90.0	95.7	52.0	84.3	92.0
SAN [12]	85.4	97.5	99.0	69.1	93.4	97.2
(Faster R-CNN)						
SAEM [40]	71.2	94.1	97.7	57.8	88.6	94.9
CAMP [37]	72.3	94.8	98.3	58.5	87.9	95.0
SCAN (ensemble) [16]	72.7	94.8	98.4	58.8	88.4	94.8
SGM [35]	73.4	93.8	97.8	57.5	87.3	94.3
BFAN (ensemble) [21]	74.9	95.2	-	59.4	88.4	-
VSRN (ensemble) [17]	76.2	94.8	98.2	62.8	89.7	95.1
PFAN (ensemble) [36]	76.5	96.3	99.0	61.6	89.6	95.2
CAMERA (single)	75.9	95.5	98.6	62.3	90.1	95.2
CAMERA (ensemble)	**77.5**	**96.3**	**98.8**	**63.4**	**90.9**	**95.8**

Table 4: Performance comparison between our proposed CAMERA and the state-of-the-art baselines in terms of the MS-COCO 5K setting. Our results are highlighted in bold.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
(VGG)						
VQA [20]	23.5	50.7	63.6	16.7	40.5	53.8
SCG [31]	56.6	84.5	92.0	39.2	68.0	81.3
(ResNet)						
CMPM + CMPC [44]	31.1	60.7	73.9	22.9	50.2	63.8
VSE++ [5]	41.3	71.1	81.2	30.3	59.4	72.4
SAN [12]	65.4	89.4	94.8	46.2	77.4	86.6
(Faster R-CNN)						
SGM [35]	50.0	79.3	87.9	35.3	64.9	76.5
CAMP [37]	50.1	82.1	89.7	39.0	68.9	80.2
SCAN (ensemble) [16]	50.4	82.2	90.0	38.6	69.3	80.4
VSRN (ensemble) [17]	53.0	81.1	89.4	40.5	70.6	81.1
CAMERA (single)	53.1	81.3	89.8	39.0	70.5	81.5
CAMERA (ensemble)	**55.1**	**82.9**	**91.2**	**40.5**	**71.7**	**82.5**

4.4 Ablation Study

The Adaptive Gating Self-Attention Module. To explore how the AGSA module affects the matching results, we designed two variants: 1) **w/o AGSA** refers to the model removing the whole AGSA module in both image and text branches; and 2) **w/o Gate** denotes the model excluding the gating mechanism in two branches. Their performance in Table 5 indicates that removing either the AGSA or the gating mechanism hurts the expressiveness of features and further degrades the performance. Meanwhile, the large performance degradation brought by w/o AGSA manifests the importance of capturing intra-modal context information.

To delve into how AGSA contributes to the performance improvement in two branches, we further designed the following

Table 5: Performance comparison among our model variants on Flickr30K. The results of the full model are highlighted in bold.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CAMERA	**76.5**	**95.1**	**97.2**	**58.9**	**84.7**	**90.2**
(Adaptive Gating Self-Attention)						
w/o AGSA	68.3	90.6	95.3	51.3	79.7	87.3
w/o Gate	76.2	93.0	97.1	57.5	84.0	90.4
w/o Vis AGSA	70.2	91.1	95.4	52.5	80.6	87.4
w/o Txt AGSA	72.6	92.5	96.9	56.4	83.4	89.7
Vis GCN	72.0	92.0	96.6	54.4	81.9	89.0
Txt BiGRU	75.0	93.5	96.7	57.3	83.5	90.1
Txt 1D-CNN	74.8	93.5	96.7	57.8	83.7	90.5
(Multi-View Summarization)						
Avg Pool	74.5	93.9	96.8	57.5	84.1	90.2
Region Max	74.6	94.3	97.5	57.9	84.8	90.5
(Diversity Regularization)						
w/o Diversity	75.2	93.8	96.6	58.0	84.0	90.1

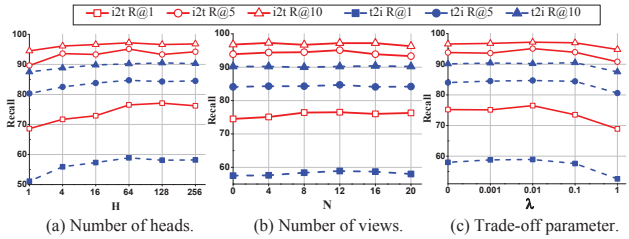
variants: 1) **w/o Vis AGSA** is the model without the visual AGSA. 2) **w/o Txt AGSA** is the model without the textual AGSA. 3) **Vis GCN** replaces the visual AGSA with the 4-layer GCN [17] for visual context modeling. 4) **Txt BiGRU** is a model in which we replaced the textual AGSA with bidirectional GRU. And 5) **Txt 1D-CNN** indicates utilizing 1D-CNN [40] as an alternative of the textual AGSA, the MLP, and the average pooling for aggregation. First, it can be seen from the comparison results summarized in Table 5 that both w/o Vis AGSA and w/o Txt AGSA achieve poor performance since they overlook the intra-modal context information modeling. Furthermore, CAMERA outperforms Vis GCN, Txt BiGRU and Txt 1D-CNN, which can be attributed to the fine-grained intra-modal interactions and the flexible control of internal information flow achieved by adaptive gating mechanism.

The Multi-View Summarization Module. To justify the effectiveness of the multi-view summarization module, we experimented with two variants of our model: 1) **Avg Pool** denotes that we directly conducted average pooling to region features for obtaining image-level features; and 2) **Region Max** means that the maximum region-sentence similarity is selected as the image-sentence similarity. By jointly analyzing their results in Table 5, we could see that CAMERA displays consistent improvements over them, especially in terms of R@1, verifying the crucial influence of the multi-view summarization module.

The Diversity Regularization. To validate the significance of diversity regularization, we carried out the experiment by removing the diversity regularization term in Eqn.(21), dubbed **w/o Diversity**. The results in Table 5 indicate that properly constraining the diversity of summarization matrix can improve the performance. Meanwhile, w/o Diversity still outperforms Avg Pool and Region Max in terms of R@1 though there is major redundancy in the summarization matrix. This further demonstrates the superiority of our proposed multi-view summarization module and the effectiveness of aggregating region features according to their importance scores.

4.5 Parameter Sensitivity

We carried out experiments to explore how the number of heads H in AGSA, the number of summarization views N, and the trade-off parameter λ affect the matching performance. By analyzing the results shown in Figure 4, we have the following observations: 1) The matching performance improves with the increasing number of heads, as shown in Figure 4(a). The observed results make sense since more heads can lead to more comprehensive interaction. When the number increases to 64, the performance reaches the saturation point. Therefore, considering the comprehensive performance and efficiency, we set the number of heads to 64. 2) As illustrated in Figure 4(b), aggregating region features from different views can improve matching performance to some extent, yet too many views are unnecessary. This may be because the information contained in an image is limited. And 3) from Figure 4(c), we could see that our model obtains the best performance when $\lambda = 0.01$. Besides, larger λ can degrade performance due to the lack of primary information in some views and the weakened ranking constraint. In a word, the performance of our model fluctuates slightly around the optimal settings, which verifies the robustness of our model.



(a) Number of heads. (b) Number of views. (c) Trade-off parameter.

Figure 4: Performance of our CAMERA with different (a) numbers of heads in AGSA, (b) numbers of summarization views, and (c) trade-off parameters on Flickr30K.

4.6 Visualization of Multi-View Summarization

To intuitively assess how each region contributes to the final image representation, we utilized the visualization strategy similar to VSRN [17]. Concretely, for a summarization vector $\tilde{\mathbf{S}}_{[:,j]} \in \mathbb{R}^R$ of j-th view, where R is the number of regions, we ranked its elements in the descending order. Afterwards, we assigned an attention score a_i to each region i according to its rank r_i. The score is calculated by $a_i = \beta(R - r_i)^2$, where β is a parameter used to emphasize the highly ranked regions[5]. Finally, the attention score at each pixel is calculated by adding up scores of all regions it belongs to.

Several visualization results are shown in Figure 5. From these results, we could see that our summarization module indeed understands an image from different views. For example, in Figure 5(a), the "seagulls" and "truck" in the first view receive more visual attention, while the second view pays more attention to the details of the vehicle type and color, *i.e.*, "Orange SUV". And more entities ("black sweater", "cap", "paper" and "wineglass") are considered in the first view of Figure 5(c) than those ("black cat" and "wine") of the second. Besides, in addition to the main characters ("girl", "lady" and "boy") of the first view in Figure 5(b) and Figure 5(d), more attention is paid to "walls" and "trees" in their second views. Thus it can be seen that our summarization module can represent an image according to different understanding preferences. Furthermore, it can not only emphasize the premier regions, but also ignore trivial ones in some specific views.

[5]In this work, we set β to 50.

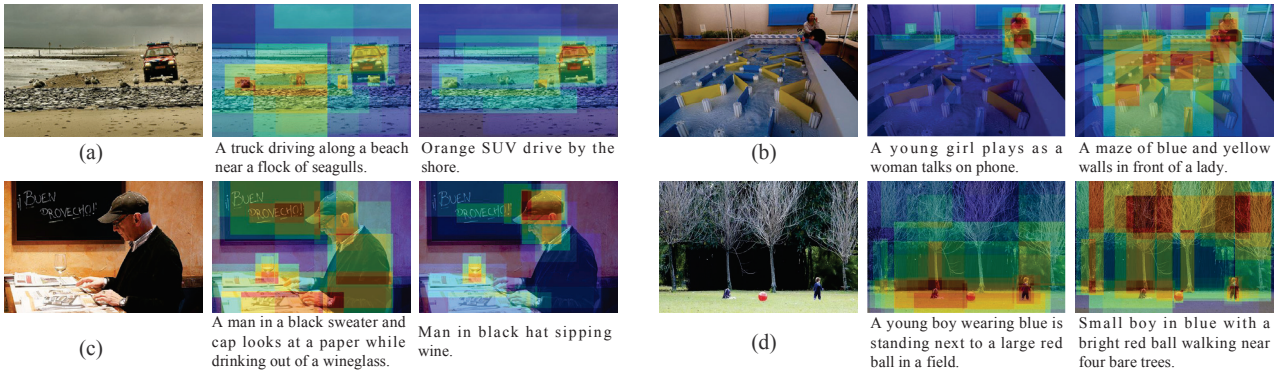
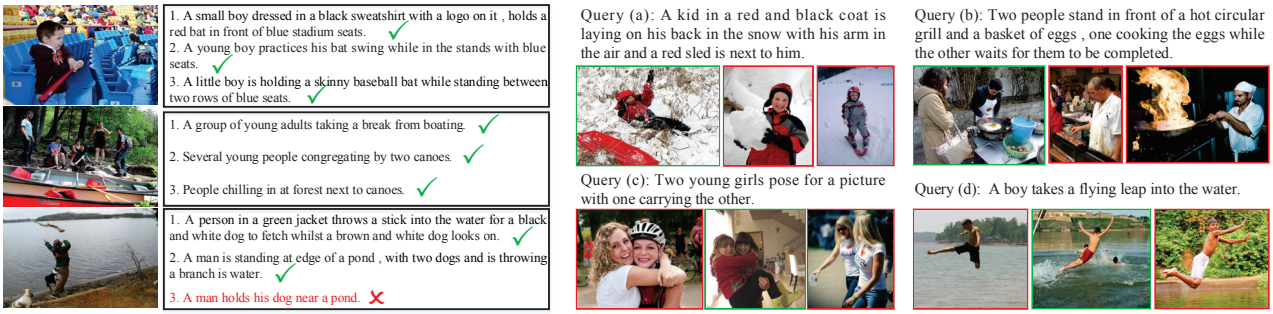


Figure 5: Visualization of our multi-view summarization module on Flickr30K. For each group, we respectively showed the raw image and two attention maps of different views with the matched sentences from left to right.



(a) Image-to-Text retrieval results.

(b) Text-to-Image retrival results.

Figure 6: Qualitative results of CAMERA on Flickr30K. The top-3 retrieved results are shown for each query. Green check indicates the ground-truth sentences, while red cross marks for wrong results. Moreover, the ground-truth images are outlined in green boxes, while the false ones are in red.

4.7 Qualitative Results

To gain deep insights into our proposed CAMERA, we listed some examples of image-to-text retrieval and text-to-image retrieval on Flickr30K. Figure 6 shows that our model is robust to complex images and long sentences, achieving promising retrieval results. Although our model misunderstands the relation between the man and the black dog in the third example of Figure 6(a), the entities "man", "dog" and "pond" are still semantically matched. This may be attributed to the ambiguous spacial positions or the excessive overlap of bounding boxes between the "man" and the "dog". As for text-to-image retrieval in Figure 6(b), the top-1 results of Query (c) and Query (d) are incorrect but reasonable, because there are few semantic differences between our results and the ground-truth images. The above qualitative results demonstrate that our CAMERA can achieve consistent representation and accurate matching. Furthermore, our model discriminates some details among unmatched pairs, *e.g.*, "laying on his back in the snow" in Query (a). This may be owing to the adaptive context capturing and the comprehensive summarization of multiple views.

5 CONCLUSION AND FUTURE WORK

In this paper, we present a novel CAMERA model to tackle the multi-view description problem for image-text matching. Concretely, we incorporate self-attention and gating mechanism to adaptively extract context-enhanced fragment features. According to different understanding preferences, we design a multi-view summarization module to summarize visual regions from different views. Besides, we introduce a diversity regularization to mitigate the information redundancy. Extensive experiments on Flickr30K and MS-COCO datasets have demonstrated the superiority of our model to a wide range of state-of-the-art methods.

In the future, we will explicitly explore the spatial and semantic relations, the high-level semantic concepts, as well as their dependency and interplay from different views. Moreover, for a more comprehensive multimodal understanding, it may be necessary to adaptively perform fine-grained cross-modal interaction according to the complexity of images and texts.

ACKNOWLEDGMENTS

This work is supported by the Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014; the National Natural Science Foundation of China, No.:U1936203; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23. This work is also supported by the special fund for distinguished Professors of Shandong Jianzhu University.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6077–6086.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2425–2433.

[3] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-Modal Image-Text Retrieval with Semantic Consistency. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1749–1757.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 1–13.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 2121–2129.

[7] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. In *Proceedings of the IEEE ConfCerence on Computer Vision and Pattern Recognition*. IEEE, 6639–6648.

[8] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7181–7189.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[10] Yupeng Hu, Peng Zhan, Yang Xu, Jia Zhao, Yujun Li, and Xueqing Li. 2020. Temporal representation learning for time series classification. *Neural Computing and Applications* (2020), 1–14.

[11] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware Image and Sentence Matching with Selective Multimodal LSTM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2310–2318.

[12] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. 2019. Saliency-Guided Attention Network for Image-Sentence Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5754–5763.

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image Retrieval using Scene Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3668–3678.

[14] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd Interncational Conference on Learning Representations*. 1–15.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[16] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. Springer, 201–216.

[17] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4654–4662.

[18] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-Aware Textual-Visual Matching with Latent Co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1890–1899.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[20] Xiao Lin and Devi Parikh. 2016. Leveraging Visual Question Answering for Image-Caption Ranking. In *Proceedings of the European Conference on Computer Vision*. Springer, 261–277.

[21] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 3–11.

[22] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 15–24.

[23] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-Modal Moment Localization in Videos. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 843–851.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 289–297.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations*. 1–12.

[26] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 299–307.

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. ACL, 2383–2392.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 91–99.

[29] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, 379–389.

[30] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial Representation Learning for Text-to-Image Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5814–5824.

[31] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 5182–5189.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 5998–6008.

[33] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 154–162.

[34] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5005–5013.

[35] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-Modal Scene Graph Matching for Relationship-Aware Image-Text Retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1508–1517.

[36] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position Focused Attention Network for Image-Text Matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 3792–3798.

[37] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5764–5773.

[38] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* (2019), 1–14.

[39] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1437–1445.

[40] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2088–2096.

[41] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy Parsing for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2621–2629.

[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[43] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proceedings of the 4th International Conference on Learning Representations*. 1–13.

[44] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. Springer, 686–701.