SMAN: Stacked Multi-Modal Attention Network for Cross-Modal Image-Text Retrieval

Zhong Ji*, *Member, IEEE,* Haoran Wang*, Jungong Han, Yanwei Pang, *Senior Member, IEEE*

***Abstract*—This paper focuses on tackling the task of cross-modal image-text retrieval, which has been an interdisciplinary topic in both computer vision and natural language processing communities. Existing global representation alignment based methods fail to pinpoint the semantically meaningful portion of images and texts, while the local representation alignment schemes suffer from the huge computation burden for aggregating the similarity of visual fragments and textual words exhaustively. In this study, we propose a Stacked Multi-modal Attention Network (SMAN) that makes use of stacked multi-modal attention mechanism to exploit the fine-grained inter-dependencies between image and text, thereby mapping the aggregation of attentive fragments into a common space for measuring cross-modal similarity. Specifically, we sequentially employ intra-modal information and multi-modal information as guidance to perform a multiple-step attention reasoning so that the fine-grained correlation between image and text can be modeled. As a consequence, we are capable of discovering the semantically meaningful visual regions or words in sentence, which contributes to measuring the cross-modal similarity in a more precise manner. Moreover, we present a novel bi-directional ranking loss that enforces the distance among pairwise multi-modal instances to be closer. Doing so allows us to make full use of pairwise supervised information to preserve the manifold structure of heterogeneous pairwise data. Extensive experiments on two benchmark datasets demonstrate that our SMAN consistently yields competitive performance compared to state-of-the-art methods.**

***Index Terms*—Vision and Language, Cross-modal Retrieval, Attention Mechanism, Multi-modal Learning**

I. Introduction

Cross-modal retrieval (CMR) [46], [15], [11], [18], [1], [3], [10] has been a fundamental and challenging topic in multimedia community, which benefits a variety of relative applications including image captioning [63], [52], visual question answering (VQA) [33], cross-modal hashing [62], [12], [13], [14], text-to-image Synthesis [19], and scene recognition [20]. Given a query instance from one modality, it aims at retrieving its counterpart from another modality. Among cross-modal retrieval tasks, image-text retrieval (also called image-text matching) plays a crucial role in bridging the gap between image and language understanding. Specifically, it is designed to search for images that are most relevant to the topic of a textual query, or captions that precisely describe the content of a visual query. However, solving the CMR problem is not easy, because data from different modalities separately reside in heterogeneous feature spaces, thus giving rise to difficulties in measuring the semantic relevances between cross-modal instances.

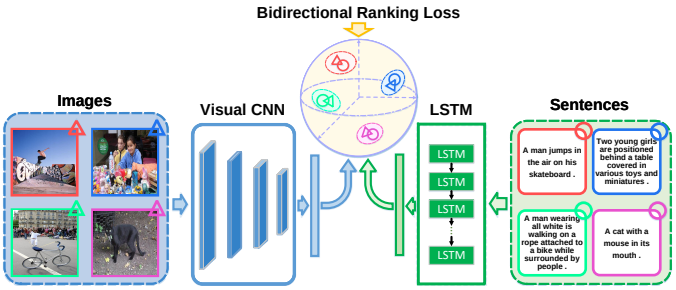


Fig. 1. Conceptual illustration of general deep image-text retrieval framework.

Recently, there has been a surge of work [29], [49], [44], [50], [17], [16], [19], [64] proposed to tackle the image-text retrieval problem. Under the umbrella of deep learning, the current predominant schemes opt to learn modality-specific deep features in a common space for both modalities. More concretely, they usually adopt a two-branch framework (as shown in Fig.1) carrying out two basic steps - visual branch (*e.g.*, Convolutional Neural Network (CNN)) and textual branch *(e.g.*, Long Short-Term Memory (LSTM)) extract visual and textural features respectively, followed by deploying an optimized objective (*e.g.*, bidirectional triplet ranking loss) to learn the joint embeddings. Although thrilling progresses [27], [9], [34], [56] have been achieved, due to the existence of heterogeneity gap, the cross-modal retrieval performance is still far from satisfactory. Therefore, the core issue for reducing the gap between image and text can be summarized as: *how to improve the discrimination of latent embeddings to align heterogeneous data in a common semantic space ?*

Most existing approaches [50], [49], [56] elect to learn global features for representing image and text, respectively. But this line of work neglects that the global similarity is commonly obtained by aggregating the local similarities between visual and textual instances (objects in an image and words in a text)[26], which results in the limited performance. Distinct from above approaches, another stride of work [29], [59] exhaustively aggregate similarity of all possible pairs of visual objects and textual words to calculate the global cross-modal similarity. Compared with the global feature based methods,

* Zhong Ji and Haoran Wang are the corresponding authors.

Manuscript received xxxx; revised xxxx. This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 61771329 and 61632018, and the Natural Science Foundation of Tianjin under Grant 19JCYBJC16000.

Z. Ji, H. Wang, and Y. Pang are with the School of Electrical and Information Engineering, Tianjin University, Tinjin 300072, China (e-mails: jizhong@tju.edu.cn; haoranwang@tju.edu.cn; pyw@tju.edu.cn)

J. Han is with the Data Science group, University of Warwick, Coventry CV4 7AL, UK (e-mail: jungong.han@warwick.ac.uk)

this category of work seems more interpretable. However, it is suboptimal to measure similarity between all possible fragments from both modalities, because most pairs of cross-modal fragments describing different semantic concepts are actually meaningless to obtain the global similarity, while consuming huge amount of senseless computation resources. Besides, these methods usually rely on additional object detectors [60] or attribute predictor [27] to acquire more powerful visual representations, which requires expensive human annotations.

As a recent significant advance of deep learning, attention mechanism [4], [52] provides us with an effective solution to alleviate the above problems by selectively attending to the meaningful components of both modalities, *i.e.*, the informative image regions and textual words. Accordingly, some recent studies [44], [26], [34] devote to introducing attention mechanism to model cross-modal correlation. For instance, [44] and [22] employ the intra-modal information to guide the attention detection on the global features, which is capable of focusing on the semantically salient image regions and textual words. Despite achieving promising performance, it neglects the cross-modal semantic correlation included among images and texts. Intuitively, compared with intra-modal data, leveraging mutually complementary multi-modal information is more reasonable to enhance the discrimination of representations for both modalities and preserve the cross-modal semantic consistency. One straightforward solution is aggregating the local similarities of all pairs of visual and textual fragments according to the attention weights measured by the semantic relevance among them, which can be achieved by deploying the cross attention [34] or co-attention [35] mechanism. However, these approaches are implemented on the premise of explicitly calculating all image-text fragments, which requires a huge amount of computing resources and lack efficiency. Additionally, the incorrect matching between textual words and visual regions may be inappropriately introduced and accumulated during the attention learning procedure, which leads to the inefficiency of cross-modal correlation modeling.

To address the above problems, instead of computing the local similarities of all pairs of image regions and textual words exhaustively, we generate the semantic-aware context features of both modalities via aggregating multiple importance-weighted representations of visual and textual fragments respectively, achieved by resorting to a stacked attention mechanism [55]. As a departure from previous attention models, our approach allows multi-step attention reasoning to learn hierarchical representations for both modalities. Specifically, the entire stacked attention mechanism is carried out in two steps. First, we simultaneously encode the sequential and semantic information of both modalities to produce the intra-modal context feature, which acts as the query guidance of self-attention to discover the relations between image regions and textual words separately. Second, taking the semantic-aware visual and textual features output by the first step as refined intra-modal guidance, we specifically utilize a multi-modal fusion gate to integrate them into the multi-modal memories, following by leveraging them as query guidance to perform multi-modal attention reasoning. Consequently, we can acquire pairwise specific representations for any image-text group, which has capacity to concentrate on more meaningful parts for both modalities by leveraging the semantic complementarity between multi-modal data.

Moreover, due to existence of end-to-end training manner in deep learning, the distance metric in common space can actually be converted to the process of representation learning [56]. In the community of image-text retrieval, the most prevailing objective function is the bidirectional ranking loss [29], [44], [27], [9], which encourages the distance between the matched heterogeneous samples to be closer than those of the unmatched ones. Recently, there has been several bidirectional ranking loss variants that learn more efficient distance metric. For instance, [50] incorporates the intra-modal distance constraint into the original bidirectional ranking loss. Alternatively, [40] replaces the constraint of unmatched heterogeneous pair with that of homogeneous pair. For the form of objective function, [51] turns the max margin into a softmax-like form. However, most existing methods are designed to impose constraint on the distance between matched and unmatched instances, while neglecting the pairwise correspondence relationship between the matched ones. To this end, following [57] that takes every image/text group as one category, in contrast to the original bidirectional ranking loss, we develop an intra-pair ranking loss by additionally requiring the distances among intra-class instances to be shorter than a pre-defined margin, which further improves the discrimination of learned representations. Extensive experiments show remarkable improvements of our proposed intra-pair ranking loss over the original triplet bi-directional loss.

It is worthwhile to highlight several features of the proposed approach here:

- We introduce a stacked multi-modal attention module and incorporate it into a deep visual-semantic embedding framework, dubbed stacked multi-modal attention network (SMAN), as illustrated in Fig.2., In SMAN, the semantically meaningful visual fragments and textual words can be located simultaneously through a multi-step cross-modal association reasoning.
- We present a pairwise bi-directional constraint as the training objective of SMAN model for further preserving the original manifold structure of cross-modal data in the common space by enforcing the semantic consistency between pairwise images and texts.
- Our experimental results demonstrate that the performance of the proposed SMAN compares favorably with the state-of-the-art image-text retrieval methods on two public benchmark datasets: Flickr30k [45] and MSCOCO [36].

II. Related Work

A. Image-Text Retrieval

Recently, a rich line of studies have been proposed for cross-modal image-text retrieval. They roughly fall into two categories: 1) cross-modal similarity learning based approaches [50], [41], [26], [35] and 2) embedding space learning based approaches [49], [50], [57], [27], [9].

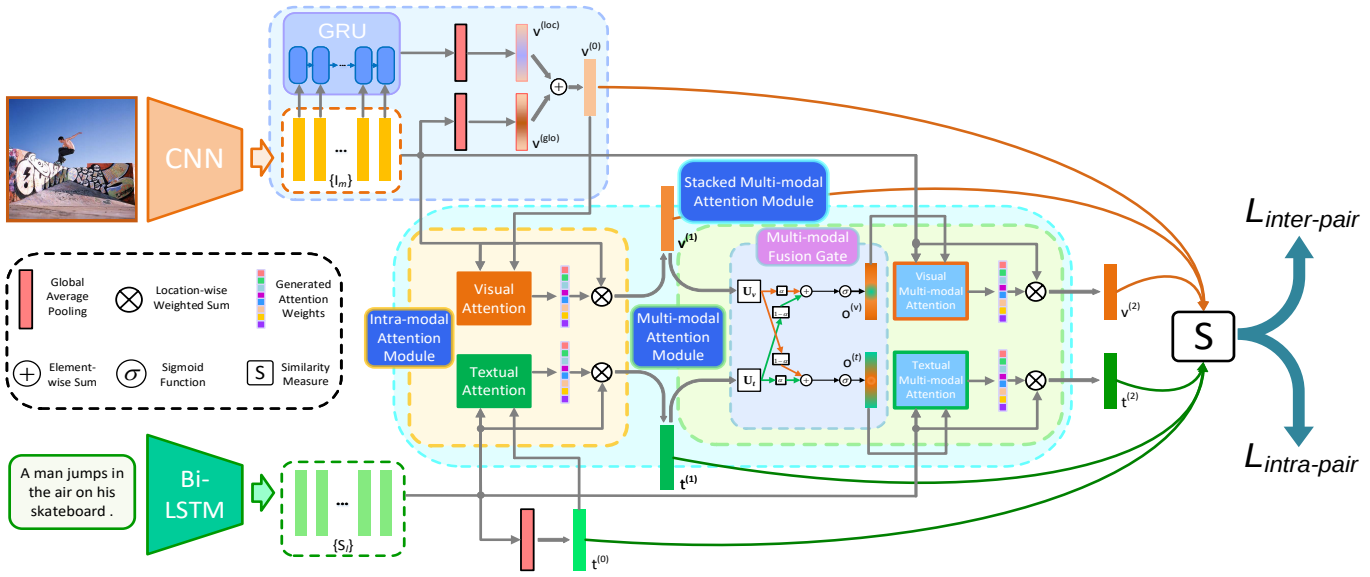


Fig. 2. The overview architecture of our proposed stacked multi-modal attention network (SMAN). The SMAN takes visual spatial features $\{\mathbf{I}_m\}$ and word-level textual features $\{\mathbf{S}_l\}$ as input from both modalities, respectively. Taking visual modality for instance (the same goes for textual modality): First, we simultaneously implement mean-pooling operation on $\{\mathbf{I}_m\}$ to generate the spatial semantic vector $\mathbf{v}^{(glo)}$ and employ GRU (Gated Recurrent Units) to produce the spatial location vector $\mathbf{v}^{(loc)}$, following by utilizing element-wise summation to obtain the global context feature $\mathbf{v}^{(0)}$. Next, the intra-modal visual attention module takes the global visual feature as guidance to re-weight the importance of visual spatial features, generating intra-modal visual context vector $\mathbf{v}^{(1)}$. Similarly, after processed by multi-modal fusion gate, we feed the intra-modal visual attention module with refined visual multi-modal memory vector $\mathbf{o}^{(v)}$ as guidance to produce the multi-modal visual context vector $\mathbf{v}^{(2)}$. Finally, we compute the global cross-modal similarity by aggregating the similarities of all multi-level context features.

Cross-modal similarity learning based approaches. Most cross-modal similarity learning based methods aim at learning a deep network for measuring the similarity of pairwise visual and textual input data. Ma *at el.* [41] proposed to employ one visual CNN to extract global image feature, and designed another matching CNN for turning the input visual and textual feature into a joint multi-modal feature. Wang *at el.* [65] utilized element-wise product to aggregate visual and textual data, followed by several fully connected layer. By employing a logistic regression loss as the training objective, it's able to predict whether the input image-text pair is matched.

In contrast, another stride of work aggregates similarities of fragments of images and texts. For example, Li *at el.* [35] proposed a latent co-attention mechanism, in which the spatial attention related each word with corresponding image regions while the latent semantic attention adopted LSTM to align different sentence structures. This strategy demands aggregating all visual-textual fragments for measuring the similarity score at the test stage, which results in huge burden of computing resource and time.

Embedding space learning based approaches. This line of approaches indicate embedding images and texts into a common space, in which the distances between them can be directly compared using conventional distance metrics, such as Euclidean distance or Cosine distance. One typical traditional solution is adopting the Canonical Correlation Analysis (CCA) [24] that aims to learn a latent common space by maximizing the correlation between both modalities [31]. With the renaissance of deep learning, the current mainstream for handling the task of cross-modal retrieval is resorting to deep architecture of neural network. For instance, Kiros *at el.* [30] employed convolutional neural networks (CNNs) to extract visual feature and recurrent neural networks (RNNs) to extract textual feature, learning joint embeddings with a triplet bi-directional ranking loss. Yan *at el.* [54] proposed to leverage deep canonical correlation analysis (DCCA) to learn one latent common space for matching images and texts. Wang *at el.* [50] combined intra-modal and inter-modal constraints to preserve the local structure for learning the common embedding. Zheng *at el.* [57] incorporated a instance loss into a two-branch CNN architecture, which utilizes the classification task to improve latent embeddings. Faghri *at el.* [6] exploited mining hard negatives in the triplet loss function to benefit the joint representation learning. Gu *at el.* [9] explored the use of generative objectives and incorporated it into the framework of cross-modal feature embedding learning.

Although these approaches have achieved encouraging progess for image-text retrieval, they ignored the fact that the global similarity is commonly based on aggregating local similarities between visual and textual fragments. As a results, they can not exploit the fine-grained interplay between both modalities to focus on capturing the shared semantics.

B. Deep Attention Mechanism

Attention mechanism, derived from recognition system of human, refers to concentrating on certain relevant elements of an input than irrelevant parts. They are first studied in the community of natural language processing (NLP), in which attention based encoder-decoder network is developed to benefit machine translation [4], [8], [48]. Later, the textual attention mechanisms are widely adopted in many relevant tasks, such

as sentence summarization [47], sentence embedding [37]. Motivated by their successful applications in the area of NLP, attention mechanisms have also been deployed in computer vision tasks such as image captioning [52], object detection [23], video summarization [28].

Recently, several attention-based methods have been leveraged for tackling the problem of image-text retrieval. Lee *at el.* [34] discovered all latent alignments between image regions and textual words by resorting to cross attention mechanism. Distinct from [34] that infers the global cross-modal similarity via directly aggregating those of pairwise visual and textual fragments, Huang *at el.* [26] proposed a context-modulated attention module that is capable of locating various pairwise instances appearing in data from both modalities through multi-step reasoning. Analogously, Nam *at el.* [44] enhanced discriminations of both visual and textual representations by developing the dual attention module that separately employs intra-modal context as guidance to focus on informative elements of each modality.

The main distinction between our proposed SMAN and the existing attention based models lies in that we exploit stacked attention mechanism to sequentially adopt the intra-modal context and multi-modal context as guidance to perform multi-step reasoning respectively. In previous work, the most relevant one to ours is [26] that performs multi-step attention reasoning. Different from [26] that performed attention on various instances (visual objects and textual words) from the same semantic level, our SMAN is able to leverage multi-level semantically complementary information to capture complex cross-modal associations. Besides, unlike the former only adopts the global CNN feature to extract visual information, our model can further encode the spatial location information of images, which helps to supply more effective guiding knowledge for multi-modal attention learning.

III. Stacked Multi-Modal Attention Network for Image-Text Retrieval

In this section, we will elaborate on our proposed stacked multi-modal attention model as the following three aspects: 1) representation learning for image and text, 2) stacked multi-modal attention module for associating image and text, 3) objective function for visual-textual embedding learning.

A. Representation Learning for Image and Text

1) Visual Representation: We employ ResNet-152 [25] pre-trained on ImageNet [32] as image feature encoder. We first resize images to 448×448 and then feed them into the CNN. For the sake of obtaining visual feature vectors of different regions, we take the feature map of ResNet152 (res5c) before the final average pooling layer as local features $\mathbf{I} \in \mathbb{R}^{7\times7\times2048}$. Then, according to the spatial position, we can represent an input image with a set of the visual descriptors, which are denoted by $\{\mathbf{I}_1, ..., \mathbf{I}_M\}$, where $M = 49$ denotes the number of image regions and the i-th image region is denoted by a 2048-dimensionality feature vector $\mathbf{I}_i$ $(i \in [1, M])$.

2) Textual Representation: Given one-hot encoding of L input words $\{\mathbf{w}_1, ..., \mathbf{w}_L\}$, we first embed the words into a vector space by $\mathbf{e}_l = \mathbf{P}\mathbf{w}_l$ where $\mathbf{P}$ is the embedding matrix. For the text encoder, we use bi-directional LSTMs as encoder to generate the textual features. Given a textual caption, we employ basic tokenizing to split it into words, and then sequentially feed them into a bi-directional LSTM:

$$\begin{aligned} \mathbf{h}_l^f &= LSTM^f(\mathbf{e}_l, \mathbf{h}_{l-1}^f), \\ \mathbf{h}_l^b &= LSTM^b(\mathbf{e}_l, \mathbf{h}_{l-1}^b), \end{aligned} \tag{1}$$

where $\mathbf{h}_l^f$ and $\mathbf{h}_l^b$ denotes the hidden states of forward and backward LSTMs at time l, respectively. Consequently, we obtain a series of feature vectors $\{\mathbf{S}_1, ..., \mathbf{S}_L\}$ by averaging the forward hidden state $\mathbf{h}_l^f$ and forward hidden state $\mathbf{h}_l^b$ at each time step, *i.e.*, $\mathbf{S}_l = \mathbf{h}_l^f + \mathbf{h}_b^f \quad (l \in [1, L])$, which summarizes semantic information of the l-th word in the context of the whole sentence.

B. Stacked Multi-Modal Attention Module

In this section, we introduce our stacked multi-modal attention module that performs multi-step attention reasoning for visual-semantic embedding. Concretely, the stacked multi-modal attention module contains a two-step reasoning. In the first step, we solely focus on leveraging the intra-modal information to improving the latent embeddings for both modalities, which is achieved by performing self-modal attentions on original visual and textual features simultaneously. Then, in the second step, based on employing the attentive visual and textual features guided by intra-modal information in the first step, we take advantage of the semantic complementarity between heterogeneous data to adopt inter-modal information to implement multi-modal attention on visual and textual features, which contributes to further mining more obtainable semantic association between images and texts. In the next section, we will elaborate on our proposed attention mechanisms with details, which serve as a significant module to compose our entire SMAN model. For clarity and simplicity, we divide the stacked multi-modal attention module into two separate parts and define them respectively as below: 1) **intra-modal attention** and 2) **multi-modal attention**. Note that, to avoid notational clutter, we omit the bias term b in the following exposition, but they actually exist in our model.

1) Intra-Modal Attention: **Visual Attention.** Just as its literal meaning,"attention" aims at telling which part of the input signal should be attended to. Assume we have a set of visual local features $\{\mathbf{I}_1, ..., \mathbf{I}_M\}$, the visual attention module is designed to independently exploit visual information to explore the fine-grained associating relations between multiple visual fragments, which is achieved by calculating the convex combination of the local features of image regions.

In this paper, we propose to exploit two types of visual information as query guidance of soft-attention mechanism [4], [52] to capture the interaction among the visual fragments. The first query item we employ is the visual global context feature, which has been validated to be effective for modeling the rich contextual relationships over local feature representations [44]. Specifically, given the local feature vectors $\{\mathbf{I}_1, ..., \mathbf{I}_M\}$

People standing outside an airplane with Native American art painted on the tail .
(a)

A group of people are watching a plane flying low .
(b)

A few onlookers are watching a plane fly through the sky .
(c)

Fig. 3. Illustration figures for introducing the motivation of encoding visual spatial location and semantic complementarity in the Stacked Multi-modal Attention (SMAN) module.

representing the visual features from all M regions of the input image, the visual global context vector $\mathbf{v^{(glo)}}$ is calculated by

$$\mathbf{v}^{(glo)} = \tanh(\mathbf{P}^{(0)} \frac{1}{M} \sum_{m=1}^{M} \mathbf{I}_m), \tag{2}$$

where $\mathbf{v}^{(glo)}$ is the global visual context vector encoding the global information from visual modality. $\mathbf{P^{(0)}}$ is a weight matrix aiming to embed visual feature vectors into the common space compatible with textual context vectors.

Except for the global information, we consider the spatial location information also plays critical role in enhancing the discrimination of visual representations. Take figures (a) and (b) in Fig. 3 for instance, the major objects of them are both "plane" and "people". It's really hard to discriminate them if only the global visual feature is employed, since the mean-pooling operation will lead to the loss of spatial location information. On the contrary, we can easily distinguish the two images according to the location of the "plane", which indicates that the spatial location information is an efficiently complementary for global information to enhance the visual discrimination ability. Therefore, we elect to encode the visual spatial location information to serve as the second query guidance of the visual attention module. Given the visual fragment features $\{\mathbf{I}_1, ..., \mathbf{I}_M\}$, we first organize them according to spatial order and then sequentially feed them into the GRU (Gated Recurrent Units), which is a special type of RNN that's characterized by computing efficiently and effectively, to model the spatial location information among them. Specifically, the above procedure can be defined as:

$$\mathbf{h}_m^{loc} = GRU(\mathbf{I}_m, \mathbf{h}_{m-1}^{loc}), \tag{3}$$

where $\mathbf{h}_m^{loc}$ denotes the hidden state of GRU at timestep m. Consequently, we can obtain a set of hidden state vectors $\{\mathbf{h}_1^{loc}, ..., \mathbf{h}_M^{loc}\}$ and perform mean-pooling on them to acquire the visual location feature $\mathbf{v}^{(loc)} = \frac{\sum_{m=1}^{M} \mathbf{h}_m^{loc}}{M}$, which summarizes the spatial location information in the context of the whole image.

Next, after the visual global context vector feature $\mathbf{v}^{(glo)}$ and the visual spatial location vector feature $\mathbf{v}^{(loc)}$ are both acquired, we combine them through a element-wise summation to get the fused query guidance of visual modality. Specifically, the visual query guidance vector $\mathbf{v}^{(0)}$ can be computed as:

$$\mathbf{v}^{(0)} = \mathbf{v}^{(glo)} \oplus \mathbf{v}^{(loc)}, \tag{4}$$

After that, the attention weight $\alpha_{v,m}^{(1)}$ of each image region is calculated by feeding the visual feature vectors $\{\mathbf{I}_1, ..., \mathbf{I}_M\}$ and the visual query guidance vector $\mathbf{v^{(0)}}$ into the attention function $f_{att}(\cdot, \cdot)$, which employs a 2-layer feed-forward perceptron (FFP) following by softmax function to ensure all the computed weights sum up to 1:

$$\mathbf{h}_{v,m}^{(1)} = \tanh(\mathbf{W}_v^{(1)} \mathbf{I}_m \odot \tanh(\mathbf{W}_{v,q}^{(1)} \mathbf{v}^{(0)}), \tag{5}$$

$$\mathbf{a}_{v,m}{}^{(1)} = \mathrm{softmax}(\mathbf{W}_{v,h}^{(1)} \mathbf{h}_{v,m}^{(1)}), \tag{6}$$

where $\mathbf{W}_v^{(1)}$, $\mathbf{W}_{v,q}^{(1)}$ and $\mathbf{W}_{v,h}^{(1)}$ are the parameters of perceptron, $\mathbf{h}_{v,m}^{(1)}$ represents the hidden state of visual attention function, and $\odot$ represents element-wise multiplication. Once the weights of different regions of image are computed, the intra-modal visual context vector is calculated by:

$$\mathbf{v}^{(1)} = \tanh(\mathbf{P^{(1)}} \frac{1}{M} \sum_{m=1}^{M} \alpha_{v,m}^{(1)} \mathbf{I}_m), \tag{7}$$

where $\mathbf{P^{(1)}}$ is a weight matrix aiming to embed visual feature vectors into the common space compatible with textual context vectors, so as to measure the similarity between visual representation and textual representation.

Textual Attention. Similar to the visual attention, textual attention is designed to generate an attentive textual context vector by focusing on certain specific words in the input sentence. The main distinction between textual attention and the former is that we just use mean-pooling to produce the global textual context vector as the input query guidance of attention module. Because there has been sequential information included in the fragments of textual words before we aggregate them. Concretely, given the textual feature vectors $\{\mathbf{S}_1, ..., \mathbf{S}_L\}$ representing the textual features from words in the sentence, we fuse them to form the global textual representation:

$$\mathbf{t}^{(0)} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{S}_l, \tag{8}$$

where $\mathbf{t^{(0)}}$ is a vector encoding the global information from textual modal, we name it global textual context vector. On the whole, the textual attention mechanism is very similar to our aforementioned visual attention mechanism. Specifically, the attention weights $\alpha_{t,j}^{(0)}$ are obtained from soft attention module consisting of 2-layer FFP and softmax function. And then the intra-modal textual context vector $\mathbf{t}^{(1)}$ is calculated by:

$$\mathbf{h}_{t,1}^{(1)} = \tanh(\mathbf{W}_t^{(1)} \mathbf{S}_l) \odot \tanh(\mathbf{W}_{t,q}^{(1)} \mathbf{t}^{(0)}), \tag{9}$$

$$\mathbf{a}_{t,l}^{(1)} = \mathrm{softmax}(\mathbf{W}_{t,h}^{(1)} \mathbf{h}_{t,l}^{(1)}), \tag{10}$$

$$\mathbf{t}^{(1)} = \sum\nolimits_{l=1}^{L} \mathbf{a}_{t,l}^{(1)} \cdot \mathbf{S}_l, \tag{11}$$

where $\mathbf{W}_t^{(1)}$, $\mathbf{W}_{t,q}^{(1)}$ and $\mathbf{W}_{t,h}^{(1)}$ are the parameters of FFP, $\mathbf{h}_{t,l}^{(1)}$ represents the hidden state of textual attention. Compared with visual attention, there is no need to add an embedding layer after the weighted averaging, as the textual features $\{\mathbf{S}_1, ..., \mathbf{S}_L\}$ already exist in the common space and trained by end-to-end manner.

2) Multi-Modal Attention: Compared with traditional models that directly measure the similarity between global visual features and textual features in the common space, the attentive context features output by the intra-modal attention module have the capacity to pinpoint which regions of image and words of sentence are more informative and discriminative to measure the relevance between image and sentence. However, although the intra-modal information contributes to indicating the meaningful parts of both modalities, it is not adept at capturing the cross-modal semantic correlations between heterogeneous data. For example, as depicted in Fig. 3, the textual descriptions of both Figures (b) and (c) are very close semantically. If we only rely on the intra-modal information to perform attention on textual features, it's really hard to distinguish them when measuring the cross-modal similarity, meanwhile the same problem goes for visual modality.

Our solution to alleviate the above issue is resorting to the semantic complementarity existing between heterogeneous data to enhance the representation ability for images and texts. Unlike most existing methods perform attention reasoning guided by only single-modal information [44], [37] to generate the unique representation for one visual or textual instance, our method is capable of producing pairwise representations for each image-text pairs according to their semantic association. Concretely, it indicates that, for measuring the cross-modal similarity, the representations for any instance are sensitive to its corresponding heterogeneous data. Doing so allows the representation from certain modality to benefit from each other via simultaneously leveraging the fine-grained visual clues [53] and abstract high-level textual semantics [51].

Specifically, to extract more meaningful and critical information for measuring the similarity between image and text, we resort to the stacked attention [55] mechanism to process the context features $\mathbf{v}^{(1)}$ and $\mathbf{t}^{(1)}$ generated by the intra-modal attention module. Obviously, due to they both encode the sequential and semantic information from each modality, the intra-modal context features are more discriminative for measuring cross-modal similarity than the original global features. Furthermore, the intra-modal context features can be seen as the product that the original global features are refined through the intra-modal attention module. This refinement process is, to some extend, similar to the memory mode of human which rules out redundant information and maintains the crucial memory fragments. Thus, they can be regarded as vectors containing memory of each modality. For simplicity, here we call them visual memory vector and textual memory vector, respectively.

Then, we need to integrate them into the semantically complementary multi-modal query guidance for performing the next step attention reasoning. An intuitive way to combine visual memory and textual memory is to sum the former with the latter together directly. However, we found this operation will result in relatively poor performance in practice. We assume that the potential reason is that the discriminations that visual memory and textual memory could contribute are not equivalent in most cases during training stage. Thus, summing them together directly may cause some effective information from one modality to be covered up by that from another modality. To circumvent this problem, we design a multi-modal fusion gate that can selectively balance the relative importance of visual memory and textual memory. As illustrated in Fig.2, after obtaining the visual memory vector and textual memory vector , we feed them into the fusion gate and the procedure can be formulated as:

$$\begin{aligned}
\hat{\mathbf{v}} &= \mathbf{U}_v(\mathbf{v}^{(0)} + \mathbf{v}^{(1)}), \\
\hat{\mathbf{t}} &= \mathbf{U}_t(\mathbf{t}^{(0)} + \mathbf{t}^{(1)}), \\
\mathbf{o}^{(v)} &= \sigma(\alpha\,\hat{\mathbf{v}} + (1-\alpha)\,\hat{\mathbf{t}}), \\
\mathbf{o}^{(t)} &= \sigma(\alpha\,\hat{\mathbf{t}} + (1-\alpha)\,\hat{\mathbf{v}}),
\end{aligned} \tag{12}$$

where $\mathbf{U}_v$ and $\mathbf{U}_t$ represent linear embedding matrix, α is the parameter controlling how much information of visual memory and textual memory contributes to their fused representation. The use of sigmoid function σ is to rescale each element in the fused representation to [0, 1]. $\mathbf{o}^{(v)}$ and $\mathbf{o}^{(t)}$ represent the refined multi-modal memory vectors output by the gated fusion unit.

After obtaining $\mathbf{o}^{(v)}$ and $\mathbf{o}^{(t)}$, we feed them as guiding information into the multi-modal visual attention and textual attention modules, producing the refined visual context vector $\mathbf{v}^{(2)}$ and textual context vector $\mathbf{t}^{(2)}$, dubbed multi-modal visual textual context vector, respectively. Similar to the procedure of producing intra-modal context vectors, the multi-modal visual textual context vector $\mathbf{v}^{(2)}$ and $\mathbf{t}^{(2)}$ are defined as follows, respectively:

$$\mathbf{h}_{v,m}^{(2)} = \tanh(\mathbf{W}_v^{(2)}\mathbf{I}_m \odot \tanh(\mathbf{W}_{v,q}^{(2)}\mathbf{o}^{(v)}), \tag{13}$$

$$\mathbf{a}_{v,m}{}^{(2)} = \mathrm{softmax}(\mathbf{W}_{v,h}^{(2)}\mathbf{h}_{v,m}^{(2)}), \tag{14}$$

$$\mathbf{v}^{(2)} = \tanh(\mathbf{P}^{(\mathbf{2})}\frac{1}{M}\sum_{m=1}^{M}\alpha_{v,m}^{(2)}\mathbf{I}_m), \tag{15}$$

$$\mathbf{h}_{t,l}^{(2)} = \tanh(\mathbf{W}_t^{(2)}\mathbf{S}_l) \odot \tanh(\mathbf{W}_{t,q}^{(2)}\mathbf{o}^{(t)}), \tag{16}$$

$$\mathbf{a}_{t,l}^{(2)} = \mathrm{softmax}(\mathbf{W}_{t,h}^{(2)}\mathbf{h}_{t,l}^{(2)}), \tag{17}$$

$$\mathbf{t}^{(2)} = \sum\nolimits_{l=1}^{L}\mathbf{a}_{t,l}^{(2)} \cdot \mathbf{S}_l, \tag{18}$$

where $\mathbf{W}_v^{(2)}$, $\mathbf{W}_{v,q}^{(2)}$, $\mathbf{W}_{v,h}^{(2)}$, $\mathbf{W}_t^{(2)}$, $\mathbf{W}_{t,q}^{(2)}$ and $\mathbf{W}_{t,h}^{(2)}$ are the parameters of perceptron. And $\mathbf{h}_{v,m}^{(2)}$ and $\mathbf{h}_{t,l}^{(2)}$ represents the hidden state of visual and textual multi-modal attention, respectively. Although the multi-modal context vectors $\mathbf{v}^{(2)}$ ($\mathbf{t}^{(2)}$) are formally similar to the intra-modal memory vectors $\mathbf{v}^{(1)}$ ($\mathbf{t}^{(1)}$), it should be mentioned that the later represents

each visual or textual instance with unique feature, whereas the former generate varied representations for any cross-modal data pair.

3) Remarks: Considering some recent works for cross-modal retrieval are also built based on attention mechanism, in this section, we make some comparison between our model and them for clear understanding this community. For modelling the intra-modal attention, our SMAN module is similar to Dual Attention Network (DAN) [44] and Self-Attention Embedding (SAE) [21] but with a distinct differences: Both DAN and SAE only focus on employing the global visual feature as query guidance for attention modeling, hence taking no spatial location information of images into consideration. In contrast, our module simultaneously encodes both the global semantic and sequential order information for both modalities to perform the attention learning. Moreover, to preserve the cross-modal semantic consistency, our model is conceptually associated with Stacked Cross Attention Networks (SCAN) [16] and Cross-media Relation Attention Network (CRAN) [17]. However, they differ significantly in design: 1) The CRAN performs multi-level attention reasoning separately, considering less on the associations among various semantic levels. By comparison, the intra-modal context features multi-modal context features of our SMAN module can benefit from each other by exploiting the stacked attention architecture. 2) Both SCAN and our model adopt the stacked attention mechanism. However, the former just allows for utilizing the raw local single-modal information to as guidance to perform attention on cross-modal data. In contrast, our method employs attentive multi-modal query guidance to simultaneously generate attention weights for visual and textual local features, which incorporates more semantically complementary information into the joint representation learning.

C. Objective Function for Cross-modal Matching

During training stage, we obtain the multi-level context vectors output by our SMAN model and combine them as visual context vectors $\mathbf{v}^{(k)}(k = 0, 1, 2)$ and textual context vectors $\mathbf{t}^{(k)}(k = 0, 1, 2)$, respectively. The similarity $s^{(k)}$ between visual and textual context vectors is obtained by computing their Cosine distance:

$$d^{(k)} = \frac{\mathbf{v}^{(k)} \cdot \mathbf{t}^{(k)}}{\|\mathbf{v}^{(k)}\|\|\mathbf{t}^{(k)}\|} \tag{19}$$

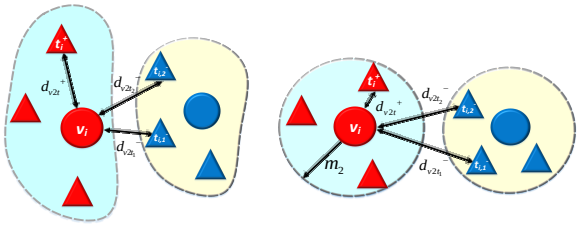
The final similarity s between a given image and sentence is calculated by

$$d = \sum_{k=0}^{2} d^{(k)} \tag{20}$$

1) Improve Bidirectional Triplet Ranking Loss with intra-pair constraint: The bidirectional triplet ranking loss is a widely adopted ranking objective for image-sentence retrieval [29], [50]. For each matched pair of an image and a sentence $(\mathbf{v}^+, \mathbf{t})$ and $(\mathbf{v}, \mathbf{t}^+)$, we select a negative image and a negative sentence to mismatched pair $(\mathbf{v}^-, \mathbf{t})$ and $(\mathbf{v}, \mathbf{t}^-)$. Then, the hinge-based triplet ranking loss can be written:

$$\begin{aligned}\mathcal{L}_{inter-pair} = \sum_{(\mathbf{v},\mathbf{t})} \{&\max[0, m_1 - d(\mathbf{v}, \mathbf{t}^+) + d(\mathbf{v}^-, \mathbf{t})] \\ &+ \max[0, m_1 - d(\mathbf{v}^+, \mathbf{t}) + d(\mathbf{v}, \mathbf{t}^-)]\},\end{aligned} \tag{21}$$

where m_1 is a margin parameter.



(a). Without pairwise distance constraint (b). With pairwise distance constraint

Fig. 4. Illustration of how pairwise distance constraint affects the procedure of learning the joint embedding space. Circles denote visual instances and rectangles denote textual instances. Same color indicates images and sentences are from the same image/sentence pair. The intra-pair constraint has capability to ensure the distances among pairwise visual and textual instances to be close.

Taking a sample from arbitrary modality as an anchor sample, we call the distance difference in the common space between its matched sample and mismatched sample as relative distance. Similarly, we call the distance between the anchor and its matched sample as absolute distance. We can observe that, by optimizing loss function defined in Eq. (21), our SMAN model is trained to focus on the common semantics that only appears in correct image-sentence pairs through the distance constraint with matched pair and mismatched pair (*e.g.* relative distance). However, since the number of negative sample pairs we can obtain via sampling randomly per epoch is much less than those of all exhaustive negative sample pairs, only imposing constraint on relative distance rather than absolute distance can not guarantee the distances among the samples from positive pair are close enough, which may lead to failed retrievals. As illustrated in Fig.4(a), we take visual instance $\mathbf{v}_i$ as the anchor sample. If its corresponding negative sample $\mathbf{t}_{i,1}^-$ does not participate in the training procedure as the negative sample of $\mathbf{v}_i$ for enough iterations, as a consequence, the relative distance between $\mathbf{v}_i$ and $\mathbf{t}_{i,1}^-$ (*e.g.* $d_{v2t_1}{}^-$) is likely to be smaller than the absolute distance between $\mathbf{v}_i$ and $\mathbf{t}_i^+$ (*e.g.* $d_{v2t}{}^+$). Intuitively, the most simple solution is to list all possible combinations of positive sample pairs and negative sample pairs exhaustively. Nonetheless, it is actually not feasible due to heavy computation burden excessively.

Based on above observations, we propose to incorporate a constraint term into the original bidirectional triplet ranking loss, which further requires the absolute distance between pairwise positive samples to be less than the margin m_2 (shown in Fig.4(b)), and it is noted that m_2 should be much smaller than m_1. Specifically, we formulate this statement as:

$$\begin{aligned}\mathcal{L}_{intra-pairk} = \sum_{(\mathbf{v},\mathbf{t})} \{&\max[0, d(\mathbf{v}, \mathbf{t}^+) - m_2] \\ &+ \max[0, d(\mathbf{t}, \mathbf{v}^+) - m_2]\end{aligned} \tag{22}$$

2) Final Objective Function and Inference Method: In summary, we combined the original bidirectional triplet ranking loss $\mathcal{L}_{inter-pair-rank}$ and our proposed intra-pair constraint term $\mathcal{L}_{intra-pair-rank}$ together as final objective function:

$$L = L_{inter-pair} + \lambda L_{intra-pair} \tag{23}$$

where λ is a tuning parameter to balance two terms of loss.

At inference time, an arbitrary image or sentence is embedded into the common space by and represented by concatenating its three-level context vectors, *i.e.* $\mathbf{r}_v = [\mathbf{v}^{(\mathbf{0})}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}]$ and $\mathbf{r}_t = [\mathbf{t}^{(0)}, \mathbf{t}^{(1)}, \mathbf{t}^{(2)}]$. Similar to Eq. (20), the relevance between them is directly measured by their Cosine distance.

IV. Experiments

In this section, we first introduce the two benchmark datasets, *i.e.*, Flickr30k and MSCOCO, following by the experimental setting in Section IV-A. Then the corresponding results on two datasets are given in Section IV-B and Section IV-C, respectively. In Section IV-D, we perform a series of ablation studies to testify how the different modules affect the performance of our model. Additionally, More experimental results are provided in Appendix, including the effect of intra-pair constraint of bidirectional ranking loss, the analysis on computational cost of our model, and the qualitative results.

A. Dataset and Setting

We evaluate our proposed SMAN model on the Flickr30K dataset [45] and MSCOCO dataset [36] for image-text retrieval.

1) Datasets:

- **Flickr30k** [45] is an image-captioning dataset consisting of 31783 images collected from the Flickr website, in which each image is annotated with five caption sentences. Following the protocol provided by [42], we split the dataset into 29783 training images, 1000 validation images and 1000 test images. The performance evaluation of bi-directional image and text retrieval is reported on 1000 test set.
- **MSCOCO** [36] is another image-captioning dataset, including 123,287 images and each image is also roughly annotated with five textual descriptions. We adopt the public dataset split proposed by [29] for MSCOCO. It contains 82,783 images for training, 5,000 images for validation and 5,000 images for testing [29]. Besides, We follow [6] to use 30,504 images that were originally in the validation set of MSCOCO for training. The experimental results are reported by either averaging over 5 folds of 1K test set.

2) Evaluation Metric: We use the widely-used R@K as evaluation metric [49], [6], defined as the possibility of at least one of ground-truth matchings appears in the top *K*-ranked retrieval results. Another metric we utilize is Med r, that denotes the median rank of the closest retrieved ground-truth sample, with a lower value be better. Besides, we follow [27] to compute an additional criterion "mR", averaging all six recall rates of R@K, which is suitable to evaluate the overall performance for cross-modal retrieval.

3) Implementation Details: All our experiments are implemented in TensorFlow toolkit with a single NVIDIA GEFORCE GTX TITAN Xp GPU. The dimension of every hidden embedding layer including word embedding, GRU units, Bi-LSTM units, attention modules and common embedding space are all set to 512. To make fair performance comparison with existing approaches, we employ random initialization to perform word embedding [6], [34], [9]. In addition, we also employ Glove [61] pre-trained on the Common-Crawl dataset to perform the word-embeddings for representing textual words. We will report their results, respectively. We train our networks by Adadelta with a learning rate 0.05, momentum 0.9, weight decay 0.0005, dropout rate 0.3, and gradient clipping at 0.1. The network is trained for 60 epochs totally, where the learning rate is dropped to 0.005 after 50 epochs. The parameter α in Eq. (12) is set to 0.9 in our experiment. The size of mini-batch is set to 128 and we follow [6] to concentrate on the hardest negatives in a mini-batch. Empirically, we set the margin parameter $m_1 = 0.2$ in Eq. (21) and $m_2 = 0.25$ in Eq. (22), respectively. The balancing parameter λ in Eq. (23) is set to 1.

4) Baseline Approaches for Comparisons:

- **SMAN-IM:** This baseline model contains only intra-modal attention part of the stacked multi-modal attention module without employing the visual location information, accompanied with both $\mathcal{L}_{inter-pair-rank}$ and intra-pair constraint term $\mathcal{L}_{intra-pair-rank}$ together.
- **SMAN-IM-LOC:** This baseline model consists of only intra-modal attention part of the stacked multi-modal attention module, accompanied with both $\mathcal{L}_{inter-pair-rank}$ and our proposed intra-pair constraint term $\mathcal{L}_{intra-pair-rank}$ together.
- **SMAN-SIM:** This baseline model consists of stacked intra-modal attention that is actually equivalent to the stacked multi-modal attention module (controlling parameter $\alpha = 1$), accompanied with both $\mathcal{L}_{inter-pair-rank}$ and our proposed intra-pair constraint term $\mathcal{L}_{intra-pair-rank}$ together.
- **SMAN-inter:** This baseline model contains stacked multi-modal attention module with only bidirectional triplet ranking loss $\mathcal{L}_{inter-pair-rank}$.
- **SMAN (random):** This is our proposed full SMAN model containing both our proposed multi-modal attention module and intra-pair constraint term $\mathcal{L}_{intra-pair-rank}$. The textual word embedding is achieved by random initialization.
- **SMAN (Glove):** This model is equivalent to SMAN except for replacing random initialization (Xavier initialization) with Glove technique to perform the word embedding.

B. Result on Flickr30k

We compare our proposed SMAN model with several state-of-the-art methods for the cross-modal image-text retrieval task. Table I lists the quantitative experimental results on Flickr30K. We can see that our proposed approach outperforms other approaches in all seven evaluation criterions, which clearly demonstrates the advantages of our approach.

TABLE I
COMPARISONS OF EXPERIMENTAL RESULTS ON FLICKR30K TESTING SET. α IS THE CONTROLLING PARAMETER IN EQ. (12). WE ADOPT RESNET-152 AND BI-LSTM TO REPRESENT THE VISUAL AND TEXTUAL MODALITIES, RESPECTUVELY.

Methods	Sentence Retrieval				Image Retrieval				mR
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
DVSA [29] (RCNN (AlexNet))	22.2	48.2	61.4	5	15.2	37.7	50.5	9	39.2
m-CNN [41] (VGG-19)	33.6	64.1	74.9	3	26.2	56.3	69.6	4	54.1
DSPE [50] (VGG-19)	40.3	68.9	79.9	-	29.7	60.1	72.1	-	58.5
2WayNet [2] (VGG-19)	49.8	67.5	-	-	36.0	55.6	-	-	-
DAN [44] (ResNet-152)	55.0	81.8	89.0	1	39.4	69.2	79.1	2	68.9
CRAN [17] (VGG-19)	38.1	70.8	82.8	-	38.1	71.1	82.6	-	63.9
CMPM [56] (ResNet-152)	49.6	76.8	86.1	-	37.3	65.7	75.5	-	65.2
VSE++ [6] (ResNet-152)	52.9	79.1	87.2	1	39.6	69.6	79.5	2	68.0
DPC [57] (ResNet-50)	55.6	81.9	89.5	1	39.1	69.2	80.9	2	69.4
SCO [27] (ResNet-152)	55.5	82.0	89.3	-	41.1	70.5	80.1	-	69.7
SMAN-inter ($\alpha = 1$)	54.9	81.9	89.3	**1**	40.2	70.6	80.1	**2**	69.5
SMAN-inter ($\alpha = 0.9$)	56.2	83.5	90.6	**1**	41.9	71.5	82.0	**2**	71.0
SMAN	56.9	84.8	91.9	**1**	43.2	73.3	**83.5**	**2**	**72.3**
SMAN (Glove)	**57.3**	**85.3**	**92.2**	**1**	**43.4**	**73.7**	83.4	**2**	**72.6**

TABLE II
COMPARISONS OF EXPERIMENTAL RESULTS ON MSCOCO 1K TESTING SET. α IS THE CONTROLLING PARAMETER IN EQ. (12). WE ADOPT RESNET-152 AND BI-LSTM TO REPRESENT THE VISUAL AND TEXTUAL MODALITIES, RESPECTUVELY.

Methods	Sentence Retrieval				Image Retrieval				mR
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
1K Testing set									
DVSA [29] (RCNN (AlexNet))	38.4	69.9	80.5	1	27.4	60.2	74.8	3	58.5
m-CNN [41] (VGG-19)	42.8	73.1	84.1	2	32.6	68.6	82.8	3	64
DSPE [50] (VGG-19)	50.1	79.7	89.2	-	39.6	75.2	86.9	-	70.1
CRAN [17] (VGG-19)	23.0	52.0	66.0	-	21.1	48.9	64.5	-	45.9
CMPM [56] (ResNet-152)	56.1	86.3	92.9	-	44.6	78.8	89	-	74.6
NAA [38] (ResNet-152)	61.3	87.9	95.4	-	47.0	80.8	90.1	-	-
VSE++ [6] (ResNet-152)	64.7	-	95.9	1	52.0	-	92.0	2	-
DPC [57] (ResNet-50)	65.6	89.8	95.5	1	47.1	79.9	90.0	2	78.0
VSEPP [43] (ResNet-152)	61.5	-	96.1	1	46.3	-	89.4	2	-
CSE [58] (ResNet-152)	56.3	84.4	92.2	1	45.7	81.2	90.6	2	-
AITE [39] (ResNet-152)	66.4	90.6	96.4	1	53.3	84.7	92.4	1	80.6
SMAN-inter ($\alpha = 1$)	66.1	89.0	94.1	**1**	55.7	85.3	90.2	**1**	80.1
SMAN-inter ($\alpha = 0.9$)	67.3	89.9	94.6	**1**	57.8	86.5	92.4	**1**	81.4
SMAN (Random)	67.9	90.6	96.2	**1**	**58.8**	87.0	**93.7**	**1**	82.4
SMAN (Glove)	**68.4**	**91.3**	**96.6**	**1**	58.5	**87.4**	93.5	**1**	**82.6**

As for our best results, the R@1 of sentence retrieval given an image query is 57.3%, achieved by our full SMAN model, where achieves 3.2% improvement comparing to the suboptimal approach, DPC [57]. Meanwhile, our SMAN arrives at 43.4% for the R@1 of image retrieval, which achieves 5.6% improvement than the suboptimal approach, SCO [27]. Furthermore, it is worth noting that when we employ the SMAN-inter ($\alpha = 1$) model and the visual location feature $v^{(loc)}$ is excluded in intra-modal visual attention, it is approximately equivalent to DAN [44] whose attention layer is equal to 2. In this situation, [44] can be essentially seen as a special case of our SMAN model.

To observe the effect of our proposed intra-pair ranking loss, we further investigate in two different configurations with SMAN-inter (α=0.9) as our baseline model. When we incorporate intra-pair constraint into the original bidirectional ranking loss, compared with the baseline model, we see 1.2% and 3.1% improvement in R@1 accuracy for sentence retrieval and image retrieval, respectively.

C. Result on MSCOCO

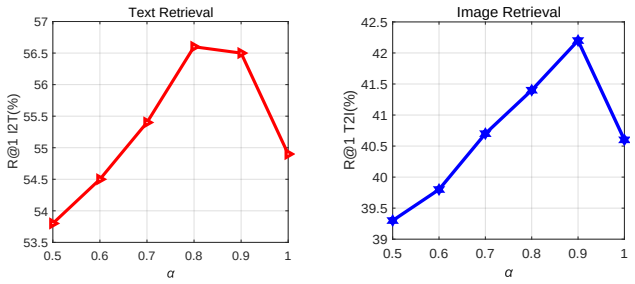
We compare our proposed SMAN with several state-of-the-art methods on the MSCOCO dataset, and present the results in Table II. From Table II, we can observe that for 1K test dataset our model SMAN-inter achieves R@1=67.3% and 57.8% with image and text as quires, respectively, which is comparable to the second best performance of AITE [39]. It demonstrates that our proposed stacked multi-modal attention module is capable of boosting the performance of our model, bringing about 1.8% and 3.8% performance gain. When adopting our proposed intra-pair constraint term, our full SMAN model outperforms the second best approach, achieving 68.4% of R@1 for text retrieval and 58.5% for image retrieval. These experimental results considerably validate the superiority of our SMAN on learning effective visual-semantic embedding, especially for matching image and text.

TABLE III
IMPACT OF VARIOUS ATTENTION MODULES ON FLICKR30K DATASET

Methods	text retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
SMAN-IM	55.5	82.9	89.2	39.8	71.2	80.7
SMAN-IM-LOC	55.9	82.6	89.1	40.3	71.4	80.9
SMAN-SIM	56.5	84.3	81.4	42.3	72.8	82.7
SMAN	**57.3**	**85.3**	**92.2**	**43.4**	**73.7**	**83.4**

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT CONTROLLING PARAMETER α ON FLICKR30K DATASET

Methods	text retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
SMAN-inter ($\alpha = 0.5$)	53.8	81.2	88.2	39.3	68.9	79.7
SMAN-inter ($\alpha = 0.6$)	54.5	82.1	89.0	39.8	69.9	80.6
SMAN-inter ($\alpha = 0.7$)	55.4	83.2	90.2	40.7	71.1	80.6
SMAN-inter ($\alpha = 0.8$)	**56.6**	83.9	90.8	41.4	**72.7**	81.8
SMAN-inter ($\alpha = 0.9$)	56.5	**84.3**	**91.5**	**42.2**	72.5	**82.5**
SMAN-inter ($\alpha = 1.0$)	54.9	83.0	89.4	40.6	71.1	80.7



(a) Text retrieval (b) Image retrieval

Fig. 5. Impact of varied controlling parameter α on Flickr30K dataset. The sub-figure (a) and (b) denote the image-to-text retrieval result (I2T) and text-to-image retrieval result (T2I), respectively.

D. Ablation Study

In this section, we present several quantitative results of ablation models and analyze the effect of each component in our model.

1) Effect of various attention components: To validate the effectiveness of our proposed stacked multi-modal attention mechanism for image-text retrieval, we implement another two attention schemes for comparison. These two attention schemes are utilized in DAN [44], which only employ intra-modal information for performing attention, and the only difference between them is how many steps their memories update for. As mentioned above, we name the modified attention module with single-layer intra-modal attention and stacked multi-modal attention mechanisms as SMAN-IM and SMAN-SIM, resprectively.

From Table III, it can be observed that, compared with the first three ablation models employing intra-modal attention mechanisms, our proposed stacked multi-modal attention achieves better performance on all seven metrics. It is worth noting that, in the case of employing stacked attention mechanism equally, our proposed attention outperforms the stacked intra-modal attention by 1.4% and 2.6% in terms of text retrieval R@1 and image retrieval R@1 accuracy, respectively. These experimental results verify the our original conjecture that multi-modal memory is able to provide more powerful guiding information for image-text retrieval than intra-modal one. Additionally, we can also see that the SMAN-IM-LOC obviously outperforms the SMAN-IM model. It indicates that our intra-modal attention module really benefits from the introducing of visual spatial location information.

2) Effect of cross-modal information fusion ratio in stacked multi-modal attention module: We further investigate in the influence of parameter α in Eq. (12), which controls the fusion ratio of cross-modal information contained in stacked multi-modal attention module, and present the results in Table IV and Fig.5. We can see that the performances improve continuously with the increase of α until the performances reach their peak when $\alpha = 0.9$, and then decreases when $\alpha = 1$, *i.e.*, the multi-modal attention module degrades as the intra-modal attention module. Through the observation, we perceive that, incorporating appropriate ratio of cross-modal information into the memory can steadily improve the performance of cross-modal retrieval compared to intra-modal memory.

However, it should be noted that the performance is not always directly proportional to the ratio of cross-modal information in the memory. In contrary, when we vary α from 0.9 to 0.5, the performance degrades continuously accompanied by more cross-modal information incorporated into the multi-modal memory. In particular, when visual memory and textual memory contribute equally to constitute the multi-modal memory ($\alpha = 0.5$), the multi-modal attention module perform slightly worse than the intra-modal attention module. We assume the potential reason is that, compared with fusing cross-modal information into memory, the intra-modal attention module can provide better generalization ability than the former to overcome the relative paucity of training data. Therefore, considering the limited size of the Flickr30k training set, incorporating higher ratio of intra-modal information into the multi-modal memory can achieve better performance.

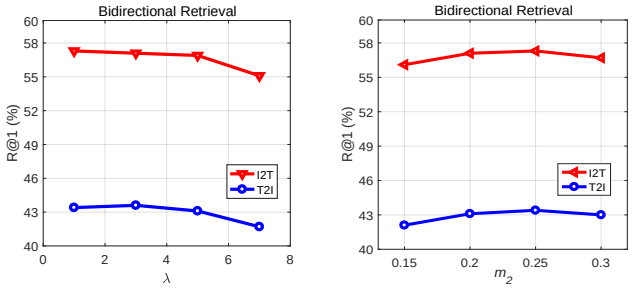
3) Effect of intra-pair constraint of Bidirectional Triplet Ranking Loss: We implement ablation study to further evaluate the influences of balancing parameter λ in Eq. (23) and intra-pair constraint margin m_2 in Eq. (22) to the maturity of the learned model. The corresponding experimental results are presented in Table V. It can be observed that when we fix the margin parameter at default value ($m_2 = 0.25$) and vary the balancing parameter λ from 1 to 5, it can seen that the performance of our model is not obviously affected by weight of the intra-pair constraint term. It demonstrates that our model is relatively not sensitive to the variation of the balancing parameter λ.

V. CONCLUSION

In this paper, we propose a novel stacked multi-modal attention based approach for learning visual-textual embeddings for image-text retrieval. Concretely, we resort to stacked attention mechanism that leverages intra-modal and multi-modal memories to perform attention on both modalities in phases, generating multi-level representations that are capable of capturing the fine-grained semantic correlations between two modalities. Besides, we propose a pairwise bi-directional loss that ensures the distance between the pairwise samples

TABLE V
THE INFLUENCE OF PARAMETERS CONCERNING INTRA-PAIR CONSTRAINT TERM ON FLICKR30K DATASET

Methods	text retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
SMAN-inter	56.5	84.3	91.5	42.2	72.5	82.5
SMAN (λ=1)	**57.3**	85.3	**92.2**	43.4	**73.7**	**83.4**
SMAN (λ=3)	57.1	**85.7**	92.1	**43.6**	73.3	83.1
SMAN (λ=5)	56.9	85.2	91.5	43.1	73.4	83.1
SMAN (λ=7)	55.1	83.4	90.2	41.7	72.1	81.9
SMAN (λ=1, m_2=0.15)	56.1	84.2	90.5	42.1	72.3	81.9
SMAN (λ=1, m_2=0.2)	57.1	85.0	91.6	43.1	72.9	83.2
SMAN (λ=1, m_2=0.3)	56.7	85.1	91.6	43.0	73.1	83.3



(a) The impact of parameter λ (b) The impact of parameter m_2

Fig. 6. Impact of the balance parameter λ and margin parameter m_2 of intra-pair constraint loss on Flickr30K dataset. The sub-figure (a) illustrates the bidirectional retrieval result influenced by λ and sub-figure (b) shows the bidirectional retrieval result influenced by m_2, respectively.

in the common space to be much closer than unmatched ones for learning more discriminative visual-textual embedding. Experimental results on two benchmark datasets have demonstrated that our proposed method achieves the state-of-the-art performance on tackling the task of cross-modal image-text retrieval. In future work, we will explore how to design more effective network architecture for representing image and text for further mining and exploiting the underlying complementary semantic information between both modalities.

REFERENCES

[1] Y. Peng, X. Huang, and Y. Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, 2019.

[2] E. Aviv and L. Wolf. Linking image and text with 2-way nets. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, , 2017, pp. 4601–4611.

[3] X. Huang, Y. Peng, and M. Yuan. Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE T. Cybern.*, Online, 2018.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. Int. Conf. Learn. Representations*, 2015.

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1335–1344.

[6] F. Fartash, D. Fleet, J. Kiros, and S. Fidler. Vse++: improved visual-semantic embeddings. In *British Machine Vision Conference*, 2018, pp. 935–943.

[7] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *ArXiv preprint arXiv:1809.02983*, 2018.

[8] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[9] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7181–7189.

[10] Y. Hua and J. Du. Deep Semantic Correlation with Adversarial Learning for Cross-Modal Retrieval. In *Proc. IEEE Conf. Elec. Info. and Emergency Commun.*, 2019, pp. 256–259.

[11] Y. He, S. Xiang, C. Kang, and J. Wang. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[12] C. Deng, Z. Chen, X. Liu, and X. Gao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, 2018.

[13] G. Song, D. Wang, and X. Tan. Deep Memory Network for Cross-Modal Retrieval Deep Semantic *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1261–1275, 2018.

[14] L. Wu, Y. Wang, and L. Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, 2018.

[15] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with CNN visual features: A new baseline *IEEE Trans. Image Process.*, vol. 47, no. 2, pp. 449–460, 2016.

[16] K. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 201–216.

[17] J. Qi, Y. Peng, and Y. Yuan. Cross-media multi-level alignment with relation attention network. In *International Joint Conference on Artificial Intelligence*, 2018, pp. 892–898.

[18] Y. Peng, W. Zhu, Y. Zhao, C. Xu, Q. Huang, H. Lu, Q. Zheng, T. Huang, and W. Gao. Cross-media analysis and reasoning: advances and directions. In *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 44–57, 2017.

[19] M. Yuan, and Y. Peng. Bridge-GAN: Interpretable Representation Learning for Text-to-image Synthesis. In *IEEE Trans. Circuits Syst. Video Technol.*, Online, 2019.

[20] Q. Wang, S. Liu, J. Chanussot, and X. Li. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[21] Y. Wu, S. Wang, G. Song, and Q. Huang. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2088–2096.

[22] Y. Song, and M. Soleymani. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1979–1988.

[23] J. Gu, H. Hu, L. Wang, Y. Wei, and J. Dai. Learning region features for object detection. In *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 381–395.

[24] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, vol. 16 no. 12, pp. 2639–2664, 2004.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 779–778.

[26] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2310–2318.

[27] Y. Huang, Q. Wu, C. Song, and L. Wang. Learning semantic concepts and order for image and sentence matching. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 6163–6171.

[28] Z. Ji, K. Xiong, Y. Pang, and X. Li. Video summarization with attention-based encoder-decoder networks. IEEE Trans. Circuits Syst. Video Technol., Online, 2019.

[29] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3128–3137.

[30] R. Kiros, R. Salakhutdinov, and R.S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Proc. Adv. Neural Inf. Process. Syst. Workshop*, 2014.

[31] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4437–4446.

[32] A. Krizhevsky, I.d Sutskever, and G. E.d Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[33] P. Anderson, X. He, C. Buehler, d. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and Top-down Attention for Image captioning and Visual Question Answering. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 6077–6086.

[34] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 201–216.

[35] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. IEEE Int. Conf. Comput. Vis*, 2017, pp. 1890–1899.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[37] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. In *Proc. Int. Conf. Learn. Representations*, 2017.

[38] C Liu, Z Mao, W Zang, and Wang B. A neighbor-aware approach for image-text matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3970–3974, 2019.

[39] R. Liu, Y. Zhao, S. Wei, L. Zheng, and Y. Yang. Modality-invariant image-text embedding for image-sentence matching. *ACM Trans. Multimed. Comput. Commun.*, vol. 15, no. 1, pp. 1–27, 2019.

[40] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew. Learning a recurrent residual fusion network for multimodal matching. In *Proc. IEEE Int. Conf. Comput. Vis*, 2017, pp. 4107–4116.

[41] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proc. IEEE Int. Conf. Comput. Vis*, 2015, pp. 2623–2631.

[42] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. Int. Conf. Learn. Representations*, 2015.

[43] N. C. Mithun, R. Panda, E. E. Papalexakis, and A. Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *ACM International Conference on Multimedia*, 2018.

[44] H. Nam, J. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 299–307.

[45] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. IEEE Int. Conf. Comput. Vis*, 2015, pp. 2641–2649.

[46] N. Rasiwasia, J. Costa Pereira, E. Coviello, E. Doyle, Lanckriet G. R., R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[47] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proc. Conf. Empirical Methods Natural Lang. Process*, 2015, pp. 379–389.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, K. Lukasz, and I. Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[49] I. Vendrov, J. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *Proc. Int. Conf. Learn. Representations*, 2016.

[50] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 5005–5013.

[51] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian. Joint global and co-attentive representation learning for image-sentence retrieval. In *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1398–1406.

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[53] Z. Ji, H. Wang, J. Han, and Y. Pang. Saliency-Guided Attention Network for Image-Sentence Matching. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5754–5763.

[54] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3441–3450.

[55] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 21–29.

[56] Y. Zhang and H. Lu. Deep cross-modal projection learning for image-text matching. In *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 686–701.

[57] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.D. Shen. Dual-path convolutional image-text embedding. In *arXiv preprint arXiv:1711.05535*, 2017.

[58] Q. You, Z. Zhang, J. Luo. End-to-End Convolutional Semantic Embeddings. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 5735–5744.

[59] Z. Niu, M. Zhou, L. Wang, X. Gao and G. Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proc. IEEE Int. Conf. Comput. Vis*, 2017, pp. 1881–1889.

[60] S. Ren, K. He, R. Girshick, J Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[61] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.

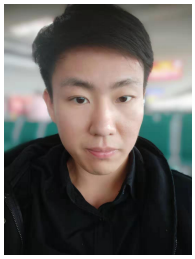
[62] T. Zhang and J. Wang. Collaborative quantization for cross-modal similarity search. In *IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2036–2045.

[63] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. In *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 1, pp. 2048–2057, 2016.

[64] Y. Wu, S. Wang, G .Song, and Q .Huang. Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. In *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4299–4312, 2019.

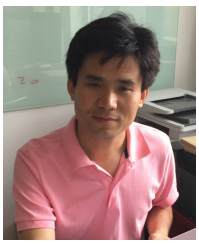
[65] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, 2019.

Zhong Ji received the Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 2008. He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. His current research interests include machine learning, computer vision, multimedia understanding, and video summarization.



Haoran Wang received the B.S. degree from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2014. He is currently pursuing the Ph.D. degree in School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision and multimodal learning.

Jungong Han is a tenured Associate Professor of Data Science at University of Warwick, UK. Previously, he was a senior lecturer with the School of Computing and Communications at Lancaster University, Lancaster, UK, and was with the Department of Computer Science at Northumbria University, UK.



Yanwei Pang received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2004. He is currently a Professor with the School of Electronic Information Engineering, Tianjin University, Tianjin, China. He has authored over 80 scientific papers. His current research interests include object detection and recognition, vision in bad weather, and computer vision.

APPENDIX A
SUPPLEMENTARY MATERIALS

A. Effect of intra-pair constraint of Bidirectional Triplet Ranking Loss

From Fig.6(a), we can observe that the performance of our model maintains stable when the balance parameter λ is in the range of 1 to 5. When λ continues to increase, we conjecture what makes the slight performance drop is that too strong intra-pair constraint may jeopardize the generalization ability of model. Moreover, as illustrated in Fig.6(b), in the case of fixing the value of the balancing parameter λ at 1 and changing the value of m_2, we find that the performance degrades slightly when m_2 drop to 5 from 10. It indicates that extremely small intra-pair constraint margin may, on the contrary, have an adverse effect on the performance of our model. The potential reason is that the model trained on dataset with limited training data cannot always guarantee to exhibit sufficient generalization ability.

B. Analysis on Computational Cost

Here, we analyse the computation cost of our model and make comparison it between the existing approaches. Assuming the total amount of image-text pairs is N, we compare the computational complexity of our model with two strides of previous methods on measuring the cross-modal similarity. Note that the comparison is made according to the common assumption that the dimensions of joint space for all models are equivalent. The most computation burden of our SMAN comes from the multi-modal attention module with computational complexity of $O(N^2)$. For comparison, the first category of works can be summarized as global joint embedding based models. The computational complexity of them is $O(N)$ since they only consider the intra-modal information to learn the modal-specific embeddings. The second stride of studies are characterized by employing the extra detectors [60] to align both modalities at local levels. Assuming the region numbers of image and those of texts are d_1 and d_2 respectively, their computational complexities are approximately equal to $O(d_1 \cdot d_2 \cdot N^2)$. Considering our SMAN outperforms most existing global embedding based models, it has capacity to achieve balance between effectiveness and efficiency.

C. Qualitative Results

1) Results of Bidirectional Image and text retrieval: To further qualitatively verify the effectiveness of our proposed model, we select several representative images and captions to show their corresponding retrieval results on Flickr30k, respectively. Fig.7 presents the qualitative results of text retrieval given image queries. For each image query, we list the top-5 retrieved sentences ranked by the similarity scores predicted by our model. Fig.8 presents the qualitative results of image retrieval given text queries. For each text query we display the top-3 retrieved images, ranking from left to right. We outline the matched results in green and mismatched results in red.

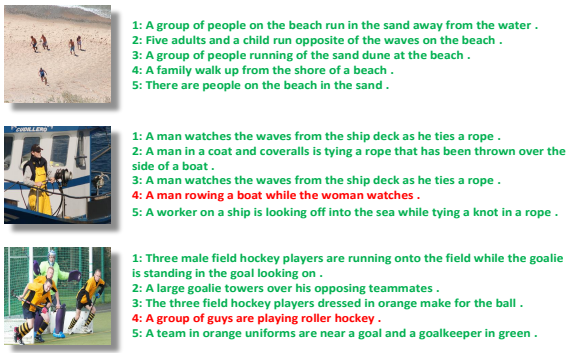


Fig. 7. Qualitative results of text retrieval given image queries on Flickr30K dataset. For each image query, the top-5 corresponding ranked sentences are presented. We observe that our SMAN model retrieves the correct results in the top ranked sentences for image queries, our model also finds some relatively reasonable mismatches. (Best viewed in color when zoomed in.)

Fig. 8. Qualitative results of image retrieval given text queries on Flickr30K dataset. For each text query, the top-3 corresponding ranked images are presented. We observe that our SMAN model retrieves the correct results in the top ranked images even for text queries. Meanwhile, it also finds some relatively reasonable mismatches. (Best viewed in color when zoomed in.)

2) Visualization of Attention: The qualitative results on Flickr30k from image-to-text and text-to-image retrieval are depicted in Fig.9 and Fig.10 with visualization of attention heatmaps, respectively. As mentioned in section IV, the DAN [44] method can be taken as a special case of our model. Considering the code of this work is not publicly released, we re-implement its attention scheme and visualize the corresponding attention heatmaps for comparison, which are listed in the second column in both Fig.9 and Fig.10. Then, in the third and fourth column, the attention heatmaps of our SMAN are exhibited by two steps, *i.e.*, intra-modal attention heatmap and multi-modal attention heatmap, separately. As illustrated in Fig.9 and Fig.10, our proposed stacked multi-modal attention can effectively discover the significant semantic information contained in both modalities through stepwise attention reasoning. More concretely, it tends to capture the main informative objects (*e.g.* people, horse, child, etc.) in a coarse-grained level at the first step, and then focuses on the more meaningful parts in a more fine-grained level at the second attention step.

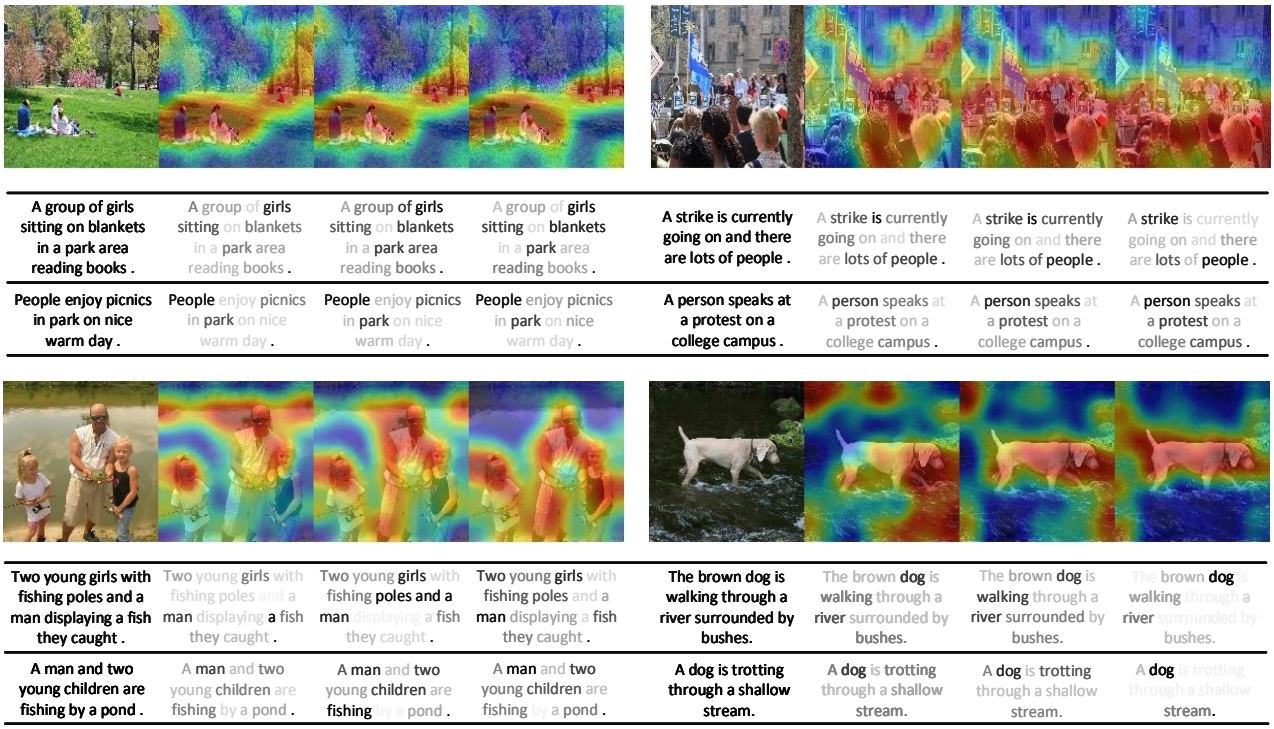


Fig. 9. Qualitative results from image-to-text retrieval with attention visualization. For each image query, the query image and its corresponding top two retrieved sentences are shown from top to bottom; the original image (sentence), the DAN [44] attention, the intra-modal of SMAN, and multi-modal attention heatmaps of SMAN are shown from left to right. (Best viewed in color when zoomed in.)

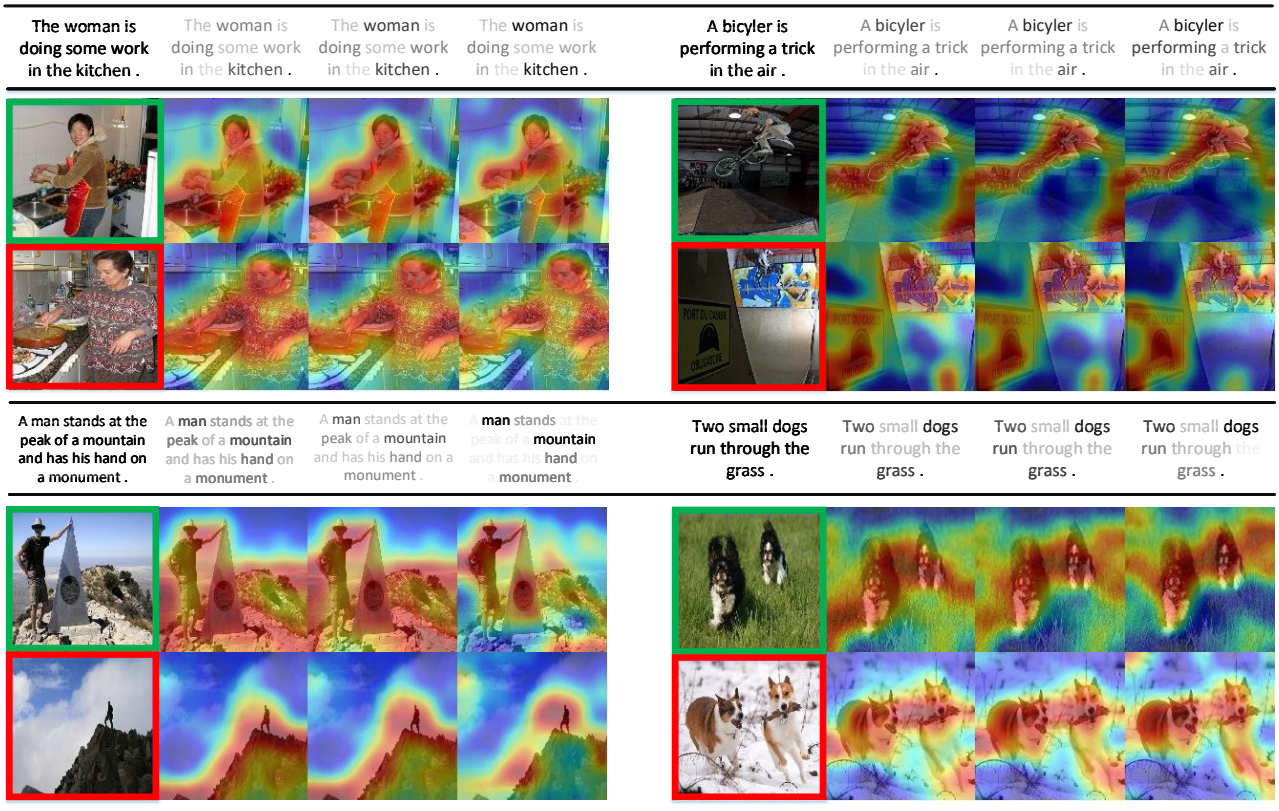


Fig. 10. Qualitative results from sentence-to-image retrieval with attention visualization. For each sentence query, the query sentence and the top two retrieved images are shown from top to bottom; the original sentence (image), the DAN [44] attention, the intra-modal and multi-modal attention heatmaps of SMAN are shown from left to right. Red and green boxes indicate mismatched and matched images, respectively. (Best viewed in color when zoomed in.)