Multi-Modal Memory Enhancement Attention Network for Image-Text Matching

ZHONG JI, ZHIGANG LIN, HAORAN WANG, AND YUQING HE
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
Corresponding author: Haoran Wang (haoranwang@tju.edu.cn)

This work was supported in part by the Natural Science Foundation of Tianjin under Grant 19JCYBJC16000, and in part by the National Natural Science Foundation of China (NSFC) under Grant 61771329.

ABSTRACT Image-text matching is an attractive research topic in the community of vision and language. The key element to narrow the “heterogeneity gap” between visual and textual data lies in how to learn powerful and robust representations for both modalities. This paper proposes to alleviate this issue to achieve the fine-grained visual-textual alignment from two aspects: exploiting attention mechanism to locate the semantically meaningful portion and leveraging the memory network to capture the long-term contextual knowledge. Unlike most existing studies solely focus on exploring the cross-modal associations at the fragment level, our designed Collaborative Dual Attention (CDA) module is able to model the semantic interdependencies from both perspectives of fragment and channel. Furthermore, considering the usage of long-term contextual knowledge contributes to compensate for detailed semantics concealed in the rarely appeared image-text pairs, we present to learn the joint representations by constructing a Multi-Modal Memory Enhancement (M3E) module. Specifically, it sequentially restores the intra-modal and multi-modal information into the memory items, and they conversely persistently memorize cross-modal shared semantics to improve the latent embeddings. By incorporating both CDA and M3E modules into a deep architecture, our approach generates more semantically consistent embeddings for representing images and texts. Extensive experiments demonstrate our model can achieve the state-of-the-art results on two public benchmark datasets.

INDEX TERMS Image-text matching, attention mechanism, memory network.

I. INTRODUCTION
In recent years, with the prevalence of deep learning in computer vision and natural language processing, vision and language understanding [1] has made tremendous progress. It includes a variety of downstream applications, such as image caption [2], visual question answering (VQA) [3], visual grounding [4] and image-text matching [5], [6]. As the central task of this community, in this paper, we focus on addressing the problem of image-text matching, which aims at seeking out the corresponding textual descriptions given images or the corresponding images given textual descriptions. Although great progress has been made, it is still quite challenging due to the existence of “heterogeneity gap” [10], which indicates that both image and text contain rich semantic information but reside in heterogeneous modalities. Thus, it is an enormous challenge to effectively measure the cross-modal similarity.

The key to narrowing the “heterogeneity gap” is learning more discriminative representations for both modalities on the joint embedding space. At the early stage, the main line [6]–[9] of research for image-text matching is to learn a joint embedding space, in which related images and texts can be mapped into the common global representations. However, for the global representations, it’s hard to pinpoint the significant portion of both modalities with meaningful semantics, i.e., image regions and textual words. To resolve this issue, some recent many approaches resort to employ attention mechanism [2], [11] on local representations, which attempts to focus on semantically salient parts of images and texts for capturing important information. Among those attention based methods, the self-attention embedding [52] network is proposed to exploit the intra-modal attention to capture the fragment dependencies among image patches according to the contextual correlations, which is characterized...
by efficient computing compared to those [48], [53] based on exhaustively calculating the pairwise similarities at local levels. However, these methods mostly focus on the exploring the relationships from the perspective of fragment information, while neglect the semantic dependencies along the channel dimension contained in features. To compensate for this problem, we design our representation module to lay emphasize on meaningful parts along both principal dimensions of features: channel and fragment axes, which aims at enhancing the discrimination for representing both modalities by blending cross-channel and fragment information together. To achieve this, we sequentially perform self-attention [44] mechanism at both channel and fragment levels, so that each of the branches can learn “what” and “where” are more semantically significant to attend in the channel and fragment axes respectively. Consequently, the image-text matching results validate our proposed module efficiently and effectively contributes to mine the meaningful parts of both modalities by learning which information to emphasize or suppress.

On the other hand, most of the existing methods are limited by sorely relying on employing the instance-level data, e.g., image-text pairs, to achieve the cross-modal alignment. Nevertheless, we know that it’s absolutely impossible that the visual-textual instance pairs are all available in the training stage in real scenario, resulting in the obstacle to associate the rarely appeared images and texts. A feasible solution to alleviate the above issue is incorporating the contextual knowledge into the learning procedure of image-text matching. In this paper, we attempt to achieve this idea by leveraging the memory network [14]–[17], [21] to store the long-term semantic relevance extracted from the heterogeneous data. It utilizes the memory components to preserve the cross-modal knowledge in memory, meanwhile extracts the associated contents from them to enhance the semantic representation ability conversely. To be specific, distinct from the previous methods that sorely employed intra-modal knowledge [14] or unidirectional cross-modal information [39] to build the memory components, our Multi-Modal Memory Enhancement (M3E) module is constituted by sequentially combining two different submodules. It first encodes the modal-specific information into the memory items to update the intra-modal memory vectors. Then, the multi-modal memory vectors are built via fusing the formers to capture the multi-modal semantics. This operation allows us to fully exploit the mutual complementarity between intra-modal and inter-modal information to preserve the semantic consistency of heterogeneous data.

To sum up, in this paper, an end-to-end architecture dubbed Multi-modal Memory Enhancement Attention Network (M3A-Net) is developed by us to address the problem of image-text matching. To learn the robust and discriminative cross-modal representations, we first build an attention based representation module by taking advantage of self-modal information. Then, a novel Multi-Modal Memory Enhancement representation module is introduced to import more contextual knowledge into the learning framework via leveraging multi-modal semantic messages.

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

1) A Collaborative Dual attention (CDA) module is proposed for learning discriminative representations for image and text. Taking fragment relationships and channel relationships into consideration sequentially, it employs the self-attention mechanism to capture the rich long-range dependencies contained in both modalities.

2) We present a Multi-Modal Memory Enhancement (M3E) module, which contributes to exploit long-term contextual knowledge to improve the joint embeddings. It is characterized by combining both intra-modal and multi-modal information to constitute the memory components, which fully utilizes the semantically complementary contents to learn the cross-modal alignment.

3) Experimental results validate that our M3A-Net is capable of achieving the state-of-the-art performance on two benchmark datasets: Flickr30k [18] and MSCOCO [19].

II. RELATED WORK

A. IMAGE-TEXT MATCHING

The core issue of most existing studies [9], [26], [27], [33], [40], [42] for image-text matching can summarized as learning the joint representations for both modalities. In [42], the framework of visual semantic embedding is proposed to align images and related texts, where the authors trained a visual recognition model with both labeled images and asemantic information form unannotated text data. Karpathy and Li [27] made an attempt to perform local similarity learning for cross-modal data by using the inferred alignments to learn to generate novel descriptions of image regions. Wang et al. [9] applied a dual-path deep network for visual-semantic embedding, accompanied by a combination of intra-modal and inter-modal constraints. Ranking loss [6], [9], [26], [33], [54] is a popular objective function to constraint the model of image-text matching. In [54], Wang et al. systematically investigated important components of both embedding and similarity networks, including loss functions and different ways of sampling positive and negative examples. Further, Fartash et al. [26] proposed to mine hard negatives in the ranking loss to benefit the joint representation learning. To learn more discriminative representations for images and sentences, Zheng et al. [33] utilized instance loss as a complement for ranking loss, which is a softmax loss that classifies an image-text pair into one of a large number of classes. Yao et al. [50] proposed a discrete robust supervised hashing model by learning a robust similarity matrix to guide the hash codes learning for cross-modal retrieval. Li et al. [55] proposed a reasoning model based on graph convolutional networks to enhanced visual representations by region relationship reasoning and global semantic reasoning. Xu et al. [51] applied graph embedding method to ensure the approximation of paired images.
and texts in joint embedding space. Similar to the above approaches, we lay emphasis on taking image-text matching as the representations learning procedure for cross-modal data.

B. DEEP ATTENTION MECHANISM

Attention mechanism aims at focusing on certain relevant parts of visual or textual inputs, which is beneficial to learn more discriminative representations. It is first applied on the area of natural language processing, in which machine translation have made great progress by leveraging attention based encoder-decoder network. As one of the seminal work, Vaswani et al. [44] proposed transformer network, based on scaled dot-product attention and multi-head attention, to draw global dependencies between input and output. Later, attention mechanisms have also been deployed in computer vision tasks such as image classification [45], image captioning [2] and object detection [46]. For instance, Hu et al. [12] introduced a Squeeze-and-Excitation (SE) Network that produces significant performance improvement for image classification. Recently, a number of approaches [47], [48] have exploited attention mechanism to benefit image-text matching. For example, Nam et al. [47] presented Dual Attentional Network that utilizes intra-modal context as a guidance to enhance discriminations of both visual and textual representations. Analogously, Lee et al. [48] explored cross modal attention mechanism to discover latent alignments between image regions and textual words. In contrast to them, our Collaborative Dual Attention (CDA) module simultaneously seeks for the attentive parts along with both channel and patch dimensions of the cross-modal representations. It is conductive to locate more semantically significant information, meanwhile just brings about slight computation burden.

C. MEMORY NETWORKS

Since Graves et al. [20] proposed Neural Turing Machine (NTM) that designs external memory to interact with the deep neural networks, memory networks have become increasingly popular in various tasks in image and language processing. Later, memory network is further improved and is able to be trained end-to-end [21]. To enhance the ability of memory networks on multi-modal learning domain, Xiong et al. [22] developed dynamic memory networks by introducing a new input module to represent images. To store prior knowledge more flexibly, Miller et al. [23] proposed Key-Value Memory Networks (KV-MemNN), which encodes prior knowledge for the considered task and to utilize possibly complex transforms between keys and values. Recently, based on KV-MemNN, Cai et al. [16] explored Memory Matching Networks to additionally integrate the contextual information. Although these studies [14], [15], [17] on vision and language understanding have achieved good results by adding external memory, most of them do not take full advantage of multi-modal information that benefits the visual-semantic embedding.

In this work, we store long-term contextual knowledge in structured memory items and make reference based on them in order to improve the representation ability for both images and texts. The main distinction between our proposed Multi-Modal Memory Enhancement (M3E) module and the existing methods lies in that we sequentially build the intra-modal and multi-modal memory components to restore more available detailed semantics, which assists to further achieve precise cross-modal alignment.

III. PROPOSED M3E-NET

In this section, we will elaborate on our proposed Multi-modal Memory Enhancement Attention Network (M3A-Net) from the following four aspects: 1) Representation learning for Image and Text, 2) Collaborative Dual Attention (CDA) module, 3) Multi-Modal Memory Enhancement (M3E) module, 4) Objective Function. For simplicity, we only introduce image branch in CDA module and M3E module. The structure of text branch is similar to the image branch.

A. REPRESENTATION LEARNING FOR IMAGE AND TEXT

1) VISUAL REPRESENTATION

We employ the ResNet-152 [24] pretrained on ImageNet [25] as visual feature encoder. We first feed image into Convolutional Neural Network (CNN) and then take the feature map before the final average pooling layer as the visual feature \( \mathbf{V} \in \mathbb{R}^{7 \times 7 \times 2048} \).

2) TEXTUAL REPRESENTATION

Given \( L \) words form a sentence, we first follow [13], [49] to represent them with one-hot vector and then implement word embedding on them. Then, we feed word vectors into bi-directional Gated Recurrent Units (GRU) and take the output of GRU as initialized textual feature \( \mathbf{S} \in \mathbb{R}^{L \times 1024} \).

B. COLLABORATIVE DUAL ATTENTION MODULE

To obtain more discriminative features, we introduce our CDA module that consist of channel attention and fragment attention, respectively. Different from most existing attention schemes [2], [21], our CDA module give consideration to both channel and fragment aspect to dig out more semantically meaningful information. As shown in Fig. 2, given an initialized visual feature \( \mathbf{x} = \mathbf{V} \in \mathbb{R}^{7 \times 7 \times 2048} \), CDA module sequentially infers a 1D channel attention map \( \mathbf{A}_c^y \in \mathbb{R}^{1 \times 1 \times 2048} \) and a 3D fragment attention map \( \mathbf{A}_f^y \in \mathbb{R}^{7 \times 7 \times h} \), where \( h \) is the number of fragment attention coefficients, \( t = \{1, 2, \cdots, T\} \) and \( T \) is the number of training images. It’s worth noting that, for textual branch, we are able to get channel attention map \( \mathbf{A}_c^e \in \mathbb{R}^{L \times 1024} \) and fragment attention map \( \mathbf{A}_f^e \in \mathbb{R}^{L \times h} \) by feeding \( \mathbf{x} = \mathbf{S}^e \in \mathbb{R}^{L \times 1024} \) into CDA module. For visual branch, the CDA module’s process can be summarized as:

\[
\mathbf{v}_c^t = \mathbf{A}_c^y(\mathbf{V}_c^t) \circ \mathbf{V}_c^t + \mathbf{V}_c^t, \tag{1}
\]

\[
\mathbf{v}_{cf}^t = N_2(\mathbf{W}_{nf}^r(\mathbf{A}_f^y(\mathbf{V}_c^t) \circ (\mathbf{v}_c^t)^T) + \text{GAP}(\mathbf{V}_c^t))). \tag{2}
\]
FIGURE 1. The architecture of our proposed M3A-Net. It takes visual feature \( V_t \) and word-level textual initial feature \( S_t \) as inputs from both modalities, respectively. Taking visual branch for instance (the same goes for textual branch): First, we feed \( V_t \) into channel and fragment attention module to obtain the attentive visual feature \( v_{t cf} \). Next, we take the first visual read vector \( r_{v1} \) as guidance to enhance the representation learning of image, generating memory based first visual feature \( v_{tm1} \). Similarly, we feed the second visual read vector \( r_{v2} \) into memory guidance block and then obtain the memory based second visual feature \( v_{tm2} \). Finally, we compute the cross-modal similarity between visual feature \( v_{tm2} \) and textual feature \( s_{tm2} \).

FIGURE 2. Diagram of collaborative dual attention module.

where \( W^c_v \in \mathbb{R}^{1 \times h} \) and \( W^e_v \in \mathbb{R}^{D \times 2048} \) are the transformation matrices of the fully-connected layer, \( D \) is the dimensionality of the joint embedding space, and \( V_t^c \in \mathbb{R}^{7 \times 7 \times 2048} \) and \( v_{cf}^e \in \mathbb{R}^D \) are the outputs of channel attention block and the whole CDA module, respectively. \( GAP \) and \( N_2 \) represent global average pooling and \( L_2 \) normalization, and symbol \( \odot \) is an element-wise multiplication. Similarly, we can obtain the output \( s_{cf}^e \in \mathbb{R}^D \) of the textual CDA module by setting the textual embedding matrix to \( W^e_s \in \mathbb{R}^{D \times 1024} \). Next, we describe the detailed architectures of these two attention sub-blocks.

**a: CHANNEL ATTENTION**

To exploit the inter-channel relationship of features, we introduce a channel attention map \( A^c_v \). Channel attention has benefited many vision tasks via capturing interdependencies among channels to selectively highlight important features and suppress unnecessary ones. Inspired by Squeeze-and-Excitation Networks [12], \( A^c_v \) is computed as:

\[
A^c_v(V_t) = \sigma(W^c_e v^e(W^c_c(GAP(V_t))))
\]

where \( W^c_e \in \mathbb{R}^{2048/r \times 2048} \), \( W^c_c \in \mathbb{R}^{2048 \times 2048/r} \), \( GAP \) represents global average pooling, \( \sigma \) and \( \gamma \) denote the sigmoid function and ReLU function. \( r \) is reduction ratio, which is used to reduce the model complexity.

**b: FRAGMENT ATTENTION**

Different from the channel attention, the fragment attention aims at attending to different regions of an instance by utilizing the locational relationship of features. Concretely, we employ a multi-head self-attention module that is composed of two layers perceptron to obtain \( h \) fragment attention maps \( A^f \). Compared with other fragment attentions,
multi-head self-attention yield diverse and refined locational weights. Given the visual output \( V_c \) of channel attention block, \( A^c_j \) is computed as:

\[
A^c_j(V^c_c) = \delta(W^c_{f2}(\tau(W^f_{f1}(V^c_c)))).
\]

(4)

where \( W^c_{f1} \in \mathbb{R}^{1024 \times 2048}, W^f_{f1} \in \mathbb{R}^{h \times 1024}, \tau \) and \( \delta \) denote the tanh function and softmax function, respectively.

C. MULTI-MODAL MEMORY ENHANCEMENT MODULE

In this section, we introduce our Multi-Modal Memory Enhancement (M3E) module that leverages memory units to remember the long-term contextual and informative knowledge for aligning both modalities. Note that we just utilize the second read controller to further extract more critical information for measuring the similarity between image and text. For simplicity, we only introduce image branch, and the structure of text branch is similar to the image branch.

1) MEMORY DEFINITION

The image memory at time step \( t \) is denoted as \( M^i_t \in \mathbb{R}^{N \times D_{ma}} = (m^i_t(i))_{i=1}^{N} \) (\( m^i_t(i) \in \mathbb{R}^{N \times D_{ma}} = (m^i_t(i))_{i=1}^{N} \) for text branch) in our M3E module, where \( N \) and \( D_{ma} \) are the number of memory locations and the vector length of each location, separately.

2) MEMORY UPDATE AND READ

Memory Networks [41] manipulates an external memory block that can flexibly read and update. Inspired by it, we design the update and read strategies that are based on content-based addressing mechanism. As shown in Fig. 1, each memory slot \( m^i_t(i) \) could be updated with the output features of CDA module to enhance previously generic memory, as well as read out to guide memory guidance block to select prediction-related visual information.

a: MEMORY UPDATE

Before reading memory, we need to write selected visual information to update memory. To control update extent, we introduce visual memory update gate \( w^v_m(i) \):

\[
w^v_m(i) = 1 - \sigma(V_{m} \cdot m_{v,m}(i)),
\]

(5)

where \( \sigma \) denote the sigmoid function, \( V_{m} \) is the output of the visual CDA module at time step \( t \). Then, the visual information can be written into the memory as follows:

\[
m^v_m(i) = (1 - w^v_m(i)) \circ m^i_t(i) + w^v_m(i) \circ V_{m},
\]

(6)

where \( i \in [1, N] \) denotes \( i \)-th memory location. Specifically, the same update strategy was employed to update text memory slots \( m^t_m(i) \), \( \circ \) denotes an element-wise multiplication.

b: MEMORY READ

When the above updating operation is finished, we utilize content-based addressing mechanism to selectively read memory slots:

\[
w^v_r(i) = \text{softmax}(S(V_{r,m}, m^v_m(i))),
\]

(7)

\[
S(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}.
\]

(8)

where the visual read weight \( w^v_r(i) \in [0, 1] \) represents normalized weight of the \( i \)-th memory slot. Then, the first visual read vector \( r^v_t \in \mathbb{R}^{D} \) can be calculated as:

\[
r^v_t = \sum_{i=1}^{N} w^v_r(i) \circ m^v_m(i).
\]

(9)

3) MEMORY BASED FEATURE ENHANCEMENT

a: FEATURE GENERATION BASED ON ONCE READ

After obtaining the first visual read vector \( r^v_t \), we propose Memory Guidance Block (see Fig. 1.) to guide feature generation. The process of visual feature generation can be defined as follows:

\[
v^1_r = W_{m11} v^f_r,
\]

(10)

\[
v^t_{m1} = \sigma(W_{m12}(r^v_t + v^t_r)) \cdot v^1_r + v^t_r,
\]

(11)

where \( W_{m11} \in \mathbb{R}^{D \times D} \) and \( W_{m12} \in \mathbb{R}^{D \times D} \) are the transformation matrices of the fully-connected layer, \( \sigma \) denotes the sigmoid function, \( v^t_{m1} \in \mathbb{R}^{D} \) is visual output feature based on once read. Analogous to visual branch, we can get textual output feature \( s^t_{m1} \in \mathbb{R}^{D} \) by using similar procedure.

b: FEATURES GENERATION BASED ON TWICE READ

For visual semantic embedding, the above proposed CDA module and memory module provide effective guidance at self-view of intra-modal. However, due to the difficulty of bridging the semantic gap and the inequality of information contained in image and text, the lack of cross modal guidance information restricts further improvement of matching performance. To circumvent this problem, we employ the multi-modal information to implement the second read operation, which is achieved based on performing the multimodal fusion of visual memory \( M^v_t \) and textual memory \( M^t_t \). As illustrated in Fig. 1, after obtaining the fused memory \( M^v_{m21} \in \mathbb{R}^{D_{ma} \times N \times N} \), we apply read controller that takes \( v^1_{m1} \) and \( s^t_{m1} \) as queries to obtain the second visual read vector \( r^v_{t2} \in \mathbb{R}^{D} \) and the textual read vector \( r^t_{t2} \in \mathbb{R}^{D} \):

\[
M^v_{m2} = M^v_{m21} \circ M^v_t,
\]

(12)

\[
w^v_{r2} = \text{softmax}(S(v^v_{r2}, m^v_{m2}(i)));
\]

(13)

\[
r^v_{t2} = \sum_{i} w^v_{r2}(i) \circ m^v_{m2}(i),
\]

(14)

where \( \circ \) denotes an slot-wise multiplication of memory and \( M^v_{m2} \in \mathbb{R}^{D_{ma} \times N \times N} = \{ m^v_{m2}(i) \}_{i=1}^{N} \). Similar to features generation procedure based on first memory reading, we feed \( r^v_{t2} \) and \( v^t_{m1} \) into the memory guidance block and the procedure is formulated as:

\[
v^2_r = W^v_{m21} v^1_r,
\]

(15)

\[
v^t_{m2} = N(\sigma(W^t_{m22}(r^v_{t2} + v^2_r)) \cdot v^1_r + v^t_{m2}),
\]

(16)

where \( W^v_{m21} \in \mathbb{R}^{D \times D} \) and \( W^t_{m22} \in \mathbb{R}^{D \times D} \) are the transformation matrices of the fully-connected layer. Besides, \( \sigma \) and \( N_2 \) denote the sigmoid function and \( L_2 \) normalization, respectively. \( v^t_{m2} \in \mathbb{R}^{D} \) is the visual output feature based on twice read. Similarly, we can get textual output feature \( s^t_{m2} \in \mathbb{R}^{D} \) by using similar procedure.
D. OBJECTIVE FUNCTION

The bidirectional triplet ranking loss [9], [26], [27] is a widely adopted ranking objective function for image-text matching. In this paper, we deploy it to impose constraint on the final features $v^t_{m2}$ and $s^t_{m2}$ to learn the joint embedding function, which is defined as follows:

$$
\mathcal{L}_{\text{rank}}(v^t_{m2}, s^t_{m2}) = \sum_{(v_{m2}, s_{m2})} \left[ \max\{0, \lambda - S(v^t_{m2}, s^t_{m2}) + S(v^t_{s2}, s^t_{s2})\} \
\quad + \max\{0, \lambda - S(v^t_{s2}, s^t_{s2}) + S(v^t_{m2}, s^t_{m2})\} \right],
$$

where $\lambda$ is a predefined margin parameter. Given a matched image-sentence pair $(v^t_{m2}, s^t_{m2})$, its corresponding negative samples are denoted as $v^t_{s2}$ and $s^t_{s2}$, respectively. Note that we also follow [26] to focus on the hardest negatives in a mini-batch.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

Two benchmark datasets are employed in our experiments. One is MSCOCO [19] that includes 123,287 images, each of which is annotated with 5 sentences. We follow the public dataset split [27], including 113,287 images for training, 1000 images for validation and 5000 images for testing. Following the most commonly used evaluation setting, we report the experimental results by averaging over 5 folds of the 1k test set and the full 5k test set, respectively. The other one is Flickr30k [18]. It consists of 317,83 images collected from the Flickr website, in which each image is associated with 5 sentences written by different Amazon Mechanical Turkers. We utilize the public protocol [7] to split the dataset into 19,783 images for training, 1000 images for validation and 1000 images for test. We report the performance evaluation of bi-directional retrieval on 1000 testing set.

For evaluation, we use the widely-used R@K as evaluation metric [8], [26], which is the abbreviation for recall at $K$ and is defined as the proportion of ground-truth matchings appears in the top $K$-ranked retrieval results. Besides, we employ an additional criterion “mR” that is suitable to evaluate the overall performance for cross-modal retrieval by averaging all six recall rates of R@K.

All our experiments are implemented in PyTorch toolkit with a single NVIDIA TITAN Xp GPU. As for image pre-processing, the input images are first resized to 256 $\times$ 256. Then, we follow [26] to use the mean feature vector for 10 crops of size of 224 $\times$ 224. For sentence, the dimensionality of word embedding space is set to 300, and the hidden state of GRU units is set to 1024. For channel attention, the reduction ratio $r$ is set to 16. The dimensionality of joint embedding space $D$ and the vector length of each memory’s slot $D_m$ is set to 1024. Empirically, we set the margin parameter $\lambda$ in Eq.(17) to 0.2. We train our model by Adam optimizer [28] with a learning rate 0.0002 and a mini-batch of 64. The network is trained for 50 epochs totally, where the learning rate is dropped to 0.00002 after 30 epochs. The default number of fragment attention coefficients $h$ and the default number of memory’s slot $N$ are set to $h = 4$ and $N = 32$, respectively.

B. COMPARISON WITH STATE-OF-THE-ART

We compare our model with several state-of-the-art approaches on datasets of MSCOCO and Flickr30k. It should be noted that the approaches adopting the bottom-up attention visual features [29] are not selected for comparison due to these approaches actually get benefit from the overlap between its training dataset [30] and the evaluation dataset [19]. Thus, it is uncertain to assure the fairness of the performance comparison for cross-modal matching.

The experimental results on the MSCOCO dataset are shown in Table 1. For 1k test dataset, it can be observed that our proposed M3A-Net achieves the best performances on all evaluation metrics. For example, it achieves 70.4% and 58.4% on R@1 with image and sentence as respective queries, which surpasses the second best performance of PVSE [13] by 1.2% and 3.2%, respectively. Particularly, M3A-Net significantly outperforms all competitor in terms of the criteria of mR. For 5k test dataset, we can observe that our model achieves the best performance in six evaluation metrics, except for that on R@10. Take R@1 for example, there are 3.7% and 5.9% performances gains against the second best PVSE approach [13] on sentence retrieval and image retrieval, respectively. The above evidences considerably validate superiority of our model on learning effective visual-semantic matching.

Table 2 shows the quantitative experimental results on Flickr30K dataset. Our M3A-Net also achieves the best performance in seven evaluation metrics. Taking R@1 for example, comparing with the second best performance, our

| Approach    | Sentence Retrieval | Image Retrieval |
|-------------|--------------------|-----------------|
|             | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| M3A-Net(Ours) | 70.4 | 91.7 | 96.8 | 58.4 | 87.1 | 94.0 |
|            | 38.3 | 65.7 | 76.9 | 64.9 | 82.1 |

| Dataset                | Measure     | 1K Test set   | 5K Test set   |
|------------------------|-------------|---------------|---------------|
|                        | R@10        | R@5           | R@10          | R@10          |
| DVSA [27]              | 22.2        | 48.2          | 61.4          | 15.2          | 37.7          | 50.5          | 58.5          |
| GMM-FV [31]            | 39.4        | 67.9          | 80.9          | 25.1          | 59.8          | 76.6          | 58.3          |
| m-CNN [32]             | 42.8        | 73.1          | 84.1          | 32.6          | 68.6          | 82.8          | 64            |
| DSPEI [9]              | 50.1        | 79.7          | 89.2          | 39.6          | 75.2          | 86.9          | 70.1          |
| DPC [33]               | 65.6        | 89.8          | 95.5          | 47.1          | 79.9          | 90.0          | 78.0          |
| 2WayNet [34]           | 55.8        | 75.2          |              | 39.7          | 63.3          |              |              |
| VSE+i+ [26]            | 64.7        | -             | 95.9          | 52.0          | 92.0          | -             |              |
| CMPM [35]              | 56.1        | 86.3          | 92.9          | 44.6          | 78.8          | 89.0          | 74.6          |
| BSSAN [37]             | 56.0        | 82.6          | 91.3          | 41.8          | 76.7          | 88.5          | 75.1          |
| CSE [36]               | 56.3        | 84.4          | 92.2          | 45.7          | 81.2          | 90.6          | 75.1          |
| UniVSE [40]            | 64.3        | 89.2          | 94.8          | 48.3          | 81.7          | 91.2          | 78.3          |
| PVSE [13]              | 69.2        | 91.6          | 96.6          | 55.2          | 86.5          | 93.7          | 82.1          |
| M3A-Net(Ours)          | 48.9        | 75.2          | 84.4          | 38.3          | 65.7          | 76.9          | 64.9          |
C. ABLATION STUDY

To further investigate the effects of the components in our M3A-Net, we present several quantitative results of ablation models. Note that our ablation studies are performed on the Flickr30k dataset. Specifically, the baseline approaches for comparisons are defined as follows:

- **BL**: This baseline model contains only the visual and textual feature encoders, i.e., ResNet-152 and GRU.
- **BL + CA**: This ablation model consists of the visual and textual feature encoders, accompanied with channel attention blocks.
- **BL + SA**: This ablation model consists of the visual and textual feature encoders, accompanied with fragment attention blocks.
- **BL + CDA**: This ablation model consists of the visual and textual feature encoders, accompanied with both channel attention blocks and fragment attention blocks.
- **BL + CDA + R1**: This ablation model consists of the visual and textual feature encoders, and channel and fragment attention modules, accompanied with memory guidance blocks based on once read.
- **BL + CDA + R1 + R2**: This ablation model consists of the visual and textual feature encoders, and channel and fragment attention modules, accompanied with memory guidance blocks based on twice read.

### TABLE 2. Comparisons of experimental results on Flickr30k testing set.

| Approach | Sentence Retrieval | Image Retrieval | mR |
|----------|--------------------|-----------------|----|
|          | R@1 R@5 R@10      | R@1 R@5 R@10   |    |
| DVSA [27] | 38.4 69.9 80.5 | 27.4 60.2 74.8 | 39.2 |
| GMM-FV [31] | 35.0 62.0 73.8 | 25.0 52.7 66.0 | 52.4 |
| m-CNN [32] | 33.6 64.1 74.9 | 26.2 56.3 69.6 | 54.1 |
| DSPE [9] | 40.3 68.9 79.9 | 29.7 60.1 72.1 | 58.5 |
| 2WayNet [34] | 49.8 67.5 - | 36.0 55.6 - | - |
| BSSAN [37] | 44.6 74.9 84.3 | 33.2 66.2 72.9 | 62.1 |
| CSE [36] | 44.6 74.3 83.5 | 36.9 69.1 79.6 | 64.7 |
| VSE++ [26] | 52.9 - 87.2 | 39.6 - 79.5 | - |
| CPMF [35] | 49.6 76.8 86.1 | 37.3 65.7 75.5 | 65.2 |
| SCO [39] | 55.5 82.0 89.3 | 41.1 70.5 80.1 | 69.7 |
| M3A-Net (Ours) | **58.1** 82.8 90.1 | **44.7** 72.4 81.1 | **71.5** |

### TABLE 3. Impact of various attention mechanisms on Flickr30k.

| Approach | Sentence Retrieval | Image Retrieval | mR |
|----------|--------------------|-----------------|----|
|          | R@1 R@5 R@10      | R@1 R@5 R@10   |    |
| BL       | 48.8 76.9 85.7     | 37.7 67.0 76.9 | 65.5 |
| BL+CA    | 53.0 78.8 87.3     | 41.2 69.6 78.9 | 68.1 |
| BL+SA(h=4) | 53.5 79.1 87.5 | 42.4 71.2 80.3 | 69.0 |
| BL+CDA(h=4) | **54.6** 80.2 **88.0** | **42.6** 71.4 **80.6** | **69.6** |

1) EFFECT OF COLLABORATIVE DUAL ATTENTION MODULE

To validate the effectiveness of our proposed CDA module for image-sentence retrieval, we implement another two attention schemes for comparison as well as offering the experimental results of the baseline module. Note that these ablation models remove memory module to simplify the model and save training time. As shown in Table 3 and Fig. 3, we can observe that the “BL + CDA” model outperforms other three comparative models in all seven evaluation metrics. Specifically, compared with the “BL” model, the “BL + CDA” model can bring about the performance boost by 5.8% and 4.9% in terms of R@1 on sentence retrieval and image retrieval, respectively. Besides, it is worth noting that the “BL + SA” model and the “BL + CA” model also out-perform the “BL” model by 3.5% and 2.7% in terms of mR, which verify that each part of the proposed Collaborative Dual Attention schemes benefit to improve performance for image-sentence retrieval. The above experimental results considerably demonstrate that our proposed attention scheme can provide more powerful guiding information for image-text retrieval by combining SE [12] based channel attention and multi-head self-attention [13] based fragment attention.

For the “BL + CDA” model in Flickr30k dataset, we further investigate in the influence of parameter h in multi-head self-attention. As illustrated in Fig. 4(a), we observe that the experimental performance (mR) increases continuously with the rise of h until it reaches the peak when $h = 4$, and then decreases when $h = 5$. We conjecture the main reason is that excessive head amounts for self-attention embeddings bring about heavy parameter amount, which may cause obstacle for model optimizing.

2) EFFECT OF MULTI-MODAL MEMORY ENHANCEMENT

To measure the influence of once read and twice read for our M3A module, we implement the “BL + CDA + R1” ablation model and “BL + CDA + R1 + R2” ablation model. Note that we set the number of multi-head self-attention embeddings h to 4 and set the size of memory slots N to 32 for image branch and text image in these two...
TABLE 4. Impact of memory module on Flickr30k.

| Approach | Sentence Retrieval | | | Image Retrieval | | |
|----------|-------------------|-------------------|------------------|------------------|------------------|
|          | R@1 | R@5 | R@10 | mR | R@1 | R@5 | R@10 | mR |
| BL+CDA   | 54.6 | 80.2 | 88.0 | 42.6 | 71.4 | 80.6 | 69.6 |
| BL+CDA+R1 | 56.2 | 81.4 | 88.5 | 43.8 | 71.8 | 80.8 | 70.4 |
| BL+CDA+R1+R2 | 58.1 | 82.8 | 90.1 | 44.7 | 72.4 | 81.1 | 71.5 |

FIGURE 4. Impact of varied controlling parameters on Flickr30k dataset. Figure (a) shows performance (mR) with different numbers of multi-head self-attention embeddings in the “BL + CDA” model. Figure (b) illustrates the effect of memory size $N$ in the “BL + CDA + R1 + R2” model.

Ablation experiments. The experimental results are shown in Table 4 and Fig. 3. It can be observed that the “BL + CDA + R1 + R2” ablation model that is selected as final model obtains the best performance and yields a result of (58.1%, 82.8%, 90.1%) and (44.7%, 72.4%, 81.1%) on (R@1, R@5, R@10) for text retrieval and image retrieval, respectively. Comparing with the “BL + CDA + R1” model, it achieves 1.9% gain on R@1 for sentence retrieval and 0.9% gain on R@1 for image retrieval. On the other hand, when applying memory module to enhance features learning, the “BL + CDA” model can collaborate with it effectively, resulting in performance gains of 1.6% and 1.2% on R@1 by performing the “BL + CDA + R1” on sentence retrieval and image retrieval, respectively. The above evidences verify that memory is beneficial to remember the previously informative knowledge on training stage and enhance semantic representation. In addition, the operation of twice read has the capability to boost retrieval performance by further mining discriminative information.

To investigate the influence of the number of memory slots $N$, we vary $N$ from 0 to 64 by fixing $h$ to 4 to conduct the experiments based on the “BL + CDA + R1 + R2” model, and then obtain the performance curve of mR, which is depicted in Fig. 4(b). It can be observed that the performance archives best results at $N = 32$. The results do not follow the empirical rule of that more memories can get better performance on visual question answering task. It indicates that appropriate number of memories can help to learn more discriminative features for image-sentence retrieval.

D. QUALITATIVE RESULTS

To further verify the effectiveness of our proposed model, the qualitative results from sentence retrieval and image retrieval are illustrated in Fig. 5 and Fig. 6, respectively. We observe that our proposed model retrieves the correct results in the top ranked sentences even for the image queries with complex scenes. Interestingly, our model also finds some reasonable mismatched sentences for image queries. Taking the last image in Fig. 5 for the example, the only mismatched sentence also captures some correct semantic concepts including “holding up” and “smart phone”.

V. CONCLUSION

In this paper, we proposed a novel Multi-modal Memory Enhancement Attention Network (M3A-Net) for achieving image-text matching. Concretely, we developed a Collaborative Dual Attention (CDA) module to fully exploit intramodal information, which performed self-attention from fragment and channel aspects to focus on the semantically meaningful parts of both modalities. Besides, to further compensate for the deficiency of the former, we proposed a Multi-Modal Memory Enhancement (M3E) module, which contributes to leverage long-term contextual knowledge to improve the joint embeddings. Experimental results on two benchmark datasets have demonstrated that our proposed approach has capacity to achieve the state-of-the-art performance for tackling the task of cross-modal image-text matching.

REFERENCES

[1] S. Dehaene, F. Pegado, L. W. Braga, P. Ventura, G. N. Filho, A. Jobert, G. Dehaene-Lambertz, R. Kolinsky, J. Morais, and L. Cohen, “How learning to read changes the cortical networks for vision and language,” *Science*, vol. 330, no. 6009, pp. 1359–1364, Nov. 2010.
[2] K. Xu, J. Ba, and R. Kiros, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[3] Z. Fang, J. Liu, Y. Li, Y. Qiao, and H. Lu, “Improving visual question answering using dropout and enhanced question encoder,” *Pattern Recognit.*, vol. 90, pp. 404–414, Jun. 2019.

[4] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, “Visual grounding via accumulated attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7746–7755.

[5] Y. Liu, Y. Guo, L. Liu, E. M. Bakker, and M. S. Lew, “CycleMatch: A cycle-consistent embedding network for image-text matching,” *Pattern Recognit.*, vol. 93, pp. 365–379, Sep. 2019.

[6] R. Kiros, R. Salakhutdinov, and R. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, 2014, pp. 1–13.

[7] 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. Represent.*, 2015, pp. 1–17.

[8] I. Vondrak, J. Kiros, S. Figler, and R. Urtasun, “Order-embeddings of images and language,” in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–12.

[9] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[10] X. Huang and Y. Peng, “Deep cross-media knowledge transfer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8837–8846.

[11] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[13] Y. Song and M. Soleymani, “Polysensome visual-semantic embedding for cross-modal retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1979–1988.

[14] G. Song and X. Tang, “Cross-modal retrieval via memory network,” in *Proc. Brit. Mach. Learn.*, 2017, pp. 1–12.

[15] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, “m3: Multimodal memory modelling for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7512–7520.

[16] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, “Memory matching networks for one-shot image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4080–4088.

[17] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid, “Visual question answering with memory-augmented networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6975–6984.

[18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flick30k entities: Collecting region-to-phrase correspondences for rich image-to-sentence models,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2641–2649.

[19] T.-Y. Lin, M. Maire, and S. Belongie, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 740–755.

[20] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing machines,” 2014, arXiv:1410.5401. [Online]. Available: http://arxiv.org/abs/1410.5401

[21] S. Sukhbaatar, J. Weston, and R. Fergus, “End-to-end memory networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[22] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for generating image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[24] H. Wu, J. Mao, Y. Zhang, J. Yang, L. Li, W. Sun, and W.-Y. Ma, “Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6699–6618.

[25] J. Weston, S. Chopra, and A. Bordes, “Memory networks,” in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[26] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, “Devise: A deep visual-semantic embedding model,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[27] A. Karpathy, A. Joulin, and F. F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[28] A. Vaswani, N. Shazeer, and N. Parmar, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[30] J. Gu, H. Hu, L. Wang, Y. Wei, and J. Dai, “Learning region features for object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 381–395.

[31] H. Nam, J.-W. Ha and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.

[32] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 212–228.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, arXiv:1301.3781. [Online]. Available: http://arxiv.org/abs/1301.3781
[50] T. Yao, Z. Zhang, L. Yan, J. Yue, and Q. Tian, “Discrete robust supervised hashing for cross-modal retrieval,” *IEEE Access*, vol. 7, pp. 39806–39814, 2019.

[51] G. Xu, X. Li, and Z. Zhang, “Semantic consistency cross-modal retrieval with semi-supervised graph regularization,” *IEEE Access*, vol. 8, pp. 14278–14288, 2020.

[52] Y. Wu, S. Wang, G. Song, and Q. Huang, “Learning fragment self-attention embeddings for image-text matching,” in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2088–2096.

[53] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, “Position focused attention network for image-text matching,” 2019, arXiv:1907.09748. [Online]. Available: http://arxiv.org/abs/1907.09748

[54] 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, Feb. 2019.

[55] K. Li, Y. Zhang, and K. Li, “Visual semantic reasoning for image-text matching,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4654–4662.

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. He has published more than 80 scientific articles. His current research interests include machine learning, computer vision, multimedia understanding, and video summarization.

ZHIGANG LIN received the B.E. degree from Xidian University, Xi’an, China, in 2018. He is currently pursuing the M.A.Sc. degree with the School of Electrical and Information Engineering, Tianjin University. His research interests include cross-modal retrieval and computer vision.

HAORAN WANG received the B.E. degree from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University. His research interests include cross-modal retrieval and computer vision.

YUQING HE received the Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 2007. He is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. He has published more than 20 scientific articles. His research interests are computer vision and image processing.

* * *