IMRAM: Iterative Matching with Recurrent Attention Memory for Cross-Modal Image-Text Retrieval*

Hui Chen[1], Guiguang Ding[1*], Xudong Liu[2], Zijia Lin[3], Ji Liu[4], Jungong Han[5]
[1]School of Software, BNRist, Tsinghua University
[2]Kwai Ads Platform; [3]Microsoft Research
[4]Kwai Seattle AI Lab, Kwai FeDA Lab, Kwai AI Platform
[5]WMG Data Science, University of Warwick
{jichenhui2012, ji.liu.uwisc, jungonghan77}@gmail.com
dinggg@tsinghua.edu.cn, liuxudong@kuaishou.com, zijlin@microsoft.com

Abstract

*Enabling bi-directional retrieval of images and texts is important for understanding the correspondence between vision and language. Existing methods leverage the attention mechanism to explore such correspondence in a fine-grained manner. However, most of them consider all semantics equally and thus align them uniformly, regardless of their diverse complexities. In fact, semantics are diverse (*i.e. *involving different kinds of semantic concepts), and humans usually follow a latent structure to combine them into understandable languages. It may be difficult to optimally capture such sophisticated correspondences in existing methods. In this paper, to address such a deficiency, we propose an Iterative Matching with Recurrent Attention Memory (IMRAM) method, in which correspondences between images and texts are captured with multiple steps of alignments. Specifically, we introduce an iterative matching scheme to explore such fine-grained correspondence progressively. A memory distillation unit is used to refine alignment knowledge from early steps to later ones. Experiment results on three benchmark datasets,* i.e. *Flickr8K, Flickr30K, and MS COCO, show that our IMRAM achieves state-of-the-art performance, well demonstrating its effectiveness. Experiments on a practical business advertisement dataset, named KWAI-AD, further validates the applicability of our method in practical scenarios.*

1. Introduction

Due to the explosive increase of multimedia data from social media and web applications, enabling bi-directional cross-modal image-text retrieval is in great demand and has become prevalent in both academia and industry. Meanwhile, this task is challenging because it requires to understand not only the content of images and texts but also their inter-modal correspondence [10].

In recent years, a large number of researches have been proposed and achieved great progress. Early works attempted to directly map the information of images and texts into a common latent embedding space. For example, Wang *et al*. [26] adopted a deep network with two branches to, respectively, map images and texts into an embedding space. However, these works coarsely capture the correspondence between modalities and thus are unable to depict the fine-grained interactions between vision and language.

To gain a deeper understanding of such fine-grained correspondences, recent researches further explored the attention mechanism for cross-modal image-text retrieval. Karpathy *et al*. [9] extracted features of fragments for each image and text (*i.e*. image regions and text words), and proposed a dense alignment between each fragment pair. Lee *et al*. [12] proposed a stacked cross attention model, in which attention was used to align each fragment with all fragments from another modality. It can neatly discover the fine-grained correspondence and thus achieves state-of-the-art performance on several benchmark datasets.

However, due to the large heterogeneity gap between images and texts, existing attention-based models, *e.g*. [12], may not well seize the optimal pairwise relationships among a number of region-word fragments pairs. Actually, semantics are complicated, because they are diverse (*i.e*. composed by different kinds of semantic concepts with different meanings, such as objects (*e.g*. nouns), attributes (*e.g*. adjectives) and relations (*e.g*. verbs)). And there generally exist strong correlations among different concepts, *e.g*. relational terms (*e.g*. verbs) usually indicate relation-

*This work was supported by the National Natural Science Foundation of China (Nos. U1936202, 61925107). Corresponding author: Guiguang Ding

ships between objects (*e.g*. nouns). Moreover, humans usually follow a latent structure (*e.g*. a tree-like structure [25]) to combine different semantic concepts into understandable languages, which indicates that semantics shared between images and texts exhibit a complicated distribution. However, existing state-of-the-art models treat different kinds of semantics equally and align them together uniformly, taking little consideration of the complexity of semantics.

In reality, when humans perform comparisons between images and texts, we usually associate low-level semantic concepts, *e.g*. objects, at the first glimpse. Then, higher-level semantics, *e.g*. attributes and relationships, are mined by revisiting images and texts to obtain a better understanding [20]. This intuition is favorably consistent with the aforementioned complicated semantics, and meanwhile, it indicates that the complicated correspondence between images and texts should be exploited progressively.

Motivated by this, in this paper, we propose an iterative matching framework with recurrent attention memory for cross-modal image-text retrieval, termed IMRAM. Our way of exploring the correspondence between images and texts is characterized by two main features: (1) an iterative matching scheme with a cross-modal attention unit to align fragments across different modalities; (2) a memory distillation unit to dynamically aggregate information from early matching steps to later ones. The iterative matching scheme can *progressively* update the cross-modal attention core to accumulate cues for locating the matched semantics, while the memory distillation unit can refine the latent correspondence by enhancing the interaction of cross-modality information. Leveraging these two features, different kinds of semantics are treated distributively and well captured at different matching steps.

We conduct extensive experiments on several benchmark datasets for cross-modal image-text retrieval, *i.e*. Flickr8K, Flickr30K, and MS COCO. Experiment results show that our proposed IMRAM can outperform the state-of-the-art models. Subtle analyses are also carried out to provide more insights about IMRAM. We observe that: (1) the fine-grained latent correspondence between images and texts can be well refined during the iterative matching process; (2) different kinds of semantics, respectively, play dominant roles at different matching steps in terms of contributions to the performance improvement.

These observations can account for the effectiveness and reasonableness of our proposed method, which encourages us to validate its potential in practical scenarios. Hence, we collect a new dataset, named KWAI-AD, by crawling about 81K image-text pairs on an advertisement platform, in which each image is associated with at least one advertisement textual title. We then evaluate our proposed method on the KWAI-AD dataset and make comparisons with the state-of-the-art models. Results show that our method performs considerably better than compared models, further demonstrating the effectiveness of our method in the practical business advertisement scenario. The source code is available at: `https://github.com/HuiChen24/IMRAM`.

The contributions of our work are three folds: 1) First, we propose an iterative matching method for cross-modal image-text retrieval to handle the complexity of semantics. 2) Second, we formulate the proposed iterative matching method with a recurrent attention memory which incorporates a cross-modal attention unit and a memory distillation unit to refine the correspondence between images and texts. 3) Third, we verify our method on benchmark datasets (*i.e*. Flickr8K, Flickr30K, and MS COCO) and a real-world business advertisement dataset (*i.e*. our proposed KWAI-AD dataset). Experimental results show that our method outperforms compared methods in all datasets. Thorough analyses on our model also well demonstrate the superiority and reasonableness of our method.

2. Related work

Our work is concerned about the task of cross-modal image-text retrieval which essentially aims to explore the latent correspondence between vision and language. Existing matching methods can be roughly categorized into two lines: (1) coarse-grained matching methods aiming to mine the correspondence globally by mapping the whole images and the full texts into a common embedding space, (2) fine-grained matching ones aiming to explore the correspondence between image fragments and text fragments at a fine-grained level.

Coarse-grained matching methods. Wang *et al*. [26] used a deep network with two branches of multilayer perceptrons to deal with images and texts, and optimized it with intra- and inter-structure preserving objectives. Kiros *et al*. [11] adopted a CNN and a Gate Recurrent Unit (GRU) with a hinge-based triplet ranking loss to optimize the model by averaging the individual violations across the negatives. Alternatively, Faghri *et al*. [4] reformed the ranking objective with a hard triplet loss function parameterized by only hard negatives.

Fine-grained matching methods. Recently, several works have been devoted to exploring the latent fine-grained vision-language correspondence for cross-modal image-text [9, 19, 6, 17, 12]. Karpathy *et al*. [9] extracted features for fragments of each image and text, *i.e*. image regions and text words, and aligned them in the embedding space. Niu *et al*. [19] organized texts as a semantic tree with each node corresponding to a phrase, and then used a hierarchical long short term memory (LSTM, a variant of RNN) to extract phrase-level features for text. Huang *et al*. [6] presented a context-modulated attention scheme to selectively attend to salient pairwise image-sentence instances. Then a

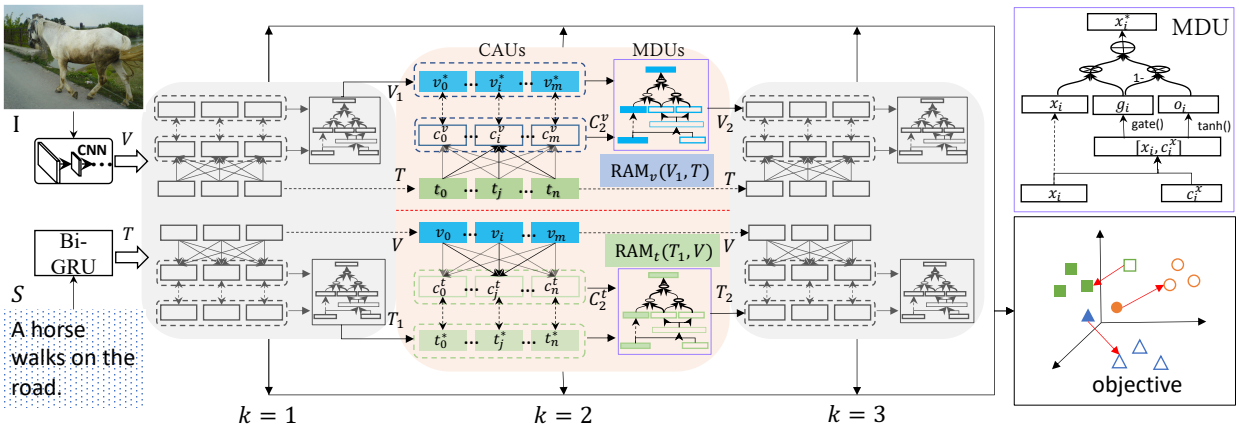

Figure 1. Framework of the proposed model.

multi-modal LSTM was used to sequentially aggregate local similarities into a global one. Nam *et al*. [17] proposed a dual attention mechanism in which salient semantics in images and texts were obtained by two attentions, and the similarity was computed by aggregating a sequence of local similarities. Lee *et al*. [12] proposed a stacked cross attention model which aligns each fragment with all other fragments from the other modality. They achieved state-of-the-art performance on several benchmark datasets for cross-modal retrieval.

While our method targets the same as [9, 12], differently, we apply an *iterative* matching scheme to refine the fragment alignment. Besides, we adopt a memory unit to distill the knowledge of matched semantics in images and texts after each matching step. Our method can also be regarded as a sequential matching method, as [17, 6]. However, within the sequential computations, we transfer the knowledge about the fragment alignment to the successive steps with the proposed recurrent attention memory, instead of using modality-specific context information. Experiments also show that our method outperforms those mentioned works.

We also noticed that some latest works make use of large-scale external resources to improve performance. For example, Mithun *et al*. [16] collected amounts of image-text pairs from the Internet and optimized the retrieval model with them. Moreover, inspired by the recent great success of contextual representation learning for languages in the field of natural language processing (ELMO [21], BERT [3] and XLNet [27]), researchers also explored to apply BERT into cross-modal understanding field [1, 13]. However, such pre-trained cross-modal BERT models[1] require large amounts of annotated image-text pairs, which are not easy to obtain in the practical scenarios. On the contrary, our method is general and unlimited to the amount of data. We leave the exploration of large-scale external data to future works.

[1]Corresponding codes and models are not made publicly available.

3. Methodology

In this section, we will elaborate on the details of our proposed IMRAM for cross-modal image-text retrieval. Figure 1 shows the framework of our model. We will first describe the way of learning the cross-modal feature representations in our work in section 3.1. Then, we will introduce the proposed recurrent attention memory as a module in our matching framework in section 3.2. We will also present how to incorporate the proposed recurrent attention memory into the iterative matching scheme for cross-modal image-text retrieval in section 3.3. Finally, the objective function is discussed in section 3.4.

3.1. Cross-modal Feature Representation

Image representation. Benefiting from the development of deep learning in computer vision, different convolution neural networks have been widely used in many tasks to extract visual information for images. To obtain more descriptive information about the visual content for image fragments, we employ a pretrained deep CNN, *e.g*. Faster R-CNN. Specifically, given an image $\boldsymbol{I}$, a CNN detects image regions and extracts a feature vector $\boldsymbol{f}_i$ for each image region $\boldsymbol{r}_i$. We further transform $\boldsymbol{f}_i$ to a d-dimensional vector $\boldsymbol{v_i}$ via a linear projection as follows:

$$\boldsymbol{v}_i = W_v \boldsymbol{f}_i + b_v \tag{1}$$

where W_v and b_v are to-be-learned parameters.

For simplicity, we denote the image representation as $\boldsymbol{V} = \{\boldsymbol{v}_i | i = 1, ..., m, \boldsymbol{v}_i \in \mathbb{R}^d\}$, where m is the number of detected regions in $\boldsymbol{I}$. We further normalize each region feature vector in $\boldsymbol{V}$ as [12].

Text representation. Basically, texts can be represented at either sentence-level or word-level. To enable the fine-grained connection of vision and language, we extract the word-level features for texts, which can be done through a bi-directional GRU as the encoder.

Specifically, for a text $\boldsymbol{S}$ with n words, we first represent each word $\boldsymbol{w}_j$ with a contiguous embedding vector $\boldsymbol{e}_j = \boldsymbol{W_e}\boldsymbol{w}_j, \forall j \in [1, n]$, where $\boldsymbol{W_e}$ is a to-be-learned embedding matrix. Then, to enhance the word-level representation with context information, we employ a bi-directional GRU to summarize information from both forward and backward directions in the text $\boldsymbol{S}$:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}}_j &= \overrightarrow{\boldsymbol{GRU}}(\boldsymbol{e}_j, \overrightarrow{\boldsymbol{h}}_{j-1}); \\ \overleftarrow{\boldsymbol{h}}_j &= \overleftarrow{\boldsymbol{GRU}}(\boldsymbol{e}_j, \overleftarrow{\boldsymbol{h}}_{j+1}) \end{aligned} \tag{2}$$

where $\overrightarrow{\boldsymbol{h}}_j$ and $\overleftarrow{\boldsymbol{h}}_j$ denote hidden states from the forward GRU and the backward GRU, respectively. Then, the representation of the word $\boldsymbol{w}_j$ is defined as $\boldsymbol{t}_j = \frac{\overrightarrow{\boldsymbol{h}}_j + \overleftarrow{\boldsymbol{h}}_j}{2}$.

Eventually, we obtain a word-level feature set for the text $\boldsymbol{S}$, denoted as $\boldsymbol{T} = \{\boldsymbol{t}_j | j = 1, ..., n, \boldsymbol{t}_j \in \mathbb{R}^d\}$, where each $\boldsymbol{t}_j$ encodes the information of the word $\boldsymbol{w}_j$. Note that each $\boldsymbol{t}_j$ shares the same dimensionality as $\boldsymbol{v}_i$ in Eq. 1. We also normalize each word feature vector in $\boldsymbol{T}$ as [12].

3.2. RAM: Recurrent Attention Memory

The recurrent attention memory aims to align fragments in the embedding space by refining the knowledge about previous fragment alignments in a recurrent manner. It can be regarded as a block that takes in two sets of feature points, *i.e.* $\boldsymbol{V}$ and $\boldsymbol{T}$, and estimates the similarity between these two sets via a cross-modal attention unit. A memory distillation unit is used to refine the attention result in order to provide more knowledge for the next alignments. For generalization, we denote the two input sets of features as a query set $\boldsymbol{X} = \{\boldsymbol{x}_i | i \in [1, m'], \boldsymbol{x}_i \in \mathbb{R}^d\}$ and a response set $\boldsymbol{Y} = \{\boldsymbol{y}_j | j \in [1, n'], \boldsymbol{y}_j \in \mathbb{R}^d\}$, where m' and n' are the numbers of feature points in $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Note that $\boldsymbol{X}$ can be either of $\boldsymbol{V}$ and $\boldsymbol{T}$, while $\boldsymbol{Y}$ will be the other.

Cross-modal Attention Unit (CAU). The cross-modal attention unit aims to summarize context information in $\boldsymbol{Y}$ for each feature $\boldsymbol{x}_i$ in $\boldsymbol{X}$. To achieve this goal, we first compute the similarity between each pair $(\boldsymbol{x}_i, \boldsymbol{y}_j)$ using the cosine function:

$$z_{ij} = \frac{\boldsymbol{x}_i^T \boldsymbol{y}_j}{||\boldsymbol{x}_i|| \cdot ||\boldsymbol{y}_j||}, \forall i \in [1, m'], \forall j \in [1, n'] \tag{3}$$

As [12], we further normalize the similarity score z as:

$$\bar{z}_{ij} = \frac{\text{relu}(z_{ij})}{\sqrt{\sum_{i=1}^{m'} \text{relu}(z_{ij})^2}} \tag{4}$$

where $\text{relu}(x) = \max(0, x)$.

Attention is performed over the response set $\boldsymbol{Y}$ given a feature $\boldsymbol{x}_i$ in $\boldsymbol{X}$:

$$\boldsymbol{c}_i^x = \sum_{j=1}^{n'} \alpha_{ij} \boldsymbol{y}_j, \quad s.t. \quad \alpha_{ij} = \frac{exp(\lambda \bar{z}_{ij})}{\sum_{j=1}^{n'} exp(\lambda \bar{z}_{ij})} \tag{5}$$

where λ is the inverse temperature parameter of the softmax function [2] to adjust the smoothness of the attention distribution.

We define $\boldsymbol{C}^x = \{\boldsymbol{c}_i^x | i \in [1, m'], \boldsymbol{c}_i^x \in \mathbb{R}^d\}$ as $\boldsymbol{X}$-grounded alignment features, in which each element captures related semantics shared by each $\boldsymbol{x}_i$ and the whole $\boldsymbol{Y}$.

Memory Distillation Unit (MDU). To refine the alignment knowledge for the next alignment, we adopt a memory distillation unit which updates the query features $\boldsymbol{X}$ by aggregating them with the corresponding $\boldsymbol{X}$-grounded alignment feature $\boldsymbol{C}^x$ dynamically:

$$\boldsymbol{x}_i^* = f(\boldsymbol{x}_i, \boldsymbol{c}_i^x) \tag{6}$$

where $f()$ is a aggregating function. We can define $f()$ with different formulations, such as addition, multilayer perceptron (MLP), attention and so on. Here, we adopt a modified gating mechanism for $f()$:

$$\begin{aligned} \boldsymbol{g}_i &= \text{gate}(W_g[\boldsymbol{x}_i, \boldsymbol{c}_i^x] + b_g) \\ \boldsymbol{o}_i &= \tanh(W_o[\boldsymbol{x}_i, \boldsymbol{c}_i^x] + b_o) \\ \boldsymbol{x}_i^* &= \boldsymbol{g}_i * \boldsymbol{x}_i + (1 - \boldsymbol{g}_i) * \boldsymbol{o}_i \end{aligned} \tag{7}$$

where W_g, W_o, b_g, b_o are to-be-learned parameters. $\boldsymbol{o}_i$ is a fused feature which enhances the interaction between $\boldsymbol{x}_i$ and $\boldsymbol{c}_i^x$. $\boldsymbol{g}_i$ performs as a gate to select the most salient information.

With the gating mechanism, information of the input query can be refined by itself (*i.e.* $\boldsymbol{x}_i$) and the semantic information shared with the response (*i.e.* $\boldsymbol{o}_i$). The gate $\boldsymbol{g}_i$ can help to filter trivial information in the query, and enable the representation learning of each query fragment (*i.e.* $\boldsymbol{x}_i$ in $\boldsymbol{X}$) to focus more on its individual shared semantics with $\boldsymbol{Y}$. Besides, the $\boldsymbol{X}$-grounded alignment features $\boldsymbol{C}^x$ summarize the context information of $\boldsymbol{Y}$ with regard to each fragment in $\boldsymbol{X}$. And in the next matching step, such context information will assist to determine the shared semantics with respect to $\boldsymbol{Y}$, forming a recurrent computation process as described in the subsequent section 3.3. Therefore, with the help of $\mathbf{C}^x$, the intra-modality relationships in $\boldsymbol{Y}$ are implicitly involved and re-calibrated during the recurrent process, which would enhance the interaction among cross-modal features and thus benefit the representation learning.

RAM block. We integrate the cross-modal attention unit and the memory distillation unit into a RAM block, formulated as:

$$\boldsymbol{C}^x, \boldsymbol{X}^* = \textbf{RAM}(\boldsymbol{X}, \boldsymbol{Y}) \tag{8}$$

where $\boldsymbol{C}^x$ and $\boldsymbol{X}^*$ are derived by Eq. 5 and 6.

3.3. Iterative Matching with Recurrent Attention Memory

In this section, we describe how to employ the recurrent attention memory introduced above to enable the iterative matching for cross-modal image-text retrieval.

Table 1. Comparison with the state-of-the-art models on Flickr8K. As results of SCAN [12] are not reported on Flickr8K, here we show our experiment results by running codes provided by authors.

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DeViSE [5]	4.8	16.5	27.3	5.9	20.1	29.6	104.2
DVSA [9]	16.5	40.6	54.2	11.8	32.1	44.7	199.9
m-CNN [15]	24.8	53.7	67.1	20.3	47.6	61.7	275.2
SCAN*	52.2	81.0	89.2	38.3	67.8	78.9	407.4
Image-IMRAM	48.5	78.1	85.3	32.0	61.4	73.9	379.2
Text-IMRAM	52.1	81.5	90.1	40.2	69.0	79.2	412.1
Full-IMRAM	**54.7**	**84.2**	**91.0**	**41.0**	**69.2**	**79.9**	**420.0**

Specifically, given an image $\boldsymbol{I}$ and a text $\boldsymbol{S}$, we derive two strategies for iterative matching grounded on $\boldsymbol{I}$ and $\boldsymbol{S}$, respectively, using two independent RAM blocks:

$$\begin{aligned} \boldsymbol{C}_k^v, \boldsymbol{V}_k &= \mathbf{RAM}_v(\boldsymbol{V}_{k-1}, \boldsymbol{T}) \\ \boldsymbol{C}_k^t, \boldsymbol{T}_k &= \mathbf{RAM}_t(\boldsymbol{T}_{k-1}, \boldsymbol{V}) \end{aligned} \tag{9}$$

where $\boldsymbol{V}_k$, $\boldsymbol{T}_k$ indicate the step-wise features of the image $\boldsymbol{I}$ and the text $\boldsymbol{S}$, respectively. And k is the matching step, and $\boldsymbol{V}_0 = \boldsymbol{V}, \boldsymbol{T}_0 = \boldsymbol{T}$.

We iteratively perform **RAM()** for a total of K steps. And at each step k, we can derive a matching score between $\boldsymbol{I}$ and $\boldsymbol{S}$:

$$F_k(\boldsymbol{I}, \boldsymbol{S}) = \frac{1}{m}\sum_{i=1}^{m} F_k(\boldsymbol{r}_i, \boldsymbol{S}) + \frac{1}{n}\sum_{j=1}^{n} F_k(\boldsymbol{I}, \boldsymbol{w}_j) \tag{10}$$

where $F(\boldsymbol{r}_i, \boldsymbol{S})$ and $F(\boldsymbol{I}, \boldsymbol{w}_j)$ are defined as the region-based matching score and the word-based matching score, respectively. They are derived as follows:

$$\begin{aligned} F_k(\boldsymbol{r}_i, \boldsymbol{S}) &= \text{sim}(\boldsymbol{v}_i, \boldsymbol{c}_{ki}^v); \\ F_k(\boldsymbol{I}, \boldsymbol{w}_j) &= \text{sim}(\boldsymbol{c}_{kj}^t, \boldsymbol{t}_j) \end{aligned} \tag{11}$$

where sim() is the cosine function that measures the similarity between two input features as Eq. 3. And $\boldsymbol{v}_i \in \boldsymbol{V}$ corresponds to the region $\boldsymbol{r}_i$. $\boldsymbol{t}_j \in \boldsymbol{T}$ corresponds to the word $\boldsymbol{w}_j$. $\boldsymbol{c}_{ki}^v \in \boldsymbol{C}_k^v$ and $\boldsymbol{c}_{kj}^t \in \boldsymbol{C}_k^t$ are, respectively, the context feature corresponding to the region $\boldsymbol{r}_i$ and the word $\boldsymbol{w}_j$. m and n are the numbers of image regions and text words, respectively.

After K matching steps, we derive the similarity between $\boldsymbol{I}$ and $\boldsymbol{S}$ by summing all matching scores:

$$F(\boldsymbol{I}, \boldsymbol{S}) = \sum_{k=1}^{K} F_k(\boldsymbol{I}, \boldsymbol{S}) \tag{12}$$

3.4. Loss Function

In order to enforce matched image-text pairs to be clustered and unmatched ones to be separated in the embedding spaces, triplet-wise ranking objectives are widely used in previous works [11, 4] to train the model in an end-to-end manner. Following [4], instead of comparing with all negatives, we only consider the *hard* negatives within a mini-batch, *i.e*. the negative that is closest to a training query:

$$\begin{aligned} \mathcal{L} = &\sum_{b=1}^{B}[\Delta - F(I_b, S_b) + F(I_b, S_{b^*})]_+ \\ &+ \sum_{b=1}^{B}[\Delta - F(I_b, S_b) + F(I_{b^*}, S_b)]_+ \end{aligned} \tag{13}$$

where $[x]_+ = max(x, 0)$, and $F(I, S)$ is the semantic similarity between I and S defined by Eq. 12. Images and texts with the same subscript b are matched examples. Hard negatives are indicated by the subscript b^*. Δ is a margin value.

Note that in the loss function, $F(\boldsymbol{I}, \boldsymbol{S})$ consists of $F_k(\boldsymbol{I}, \boldsymbol{S})$ at each matching step (*i.e*. Eq. 12), and thus optimizing the loss function would directly supervise the learning of image-text correspondences at each matching step, which is expected to help the model to yield higher-quality alignment at each step. With the employed triplet-wise ranking objective, the whole model parameters can be optimized in an end-to-end manner, using widely-used optimizers like SGD, etc.

4. Experiment

4.1. Datasets and Evaluation Metric

Three benchmark datasets are used in our experiments, including: (1) **Flickr8K**: contains 8,000 images and provides 5 texts for each image. We adopt its standard splits as [19, 15], using 6,000 images for training, 1,000 images for validation and another 1,000 images for testing. (2) **Flickr30K**: consists of 31,000 images and 158,915 English texts. Each image is annotated with 5 texts. We follow the dataset splits as [12, 4] and use 29,000 images for training, 1,000 images for validation, and the remaining 1,000 images for testing. (3) **MS COCO**: is a large-scale image description dataset containing about 123,287 images with at least 5 texts for each. As previous works [12, 4], we use 113,287 images to train all models, 5,000 images for validation and another 5,000 images for testing. Results on MS

Table 2. Comparison with state-of-the-art models on Flickr30K.

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DPC [28]	55.6	81.9	89.5	39.1	69.2	80.9	416.2
SCO [7]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
SCAN* [12]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN* [14]	71.3	90.6	96.0	**54.7**	**81.8**	**88.2**	482.6
Image-IMRAM	67.0	90.5	95.6	51.2	78.2	85.5	468.0
Text-IMRAM	68.8	91.6	96.0	53.0	79.0	87.1	475.5
Full-IMRAM	**74.1**	**93.0**	**96.6**	**53.9**	**79.4**	**87.2**	**484.2**

Table 3. Comparison with state-of-the-art models on MS COCO.

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
1K							
DPC [28]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
SCO [7]	69.9	92.9	97.5	56.7	87.5	94.8	499.3
SCAN* [12]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
PVSE [23]	69.2	91.6	96.6	55.2	86.5	93.7	492.8
VSRN* [14]	76.2	94.8	98.2	**62.8**	**89.7**	**95.1**	516.8
Image-IMRAM	76.1	95.3	98.2	61.0	88.6	94.5	513.7
Text-IMRAM	74.0	95.6	98.4	60.6	88.9	94.6	512.1
Full-IMRAM	**76.7**	**95.6**	**98.5**	**61.7**	**89.1**	**95.0**	**516.6**
5K							
DPC [28]	41.2	70.5	81.1	25.3	53.4	66.4	337.9
SCO [7]	42.8	72.3	83.0	33.1	62.9	75.5	369.6
SCAN* [12]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
PVSE [23]	45.2	74.3	84.5	32.4	63.0	75.0	374.4
VSRN* [14]	53.0	81.1	89.4	**40.5**	**70.6**	**81.1**	415.7
Image-IMRAM	53.2	82.5	90.4	38.9	68.5	79.2	412.7
Text-IMRAM	52.0	81.8	90.1	38.6	68.1	79.1	409.7
Full-IMRAM	**53.7**	**83.2**	**91.0**	**39.7**	**69.1**	**79.8**	**416.5**

Figure 2. Difference between our KWAI-AD dataset and standard datasets, *e.g*. MS COCO.

COCO are reported by averaging over 5 folds of 1K test images and testing on the full 5K test images as [12].

To further validate the effectiveness of our method in practical scenarios, we build a new dataset, named **KWAI-AD**. We collect 81,653 image-text pairs from a real-world business advertisement platform, and we randomly sample 79,653 image-text pairs for training, 1,000 for validation and the remaining 1,000 for testing. The uniqueness of our dataset is that the provided texts are not detailed textual descriptions of the content in the corresponding images, but maintain weakly associations with them, conveying strong affective semantics instead of factual semantics (seeing Figure 2). And thus our dataset is more challenging than conventional datasets. However, it is of great importance in the practical business scenario. Learning subtle links of advertisement images with related well-designed titles could not only enrich the understanding of vision and language but also benefit the development of recommender systems and social networks.

Evaluation Metric. To compare our proposed method with the state-of-the-art methods, we adopt the same evaluation metrics in all datasets as [16, 12, 4]. Namely, we adopt Recall at K (R@K) to measure the performance of bi-directional retrieval tasks, *i.e*. retrieving texts given an image query (Text Retrieval) and retrieving images given a text query (Image Retrieval). We report R@1, R@5, and R@10 for all datasets as in [12]. And to well reveal the effectiveness of the proposed method, we also report an extra metric "R@sum", which is the summation of all evaluation metrics as [6].

4.2. Implementation Details

To systematically validate the effectiveness of the proposed IMRAM, we experiment with three of its variants: (1) Image-IMRAM only adopts the RAM block grounded on images (*i.e*. only using the first term in Eq. 10); (2) Text-IMRAM only adopts the RAM block grounded on texts (*i.e*. only using the first term in Eq. 10); (3) Full-IMRAM. All models are implemented by Pytorch v1.0. In all datasets,

for each word in texts, the word embedding is initialized by random weights with a dimensionality of 300. We use a bi-directional GRU with one layer and set its hidden state (*i.e.* $\overrightarrow{\boldsymbol{h}}_j$ and $\overleftarrow{\boldsymbol{h}}_j$ in Eq. 2) dimensionality as 1,024. The dimensionality of each region feature (*i.e.* $\boldsymbol{v}_i$ in $\boldsymbol{V}$) and and each word feature (*i.e.* $\boldsymbol{t}_j$ in $\boldsymbol{T}$) is set as 1,024. On three benchmark datasets, we use Faster R-CNN pre-trained on Visual Genome to extract 36 region features for each image. For our KWAI-AD dataset, we simply use Inception v3 [24] to extract 64 features for each image.

4.3. Results on Three Benchmark Datasets

We compare our proposed IMRAM with published state-of-the-art models in the three benchmark datasets[2]. We directly cite the best-reported results from respective papers when available. And for our proposed models, we perform 3 steps of iterative matching by default.

Results. Comparison results are shown in Table 1, Table 2 and Table 3 for Flickr8K, Flickr30K and MS COCO, respectively. '*' indicates the performance of an ensemble model. '-' means unreported results. We can see that our proposed IMRAM can consistently achieve performance improvements in terms of all metrics, compared to the state-of-the-art models.

Specifically, our Full-IMRAM can significantly outperform the previous best model, *i.e.* SCAN* [12], by a large margin of 12.6%, 19.2%, 8.7% and 5.6% in terms of the overall performance R@sum in Flickr8K, Flickr30K, MS COCO (1K) and MS COCO (5K), respectively. And among recall metrics for the text retrieval task, our Full-IMRAM can obtain a maximal performance improvement of 3.2% (R@5 in Flickr8K), 6.7% (R@1 in Flickr30K), 4.0% (R@1 in MS COCO (1K)) and 3.3% (R@1 in MS COCO (5K)), respectively. As for the image retrieval task, the maximal improvements are 2.7% (R@1 in Flickr8K), 5.3% (R@1 in Flickr30K), 2.9% (R@1 in MS COCO (1K)) and 1.1% (R@1 in MS COCO (5K)), respectively. These results well demonstrate that the proposed method exhibits great effectiveness for cross-modal image-text retrieval. Besides, our models can consistently achieve state-of-the-art performance not only in small datasets, *i.e.* Flickr8K and Flickr30K, but also in the large-scale dataset, *i.e.* MS COCO, which well demonstrates its robustness.

4.4. Model Analysis

Effect of the total steps of matching, K. For all three variants of IMRAM, we gradually increase K from 1 to 3 to train and evaluate them on the benchmark datasets. Due to the limited space, we only report results on MS COCO (5K test) in Table 4. We can observe that for all variants, $K = 2$ and $K = 3$ can consistently achieve better performance than $K = 1$. And $K = 3$ performs better or comparatively, compared with $K = 2$. This observation well demonstrates that the iterative matching scheme effectively improves model performance. Besides, our Full-IMRAM consistently outperforms Image-IMRAM and Text-IMRAM for different values of K.

[2]We omit models that require additional data augmentation [18, 22, 16, 13, 1, 8].

Table 4. The effect of the total steps of matching, K, on variants of IMRAM in MS COCO (5K).

Model	K	Text Retrieval		Image Retrieval	
		R@1	R@10	R@1	R@10
Image -IMRAM	1	40.8	85.7	34.6	76.2
	2	51.5	89.5	37.7	78.3
	3	**53.2**	**90.4**	**38.9**	**79.2**
Text -IMRAM	1	46.2	87.0	34.4	75.9
	2	50.4	89.2	37.4	78.3
	3	**51.4**	**89.9**	**39.2**	**79.2**
Full -IMRAM	1	49.7	88.9	35.4	76.7
	2	53.1	90.2	39.1	79.5
	3	**53.7**	**91.0**	**39.7**	**79.8**

Table 5. The effect of the aggregating function in the proposed memory distillation unit of Text-IMRAM ($K = 3$) in Flickr30K.

Memory	Text Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
add	64.5	95.1	49.2	84.9
mlp	66.6	**96.4**	52.8	86.2
att	66.1	95.5	52.1	86.2
gate	66.2	**96.4**	52.5	86.1
ours	**68.8**	96.0	**53.0**	**87.1**

Table 6. Statistical results of salient semantics at each matching step, k, in Text-IMRAM ($K = 3$) in MS COCO.

k	nouns (%)	verbs (%)	adjectives (%)
1	99.0	32.0	35.3
2	99.0	38.8	37.9
3	**99.0**	**40.2**	**39.1**

Effect of the memory distillation unit. The aggregation function $f(\boldsymbol{x}, \boldsymbol{y})$ in Eq. 6 is essential for the proposed iterative matching process. We enumerate some basic aggregation functions and compare them with ours: (1) add: $\boldsymbol{x} + \boldsymbol{y}$; (2) mlp: $\boldsymbol{x} + \tanh(W\boldsymbol{y} + b)$; (3) att: $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$ where α is a real-valued number parameterized by $\boldsymbol{x}$ and $\boldsymbol{y}$; (4) gate: $\boldsymbol{\beta}\boldsymbol{x} + (1-\boldsymbol{\beta})\boldsymbol{y}$ where $\boldsymbol{\beta}$ is a real-valued vector parameterized by $\boldsymbol{x}$ and $\boldsymbol{y}$. We conduct the analysis with Text-IMRAM ($K = 3$) in Flickr30K in Table 5. We can observe that the aggregation function we use (*i.e.* Eq. 7) achieves substantially better performance than baseline functions.

4.5. Qualitative Analysis

We intend to explore more insights for the effectiveness of our models here. For the convenience of the explanation, we mainly analyze semantic concepts from the view of language, instead of from the view of vision, *i.e.* we treat each

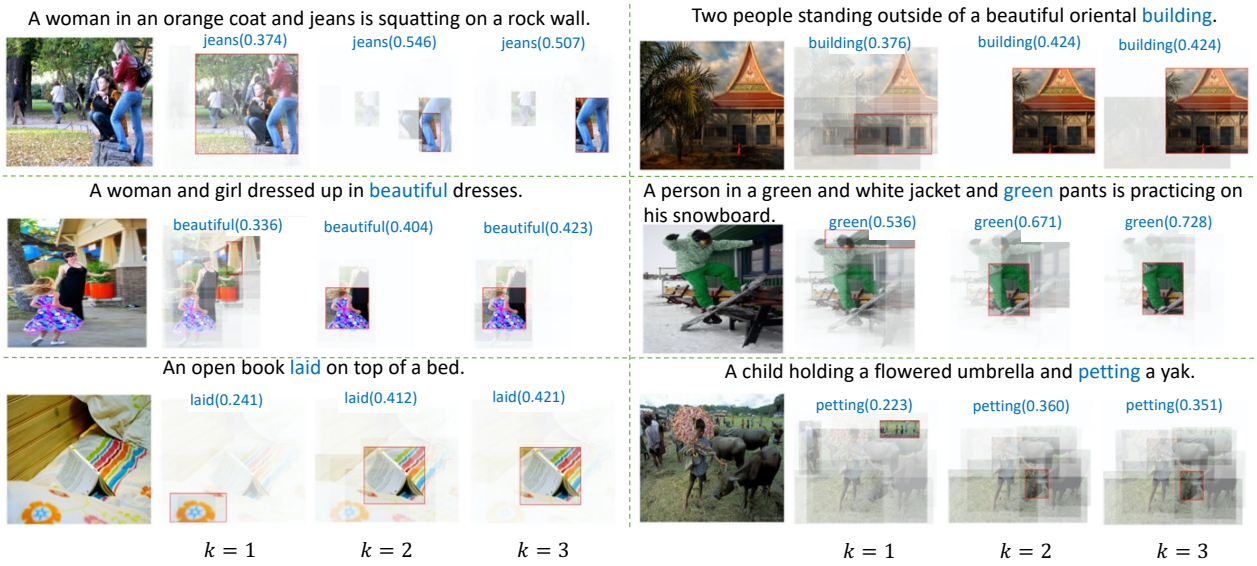

Figure 3. Visualization of attention at each matching step in Text-IMRAM. Corresponding matched words are in blue, followed by the matching similarity.

word in the text as one semantic concept. Therefore, we conduct the qualitative analysis on Text-IMRAM.

We first visualize the attention map at each matching step in Text-IMRAM ($K = 3$) corresponding to different semantic concepts in Figure 3. We can see that the attention is refined and gradually focuses on the matched regions.

To quantitatively analyze the alignment of semantic concepts, we first define a semantic concept in Text-IMRAM as a salient one at the matching step k as follows: 1) Given an image-text pair, at the matching step k, we derive the word-based matching score by Eq. 11 for each word with respect to the image, and derive the image-text matching score by averaging all the word-based scores (see Eq. 10). 2) A semantic concept is salient if its corresponding word-based score is greater than the image-text score. For a set of image-text pairs randomly sampled from the testing set, we can compute the percentage of such salient semantic concepts for each model at different matching steps.

Then we analyze the change of the salient semantic concepts captured at different matching steps in Text-IMRAM ($K = 3$). Statistical results are shown in Table 6. We can see that at the 1st matching step, nouns are easy to be recognized and dominant to help to match. While during the subsequent matching steps, contributions of verbs and adjectives increase.

4.6. Results on the Newly-Collected Ads Dataset

We evaluate our proposed IMRAM on our KWAI-AD dataset. We compare our models with the state-of-the-art SCAN models in [12]. Comparison results are shown in Table 7. We can see that the overall performance on this dataset is greatly lower than those on benchmark datasets,

Table 7. Results on the Ads dataset.

Method	Text Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
i-t AVG [12]	7.4	21.1	2.1	9.3
Image-IMRAM	**10.7**	**25.1**	**3.4**	**16.8**
t-i AVG [12]	6.8	20.8	2.0	9.9
Text-IMRAM	**8.4**	**21.5**	**2.3**	**15.9**
i-t + t-i [12]	7.3	22.5	2.7	11.5
Full-IMRAM	**10.2**	**27.7**	**3.4**	**21.7**

which indicates the challenges of cross-modal retrieval in real-world business advertisement scenarios. Results also show that our models can obtain substantial improvements over compared models, which demonstrates the effectiveness of the proposed method in this dataset.

5. Conclusion

In this paper, we propose an Iterative Matching method with a Recurrent Attention Memory network (IMRAM) for cross-modal image-text retrieval to handle the complexity of semantics. Our IMRAM can explore the correspondence between images and texts in a progressive manner with two features: (1) an iterative matching scheme with a cross-modal attention unit to align fragments from different modalities; (2) a memory distillation unit to refine alignments knowledge from early steps to later ones. We validate our models on three benchmarks (*i.e.* Flickr8K, Flickr30K and MS COCO) as well as a new dataset (*i.e.* KWAI-AD) for practical business advertisement scenarios. Experiment results on all datasets show that our IMRAM outperforms compared methods consistently and achieves state-of-the-art performance.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740, 2019.

[2] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[5] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[6] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

[7] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[8] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5754–5763, 2019.

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[10] Andrej Karpathy, Armand Joulin, and Fei Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *International Conference on Neural Information Processing Systems*, 2014.

[11] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[12] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[13] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, 2019.

[14] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[15] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[16] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E. Papalexakis, and Amit K. Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 1856–1864, 2018.

[17] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

[18] Duy-Kien Nguyen and Takayuki Okatani. Multi-task learning of hierarchical vision-language representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1899–1907, 2017.

[20] Leonid Perlovsky. Language and cognition interaction neural mechanisms. *Computational Intelligence and Neuroscience*, 2011, 2011.

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237, 2018.

[22] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. Knowledge aware semantic concept expansion for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5182–5189, 7 2019.

[23] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[25] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1556–1566, Beijing, China, july 2015.

[26] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[27] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

[28] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.