

# ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching

Yan Huang<sup>1,4</sup>      Liang Wang<sup>1,2,3,4</sup>

<sup>1</sup>Center for Research on Intelligent Perception and Computing (CRIPAC)  
National Laboratory of Pattern Recognition (NLPR)

<sup>2</sup>Center for Excellence in Brain Science and Intelligence Technology (CEBSIT)  
Institute of Automation, Chinese Academy of Sciences (CASIA)

<sup>3</sup>University of Chinese Academy of Sciences (UCAS)

<sup>4</sup>Chinese Academy of Sciences Artificial Intelligence Research (CAS-AIR)

Fyhuang, wangliang@nlpr.ia.ac.cn

## Abstract

*Image and sentence matching has drawn much attention recently, but due to the lack of sufficient pairwise data for training, most previous methods still cannot well associate those challenging pairs of images and sentences containing rarely appeared regions and words, i.e., few-shot content. In this work, we study this challenging scenario as few-shot image and sentence matching, and accordingly propose an Aligned Cross-Modal Memory (ACMM) model to memorize the rarely appeared content. Given a pair of image and sentence, the model first includes an aligned memory controller network to produce two sets of semantically-comparable interface vectors through cross-modal alignment. Then the interface vectors are used by modality-specific read and update operations to alternatively interact with shared memory items. The memory items persistently memorize cross-modal shared semantic representations, which can be addressed out to better enhance the representation of few-shot content. We apply the proposed model to both conventional and few-shot image and sentence matching tasks, and demonstrate its effectiveness by achieving the state-of-the-art performance on two benchmark datasets.*

## 1. Introduction

With the rapid growth of multimodal data, image and sentence matching has drawn much attention recently. This technique has been widely applied to the task of cross-modal retrieval, i.e., given an image query to retrieve certain sentences with similar content, and vice-versa given a sentence query. The challenge of image and sentence matching lies in how to accurately measure the cross-modal similarity between images and sentences. As shown in Figure 1, the

Figure 1. Averaged recall rate vs. minimum appearing frequency (best viewed in colors).

global similarity of a given pair of image and sentence usually depends on multiple local similarities between regions (marked by rectangle) and words (marked by bold). Most existing models [13, 21, 25, 5] measure these local similarities by training on limited pairs of image and sentence, so they statistically tend to better associate partial regions and words with higher appearing frequencies (marked by blue) during training. While for the rarely appeared region and word (marked by red), i.e., few-shot content, these models cannot well recognize or associate them.

In Figure 1, we also illustrate the performance of cross-modal retrieval by three state-of-the-art models: VSE++ [5], SCO [13] and SCAN [21] on selected test sets containing k-shot content. For each test set, we select certain pairs of image and sentence that have at least one word<sup>1</sup> whose appearing frequency is less than k. We can observe that all the models can achieve well performance when there is no few-shot ( $k = 100$ ) content. But when  $k = 10$ , their perfor-

<sup>1</sup>Only considering nouns, verbs and adjectives.

mance all drops heavily by a large gap of 6%–7%. It indicates that these methods cannot be well generalized to deal with few-shot content. Additionally, such a few-shot issue could be a bottleneck for further performance improvement, especially in practical applications where the data content could be much more imbalanced.

To alleviate this problem, in this paper, we focus on the challenging scenario as *few-shot image and sentence matching*. Different from conventional image and sentence matching, we especially study how to better match those pairwise images and sentences having rarely appeared regions and words. To the best of our knowledge, this scenario has been seldom identified or investigated. Although the problem of few-shot matching for image and word [6, 26] has been previously studied, directly adapting them to our task is infeasible. Rather than multiple separated few-shot objects and nouns, image and sentence matching usually have much more complex few-shot content, *i.e.*, objects, actions and properties in images, and nouns, verbs and adjectives in sentences. In addition, we deal with sentences rather than words which simultaneously include both few-shot words and common ones. So how to suitably model their relation and exploit it to better understand few-shot content is another issue.

To deal with these problems, we propose an Aligned Cross-Modal Memory (ACMM) model which can represent, align and memorize few-shot content in a successive manner. To well describe those rarely appeared regions and words, the model first resorts to pretrained models on external resources to obtain generic representations. Then to reduce their cross-modal heterogeneity and predict two sets of semantically-comparable interface vectors, ACMM includes a cross-modal graph convolutional network as its memory controller, which aligns region representations to word ones. Based on the interface vectors, modality-specific read and update operations are designed to alternatively interact with cross-modal shared memory items. The memory items are persistently updated across minibatches during the whole training, whose stored cross-modal shared semantic representations can be used for enhancing the representation of few-shot content. We apply the proposed model to both conventional and few-shot image and sentence matching tasks on two publicly available datasets, and demonstrate its effectiveness by achieving the state-of-the-art performance.

## 2. Related Work

### 2.1. Image and Sentence Matching

Frome *et al.* [6] propose the visual-semantic embedding framework to associate pairs of images and words. Based on this framework, Kiros *et al.* [17] later extend it for image and sentence matching. Faghri *et al.* [5] penalize the mod-

el based on the hardest negative examples in the objective function and achieve better results. In addition to the global similarity measurement, Karpathy *et al.* [15] make the attempt to learn local similarities from fragments of images and sentences. Lee *et al.* [21] use the stacked cross-modal attention to softly align regions and words. Huang *et al.* [13] first extract semantic concepts and then organize them in a semantic order which can greatly improve the performance. Different from them, we focus on the rarely studied image and sentence matching with few-shot content.

### 2.2. Neural Memory Modeling

Graves *et al.* [8] propose neural Turing machines and later extend it to a differentiable neural computer [9], in which neural networks can interact with external memory. Sukhbaatar *et al.* [38] develop memory networks which can reason with a long-term memory module via read and write. Based on the similar framework, Weston *et al.* [47] design end-to-end memory networks which require less supervision during training. Xiong *et al.* [49] improve the memory as dynamic memory networks. Different from these unimodal memory models, we propose a cross-modal shared memory which can alternatively interact with multiple data modalities. Although other work [41, 27, 37] also extend memory networks to multimodal settings, most of them are episodic memory networks that are wiped during each minibatch. While our model persistently memorizes semantic representations during the whole training procedure, to better deal with the few-shot content.

### 2.3. Few-Shot Learning

Conventional few-shot learning [34, 48, 45] usually focuses on single-label classification. Other researchers [20, 7] further study the problem in the context of multi-label classification. Hendricks *et al.* [2, 40] propose the task of few-shot image captioning, which can be regarded as sentence classification. In addition to the few-shot classification, there are also many work focusing on few-shot matching. Socher *et al.* [36] and Frome *et al.* [6] use visual-semantic matching frameworks to recognize unseen objects in images. Long *et al.* [26] study the few-shot problem in the image-attributes matching task. Rather than single or multiple words, here we aim to deal with the few-shot matching for sentences, which include not only multiple few-shot words but also other common ones, as well their relation. The most related work is [11], which initially studies this few-shot matching problem by adaptively fusing of multiple models.

## 3. Aligned Cross-Modal Memory

We illustrate our proposed Aligned Cross-Modal Memory (ACMM) for image and sentence matching in Figure 2.

Figure 2. The proposed Aligned Cross-Modal Memory (ACMM) for few-shot image and sentence matching.

To associate a given pair of image and sentence with few-shot content, the proposed ACMM includes three key steps: 1) generic representation extraction for regions and words based on large-scale external resources, 2) cross-modal graph convolutional network as aligned memory controller network to generate semantically-comparable interface vectors, and 3) modality-specific read and update operations for persistent memory items to memorize cross-modal shared semantic representations. We will present the corresponding details in the following.

### 3.1. Generic Representation Extraction

As shown in Figure 2 (a), for a pair of image and sentence, how to accurately detect and represent their regions and words, especially those few-shot ones (marked by red), is the foundation for cross-modal association. But because the number of pairwise data is quite limited, we cannot directly learn the desired representations from scratch.

So we attempt to leverage large-scale external resources, and regard already pretrained models on them as generic representation extractors for all the regions and words. In particular, we choose images from the Visual Genome dataset [19] and texts from [wikipedia.org](http://wikipedia.org) as our multimodal external resources. Both of them have been widely demonstrated to be useful in various tasks [6, 1, 40]. Although some few-shot content in regions might be not included in the Visual Genome dataset, the dataset is diverse enough and its pre-defined attributes [1] can comprehensively describe them.

Then, we use the faster-RCNN [35, 1] and Skip-Gram [30, 6] pretrained on these external resources to extract generic representations for regions and words, respectively. Given an image, the faster-RCNN detects  $l$  regions with high probabilities of containing objects, actions or properties, and outputs  $l$  corresponding  $F$ -dimensional representation vectors from the last fully-connected layer, *i.e.*,  $\{\mathbf{g}_i | \mathbf{g}_i \in \mathbb{R}^F, i=1, \dots, l\}$ . While given a sentence, the Skip-Gram encodes all the included words into  $E$ -dimensional

representation vectors, *i.e.*,  $\{\mathbf{w}_j | \mathbf{w}_j \in \mathbb{R}^E, j=1, \dots, J\}$ , where  $J$  is the length of the sentence. Note that the use of faster-RCNN and Skip-Gram for generic representation extraction might be not optimal, but we empirically find they can already achieve satisfactory performance.

### 3.2. Aligned Memory Controller Network

After obtaining the generic representations of regions and words, we need a memory controller network to generate modality-specific interface vectors to connect with shared memory items. But the generic representations are intrinsically cross-modal heterogeneous, so their directly generated interface vectors tend to be semantically-incomparable. Thus it is very difficult for the memory to recognize and store the desired shared semantic information from them. To handle this issue, we propose an aligned memory controller network based on a cross-modal Graph Convolutional Network (cm-GCN), which explicitly performs cross-modal alignment between the representations of regions and words.

**Semantic Relation Modeling.** We first model the semantic relation among regions and words, respectively, which aims to exploit the potential clues between few-shot content and common one. In particular for words, considering that they are naturally organized in the sequential order in the sentence, we use a bidirectional Gated Recurrent Unit (GRU) network [3] to model their sequential dependency relation, as shown in Figure 2 (b). We sequentially feed the representations of all the words into the bidirectional GRU and regard the corresponding hidden states as their new representations, *i.e.*,  $\{\mathbf{s}_j | \mathbf{s}_j \in \mathbb{R}^H, j=1, \dots, J\}$ , abbreviated as  $\mathbf{S} \in \mathbb{R}^{J \times H}$ . While for regions, we model their relation based on their appearance similarity using a conventional Graph Convolution Network (GCN) [44]. In particular, we first measure the appearance similarity between each pairwise regions to build a similarity graph, in which pairs of appearance-similar regions will have edges with high scores. Based on the graph, we can perform graph convolu-

tion on region representations to obtain new representations, *i.e.*,  $\mathbf{a}_i | \mathbf{a}_i \in \mathbb{R}^F$ ,  $i=1, \dots, J$ , abbreviated as  $\mathbf{A} \in \mathbb{R}^{J \times F}$ . Considering that both GRU and GCN are widely used models, here we omit their detailed formulations for simplicity.

**Cross-modal Alignment.** The unimodal graph convolution above can be viewed as performing a transformation from the original region space to another one. During this procedure, the number of regions remains unchanged, and each region is aligned to itself by considering the contributions from others. Inspired by this, the desired cross-modal alignment can also be formulated as graph convolution but in a cross-modal setting, which performs a cross-modal transformation from region to word spaces. The major difference is that the number of regions does not equal to the number of words.

To implement this, we first construct a cross-modal similarity graph by measuring the cross-modal similarity between each pairwise region and word with two modality-specific mappings. The size of obtained similarity matrix is not squared so that the number of aligned regions will be equivalent to the number of words. The detailed formulations are:

$$g(\mathbf{s}_j, \mathbf{a}_i) = (\mathbf{s}_j)^T (\mathbf{a}_i), \mathbf{G}_{ji} = \frac{e^{g(\mathbf{s}_j, \mathbf{a}_i)}}{\sum_i e^{g(\mathbf{s}_j, \mathbf{a}_i)}}, \mathbf{V} = \mathbf{GAW}$$

where  $(\mathbf{s}_j) = \mathbf{P}\mathbf{s}_j$ ,  $\mathbf{P} \in \mathbb{R}^{H \times H}$ , and  $(\mathbf{a}_i) = \mathbf{Q}\mathbf{a}_i$ ,  $\mathbf{Q} \in \mathbb{R}^{H \times F}$  denote two modality-specific mappings for cross-modal similarity measurement,  $\mathbf{G} \in \mathbb{R}^{J \times J}$  is the normalized cross-modal similarity matrix,  $\mathbf{W} \in \mathbb{R}^{F \times H}$  is the weight matrix, and  $\mathbf{V} \in \mathbb{R}^{J \times H}$  is the aligned region representations.

**Interface Vectors.** Note that  $\mathbf{V}$  and the word representations  $\mathbf{S}$  not only have the same size, but also are semantically aligned. For the  $j$ -th row in  $\mathbf{V}$ , denoted as  $\mathbf{v}_j$ , it is an aggregated representation weighted by cross-modal similarities between the  $j$ -th word and all the regions. Therefore,  $\mathbf{v}_j$  can be viewed as a visual representation of the  $j$ -th word, sharing the same semantic meaning with the word representation  $\mathbf{s}_j$ . Based on the aligned representations, we can obtain two sets of semantically-comparable interface vectors:

$$\begin{aligned} \mathbf{k}^{Vr}, \mathbf{v}^r, \mathbf{k}^{Vw}, \mathbf{v}^w, \mathbf{e}^V, \mathbf{u}^V &= \mathbf{t}^V(\mathbf{v}_j), \\ \mathbf{k}^{Sr}, \mathbf{s}^r, \mathbf{k}^{Sw}, \mathbf{s}^w, \mathbf{e}^S, \mathbf{u}^S &= \mathbf{t}^S(\mathbf{s}_j) \end{aligned}$$

where  $\mathbf{t}^V(\cdot)$  and  $\mathbf{t}^S(\cdot)$  are two linear mappings for region and word, respectively. For the region,  $\mathbf{k}^{Vr} \in \mathbb{R}^W$ ,  $\mathbf{v}^r = \text{oneplus}(\mathbf{v}^r) \in [1, \dots]$ ,  $\mathbf{k}^{Vw} \in \mathbb{R}^W$ ,  $\mathbf{v}^w = \text{oneplus}(\mathbf{v}^w) \in [1, \dots]$ ,  $\mathbf{e}^V = \text{sigmoid}(\mathbf{e}^V) \in [0, 1]^W$ , and  $\mathbf{u}^V \in \mathbb{R}^W$  are its memory read key, read strength, write key, write strength, erase vector, and write vector, respectively. They are all used to interact with memory items, and the corresponding details will be explained in the following.

### 3.3 Memory Read and Update

Based on the two sets of interface vectors, we design shared memory items represented as a matrix  $\mathbf{M} \in \mathbb{R}^{N \times W}$  to store cross-modal shared semantic representations. As shown in Figure 2 (c), each memory item  $\mathbf{M}_i \in \mathbb{R}^W$  could be alternatively updated by modality-specific interface vectors with similar semantic meanings, as well as read out to enhance previously obtained generic representations.

**Memory Read.** We use a content-based addressing mechanism to determine to read which memory items:

$$(\mathbf{k}, \mathbf{M}_i) = \frac{e^{s(\mathbf{k}, \mathbf{M}_i)}}{\sum_i e^{s(\mathbf{k}, \mathbf{M}_i)}}, s(\mathbf{k}, \mathbf{M}_i) = \frac{\mathbf{k} \cdot \mathbf{M}_i}{|\mathbf{k}| |\mathbf{M}_i|}$$

where  $\mathbf{k}$  is read key,  $\mathbf{v}$  is read strength, and  $s(\cdot, \cdot)$  measures the cosine similarity. The read weight  $(\mathbf{k}, \mathbf{M}_i) \in [0, 1]$  defines a normalized weight over the  $i$ -th memory item.

Then we can read memory by alternatively regarding the obtained read keys of region and word as queries:

$$\begin{aligned} \mathbf{r}^V &= \sum_i w_i^{Vr} \mathbf{M}_i, w_i^{Vr} = (\mathbf{k}^{Vr}, \mathbf{M}_i, \mathbf{v}^r) \\ \mathbf{r}^S &= \sum_i w_i^{Sr} \mathbf{M}_i, w_i^{Sr} = (\mathbf{k}^{Sr}, \mathbf{M}_i, \mathbf{s}^r) \end{aligned}$$

where  $w_i^{Vr}$  and  $w_i^{Sr}$  are two read weights for region and word, respectively.  $\mathbf{r}^V \in \mathbb{R}^W$  and  $\mathbf{r}^S \in \mathbb{R}^W$  are two read vectors which can be regarded as memory-enhanced representations of region and word, respectively.

**Memory Update.** Memory update includes how we write and delete the desired shared semantic representations. To determine to write which memory items, we first compute the cross-modal write weights by comparing write keys with memory items through content-based addressing:

$$w_i^{Vw} = (\mathbf{k}^{Vw}, \mathbf{M}_i, \mathbf{v}^w), w_i^{Sw} = (\mathbf{k}^{Sw}, \mathbf{M}_i, \mathbf{s}^w)$$

Note that without the cross-modal pre-alignment, the two write keys are likely to be semantic-incomparable. So we cannot guarantee they can write into similar memory items at nearby locations. Thus the shared semantic representations cannot be uncovered or stored here. After obtaining the write weights, we can selectively update memory items by: 1) adding write vectors  $\mathbf{u}^V$  and  $\mathbf{u}^S$ , *i.e.*, new semantic representations, and 2) deleting old memory gated by erase vectors  $\mathbf{e}^V$  and  $\mathbf{e}^S$ , *i.e.*, how much memory to be deleted:

$$\begin{aligned} \mathbf{M}_i &= \mathbf{M}_i \odot (1 - w_i^{Vw} \mathbf{e}^V) + w_i^{Vw} \mathbf{u}^V, \\ \mathbf{M}_i &= \mathbf{M}_i \odot (1 - w_i^{Sw} \mathbf{e}^S) + w_i^{Sw} \mathbf{u}^S \end{aligned}$$

where  $\odot$  is element-wise multiplication. The memory first updates its memory items with the extracted information from the region and then from the word. In fact, the update order can be alternative and does not show a significant impact on the final performance.

**Discussion.** Our cross-modal memory is initially inspired by [9], but different from them, ours is implemented in a cross-modal way. It focuses on alternative interaction between shared memory items and different data modalities, and especially designs the corresponding aligned controller network. We could alternatively use two sets of modality-specific memory items to separately process region and word, respectively. But this strategy cannot well exploit the homogeneity and complementarity of region and word, and thus tends to degenerate the performance as shown in Section 4.3.

In addition, our memory is persistent during the whole training process, *i.e.*, we do not wipe the learned memory for each minibatch as other memory models [8, 9, 47, 49, 41], with the goal to memorize rarely appeared content. We also do not include the mechanism of dynamic memory allocation, since we experimentally find it will slightly degenerate the performance. It might because this operation automatically removes some rarely accessed but useful memory items associated with few-shot content.

### 3.4 Model Learning

After obtaining memory enhanced representations for all regions and words  $\mathbf{r}_j^V | \mathbf{r}_j^S \quad \mathbf{R}^H \quad j=1, \dots, J$  and  $\mathbf{r}_j^S | \mathbf{r}_j^V \quad \mathbf{R}^H \quad j=1, \dots, J$ . We next perform cross-modal association analysis by first defining the global similarity score of image and sentence as a combination of two averaged cosine similarities:

$$s = \lambda \sum_j s(\mathbf{r}_j^V, \mathbf{r}_j^S)/J + (1-\lambda) \sum_j s(\mathbf{v}_j, \mathbf{s}_j)/J \quad (1)$$

where  $\lambda$  is a balancing parameter, and the two items measure two-stage similarities after and before memory, respectively. When  $\lambda=0$ , it means the model has to memorize from semantically-incomparable regions and words. When  $\lambda>0$ , it indicates that we can pre-align the regions and words. We experimentally find that setting  $\lambda=0.5$  can achieve good performance. Based on the defined similarity score, we use the ranking loss to encourage the similarity score of matched image and sentence to be larger than those of mismatched ones:

$$L = \max_k [0, m - s_{ii} + s_{ik}]_+ + \max_k [0, m - s_{ii} + s_{ki}]_+$$

where  $m$  is a margin parameter,  $[\cdot]_+ = \max(\cdot, 0)$ ,  $s_{ii}$  is the score of the matched  $i$ -th image and  $i$ -th sentence,  $s_{ik}$  is the score of the mismatched  $i$ -th image and  $k$ -th sentence, and vice-versa with  $s_{ki}$ .

## 4. Experimental Results

To demonstrate the effectiveness of the proposed model, we perform experiments in terms of conventional and few-shot image and sentence matching tasks on two publicly available datasets.

### 4.1. Datasets and Protocols

The details of two experimental datasets and their corresponding protocols are as follows. 1) Flickr30k [51] consists of 31783 images collected from the Flickr website. Each image has 5 human annotated sentences. We use the public validation and test splits, which contain 1000 and 1000 images, respectively. 2) MSCOCO [23] consists of 82783 training and 40504 validation images, each of which is associated with 5 sentences. We use the public validation and test splits, with 4000 and 1000 (or 5000) images, respectively. When using 1000 images for test, we perform the validation on 5-fold and report the averaged results.

### 4.2. Implementation Details

The commonly used evaluation criteria for image and sentence matching are “R@1”, “R@5” and “R@10”, *i.e.*, recall rates at the top 1, 5 and 10 results. Following [13], we also use the additional criterion of “mR” by averaging all the recall rates to evaluate the overall performance.

During the generic representation extraction, the number of detected regions in each image is  $I=36$ , the dimension of region representation vectors is  $F=2048$ , the number of words  $J$  equals to the length of each sentence, and the dimension of word representation vectors is  $E=300$ . We set the max length for all the sentences as 50, and pad shorter sentences with zero values. The dimension of hidden states in the bidirectional GRU is  $H=1024$ . The margin parameter is empirically set as  $m=0.2$ . The number and dimension of memory items are  $N=128$  and  $W=256$ , respectively. We empirically find that further increasing the memory number results in a convergence of the performance.

During model learning, we use stochastic gradient descent for parameter optimization, with a learning rate of 0.0005 and gradient clipping at 2. The model is iteratively trained for 30 epochs to guarantee its convergence. In each epoch, the model is learnt in a minibatch manner, with a batch size of  $B=128$ . During each minibatch, our memory totally needs  $B \times J \times 2$  times update. To accelerate the computational speed, use the NVIDIA DGX-1 AI Supercomputer.

### 4.3. Ablation Study

To comprehensively verify the effectiveness of the proposed model, we compare its various ablation models as follows. 1) “align” only performs the cross-modal alignment but does not use its following memory items, and “align (w/o relation)” further removes the modeling of relation with the visual GCN and bidirectional GRU. 2) “mem (w/o shared)” and “mem” alternatively uses two and one sets of modality-specific memory items to process unaligned region and word representations, respectively. 3) “align + mem” is our full model that first aligns region representations to word ones, and then enhances both of them with the

Table 1. Conventional image and sentence matching by ablation models on the Flickr30k and MSCOCO (5000 test) datasets.

Method	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
align (w/o relation)	53.2	80.8	90.3	40.6	69.1	78.3	68.7	32.0	61.1	73.6	22.5	48.6	61.8	49.9
align	65.3	90.6	94.7	47.7	76.7	84.1	76.5	43.3	75.0	85.7	32.4	61.4	73.9	62.0
mem (w/o shared)	0.1	0.5	1.0	0.2	0.7	1.2	0.6	0.0	0.1	0.2	0.0	0.1	0.2	0.1
mem	1.3	6.5	12.6	1.0	4.7	8.3	5.7	0.2	1.2	2.4	0.3	1.3	2.6	1.3
align + mem (w/o shared)	64.8	88.5	93.7	42.6	72.3	81.2	73.9	45.1	76.1	86.0	30.0	58.3	71.4	61.2
align + mem	<b>80.0</b>	<b>95.5</b>	<b>98.2</b>	<b>50.2</b>	<b>76.8</b>	<b>84.7</b>	<b>80.9</b>	<b>63.5</b>	<b>88.0</b>	<b>93.6</b>	<b>36.7</b>	<b>65.1</b>	<b>76.7</b>	<b>70.6</b>

Figure 3. Histograms of cosine similarities between cross-modal write weight vectors in unaligned memory and aligned memory, respectively (best viewed in colors).

stored semantic representations in the memory. Due to the space limitation, we put the analysis of other ablation models related to dynamical memory allocation and Skip-Gram initialization in supplementary material. We use the mentioned ablation models to perform the experiment of image and sentence matching, and compare their performance on the Flickr30k and MSCOCO datasets (5000 test) in Table 1. From this table, we can obtain the following conclusions.

**Cross-modal Alignment.** Only performing the cross-modal alignment with relation modeling (as “align”) can already achieve good performance. When using the cross-modal alignment in the aligned controller network, aligned memory (as “align + mem”) can further improve the performance of unaligned memory (as “mem”). To better illustrate this, we compute cosine similarities between pairs of cross-modal write weight vectors ( $\mathbf{w}^{Vw}$  and  $\mathbf{w}^{Sw}$ ), and then draw similarity histograms by both unaligned memory and aligned memory in Figure 3. We can see that most similarities by aligned memory are around 0.8 and much higher than around 0.15 by unaligned memory. It indicates the cross-modal alignment is able to write cross-modal information into similar memory items at nearby locations to store shared semantic representations.

**Shared Memory.** Without the cross-modal alignment, using either modality-specific memory (as “mem (w/o shared)”) or shared memory (as “mem”) can not achieve well performance. But when using the cross-modal alignment, shared memory (as “align + mem”) performs much

Figure 4. Two-dimensional visualization of learned memory items. Few-shot content are marked as red (best viewed in colors).

better than the modality-specific memory (as “align + mem (w/o shared)”). To illustrate what the shared memory actually learns, we reduce the dimensionality of memory vectors with PCA, and show their two-dimensional representations (nodes) in Figure 4. We can see that all the nodes distribute in a divergent shape, in which the right nodes are more compact while the left ones are more scattered. To figure out the semantic meanings of these nodes, we take several representative nodes (with arrows) as queries to retrieve pairwise images and sentences. We find the compact nodes are more likely to represent commonly appeared content, while the scattered ones tend to retrieve images and sentences with few-shot content (marked by red).

#### 4.4 Few-Shot Image and Sentence Matching

In this section, we aim to especially demonstrate the effectiveness of our proposed model on handling pairs of images and sentences containing rarely appeared regions and words. To achieve this goal, we perform a challenging experiment in terms of few-shot image and sentence matching. In particular, we perform the test in a k-shot matching ( $k \in \{0, 5, 10\}$ ) manner. On each dataset, we only select partial pairs of images and sentences from the standard test set to constitute a new k-shot test set, in which each sentence contains at least one word whose appearing frequency in the training set is less than or equals to k. Note the training stage of few-shot image and sentence matching is the same as that in the conventional matching, the only difference is using different test sets.

Table 2. Few-shot image and sentence matching on the Flickr30k and MSCOCO (5000 test) datasets.

k	N	Method	Flickr30k dataset							MSCOCO dataset						
			Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0	204/ 516	VSE++ [5]	48.2	79.2	85.7	31.9	60.3	71.1	62.7	39.2	71.8	82.1	22.9	49.0	62.6	54.6
		SCO [13]	48.8	77.4	85.7	31.4	58.8	71.6	62.3	40.2	71.6	81.3	24.0	49.8	63.8	55.1
		SCAN [21]	54.8	86.3	91.1	35.3	59.8	71.6	66.5	40.6	73.9	85.9	25.6	49.4	60.3	55.9
		GVSE [11]	62.5	86.9	92.3	46.1	73.5	82.4	73.9	47.2	76.6	88.4	31.2	61.2	70.5	62.5
		<b>ACMM</b>	<b>73.8</b>	<b>94.6</b>	<b>98.2</b>	<b>42.2</b>	<b>68.6</b>	<b>78.4</b>	<b>76.0</b>	<b>62.3</b>	<b>86.3</b>	<b>91.2</b>	<b>27.3</b>	<b>52.3</b>	<b>64.0</b>	<b>63.9</b>
1	321/ 754	VSE++ [5]	50.4	78.6	86.9	33.0	59.5	71.7	63.3	40.7	73.1	82.5	24.8	52.1	64.1	56.2
		SCO [13]	50.4	78.6	88.1	33.3	59.8	70.4	63.4	41.8	72.8	82.4	24.3	51.5	64.9	56.2
		SCAN [21]	56.7	86.1	90.9	37.4	59.2	72.3	67.1	42.5	74.6	86.3	26.4	50.8	61.8	57.1
		GVSE [11]	62.3	88.9	92.9	46.4	73.5	83.2	74.5	49.7	77.1	88.4	32.2	63.5	72.4	63.9
		<b>ACMM</b>	<b>73.0</b>	<b>91.3</b>	<b>96.4</b>	<b>40.5</b>	<b>66.7</b>	<b>77.6</b>	<b>74.2</b>	<b>62.8</b>	<b>86.2</b>	<b>91.9</b>	<b>28.0</b>	<b>53.7</b>	<b>66.2</b>	<b>64.8</b>
5	678/ 973	VSE++ [5]	52.1	80.1	88.0	32.0	60.2	72.3	64.1	41.2	72.7	82.2	23.3	50.6	63.0	55.5
		SCO [13]	52.2	80.3	88.6	33.8	60.9	71.5	64.6	40.9	71.8	81.6	25.4	52.8	65.9	56.4
		SCAN [21]	62.2	87.8	93.4	37.0	64.2	74.3	69.8	42.5	74.6	86.1	25.9	50.5	62.4	57.0
		GVSE [11]	63.8	90.3	94.0	45.4	75.2	85.0	75.6	50.2	78.0	88.1	31.6	63.7	73.4	64.2
		<b>ACMM</b>	<b>76.6</b>	<b>93.2</b>	<b>97.6</b>	<b>42.3</b>	<b>68.0</b>	<b>76.8</b>	<b>75.8</b>	<b>62.2</b>	<b>86.8</b>	<b>92.4</b>	<b>28.1</b>	<b>53.7</b>	<b>65.9</b>	<b>64.9</b>

We make comparisons with three recent state-of-the-art methods in terms of VSE++ [5], SCO [13], SCAN [21], and GVSE [11]. For each compared method, we use its reported best model, and perform test on the k-shot test sets. The comparison results are shown in Table 2, in which N denotes the number of rarely appeared words in k-shot test sets on the two datasets. We can see that in the challenging 1-shot matching, our model can greatly outperform all the compared methods, and achieve much better performance than the best compared SCAN by 7.1% and 7.7% (in mR) on the two datasets, respectively. These evidences show that our model can better recognize and associate those rarely appeared regions and words even though they are presented only once during training. Additionally, when N becomes larger as k increases, our model can consistently achieve better performance. This proves its good generalization ability under various conditions.

#### 4.5. Conventional Image and Sentence Matching

Although our model is especially motivated to deal with the few-shot matching problem, it can be naturally applied to the conventional image and sentence matching. We compare our model with recently published methods on the standard test sets of Flickr30k and MSCOCO datasets in Tables 3 and 4. We denote “ACMM” as an ensemble version of our proposed model, which averages two predicted similarity matrices by setting  $\alpha=0.5$  and  $\beta=0.8$  for final evaluation.

From the table we can see that our model outperforms the current state-of-the-art models in all 7 evaluation criteria on both the Flickr30k and MSCOCO datasets. It is mainly because our memory can store useful cross-modal shared semantic representations, and thus better associate those rarely appeared regions and words in the standard test sets. Note that our model shows much larger improvements on the Flickr30k dataset than the MSCOCO dataset. It mainly results from that the fewer training data of Flickr-

Table 4. Conventional image and sentence matching on the MSCOCO (5000 test) dataset.  $\mathbf{\hat{\cdot}}$  indicates ensemble methods.

Method	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [16]	11.8	32.5	45.4	8.9	24.9	36.3	26.6
FV [18]	17.3	39.0	50.2	10.8	28.3	40.1	31.0
VQA [24]	23.5	50.7	63.6	16.7	40.5	53.8	41.5
OEM [39]	23.3	50.5	65.0	18.0	43.6	57.6	43.0
CSE [50]	27.9	57.1	70.4	22.2	50.2	64.4	48.7
DPCNN [52]	41.2	70.5	81.1	25.3	53.4	66.4	56.3
VSE++ [5]	41.3	69.2	81.2	30.3	59.1	72.4	58.9
LIM [10]	42.0	-	84.7	31.7	-	74.6	-
SCO [13]	42.8	72.3	83.0	33.1	62.9	75.5	61.6
SCO++ [14]	45.7	76.0	86.4	36.8	67.0	78.8	65.1
GVSE [11]	49.9	77.4	87.6	38.4	68.5	79.7	66.9
SCAN [21]	50.4	82.2	90.0	38.6	69.3	80.4	68.5
<b>ACMM</b>	63.5	88.0	93.6	36.7	65.1	76.7	70.6
<b>ACMM</b>	<b>66.9</b>	<b>89.6</b>	<b>94.9</b>	<b>39.5</b>	<b>69.6</b>	<b>81.1</b>	<b>73.6</b>

30k cannot guarantee the previous models can well recognize regions and words. But our model can better exploit the auxiliary resources to better describe them. We can see that our model has much larger performance improvements on the task of image annotation than image retrieval. It might be because the image annotation focuses more on how to learn a suitable semantic space for sentences, and the semantic space is usually more discriminative than the visual space learned by image retrieval.

#### 4.6. Error Analysis

Although our proposed model can achieve well performance in both few-shot and conventional image and sentence matching tasks, it still has limitations on generalizing to arbitrary complex content. To explore its capability, we select several representative failure cases by the proposed model in Figure 5, where the numbers in the top left corner are returned rankings (the smaller, the better) of sentence-based image retrieval. We can see that all their rankings

Table 3. Conventional image and sentence matching on the Flickr30k and MSCOCO (1000 test) datasets.  $\dagger$  indicates ensemble methods.

Method	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
m-RNN [29]	35.4	63.8	73.7	22.8	50.7	63.1	51.6	41.0	73.0	83.5	29.0	42.2	77.0	57.6
FV [18]	35.0	62.0	73.8	25.0	52.7	66.0	52.4	39.4	67.9	80.9	25.1	59.8	76.6	58.3
DVSA [16]	22.2	48.2	61.4	15.2	37.7	50.5	39.2	38.4	69.9	80.5	27.4	60.2	74.8	58.5
MNLM [17]	23.0	50.7	62.9	16.8	42.0	56.5	42.0	43.4	75.7	85.8	31.0	66.7	79.9	63.8
m-CNN [28]	33.6	64.1	74.9	26.2	56.3	69.6	54.1	42.8	73.1	84.1	32.6	68.6	82.8	64.0
RNN+FV [22]	34.7	62.7	72.6	26.2	55.1	69.2	53.4	40.8	71.9	83.2	29.6	64.8	80.5	61.8
OEM [39]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
VQA [24]	33.9	62.5	74.5	24.9	52.6	64.8	52.2	50.5	80.1	89.7	37.0	70.9	82.9	68.5
RTP [33]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
DSPE [42]	40.3	68.9	79.9	29.7	60.1	72.1	58.5	50.1	79.7	89.2	39.6	75.2	86.9	70.1
sm-LSTM [12]	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
2WayNet [4]	49.8	67.5	-	36.0	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
CSE [50]	44.6	74.3	83.8	36.9	69.1	79.6	64.7	56.3	84.4	92.2	45.7	81.2	90.6	75.1
RRF [25]	47.6	77.4	87.1	35.4	68.3	79.9	66.0	56.4	85.3	91.5	43.9	78.1	88.6	73.9
DAN [31]	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-
CHAIN-VSE [46]	-	-	-	-	-	-	-	59.4	88.0	94.2	43.5	79.8	90.2	75.9
DPCNN [52]	55.6	81.9	89.5	39.1	69.2	80.9	69.4	65.6	89.8	95.5	47.1	79.9	90.0	78.0
VSE++ [5]	52.9	79.1	87.2	39.6	69.6	79.5	68.0	64.6	89.1	95.7	52.0	83.1	92.0	79.4
LIM [10]	-	-	-	-	-	-	-	68.5	-	97.9	56.6	-	94.5	-
SCO [13]	55.5	82.0	89.3	41.1	70.5	80.1	69.7	69.9	92.9	97.5	56.7	87.5	94.8	83.2
SCO++ [14]	58.0	84.5	90.5	43.9	72.9	81.6	71.9	71.3	93.8	98.0	58.2	88.8	95.3	84.2
SCAN [21]	67.4	90.3	95.8	48.6	77.7	85.2	77.5	72.7	94.8	98.4	58.8	88.4	94.8	84.7
GVSE [11]	68.5	90.9	95.5	50.6	79.8	87.6	78.8	72.2	94.1	98.1	60.5	89.4	95.8	85.0
ACMM	80.0	95.5	98.2	50.2	76.8	84.7	80.9	81.9	98.0	99.3	58.2	87.3	93.9	86.4
ACMM	85.2	96.7	98.4	53.8	79.8	86.8	83.5	84.1	97.8	99.4	60.7	88.7	94.9	87.6

Figure 5. Failure cases of our proposed model. Rarely appeared words are marked as red (best viewed in colors).

are very high, and some of them are even several hundreds. We find they mostly include very complex visual content, which are described by at least 3 few-shot words (marked as red) in sentences. Although our model is able to extract the generic representation for each few-shot word, but the co-occurrence of too many few-shot words might easily confuse our model. A possible solution is to use external knowledge bases [43, 32] to provide more useful cues to well capture the intrinsic relation among few-shot words.

## 5. Conclusions and Future Work

In this work, we have proposed the Aligned Cross-Modal Memory (ACMM) for the rarely studied scenario namely few-shot image and sentence. The main contributions of this work are: 1) cross-modal aligning regions to words with a cross-modal graph convolutional network, and 2) memorizing cross-modal shared semantic representations with persistent memory items. We have comprehensively investigated the influence of different modules on the final perfor-

mance, and verified the effectiveness of our proposed model by achieving significant performance improvements.

In the future, we will extensively study how the hyper-parameters in the proposed model affect the final performance, instead of simply using the default ones now.

## Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YF-B1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Capital Science and Technology Leading Talent Training Project (Z181100006318030), Beijing Science and Technology Project (Z181100008918010), HW2019SOW01, and CAS-AIR. This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.



## References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.
- [2] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2016.
- [3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [4] Aviv Eisenschlat and Lior Wolf. Linking image and text with 2-way nets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4601–4611, 2017.
- [5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.
- [6] 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.
- [7] Yanwei Fu, Yongxin Yang, Tim Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-label zero-shot learning. *arXiv preprint arXiv:1503.07790*, 2015.
- [8] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- [9] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471, 2016.
- [10] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *arXiv preprint arXiv:1711.06420*, 2017.
- [11] Yan Huang, Yang Long, and Liang Wang. Few-shot image and sentence matching via gated visual-semantic matching. In *AAAI Conference on Artificial Intelligence*, 2019.
- [12] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.
- [13] Yan Huang, Qi Wu, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.
- [14] Yan Huang, Qi Wu, Wei Wang, and Liang Wang. Image and sentence matching via semantic concepts and order learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [15] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2014.
- [16] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
- [17] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*, 2015.
- [18] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.
- [19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [20] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1576–1585, 2018.
- [21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, pages 0–0, 2018.
- [22] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850, 2016.
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.
- [24] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277, 2016.
- [25] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew. Learning a recurrent residual fusion network for multimodal matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4107–4116, 2017.
- [26] Yang Long, Li Liu, Yuming Shen, Ling Shao, and J Song. Towards affordable semantic searching: Zero-shot. retrieval via dominant attributes. In *AAAI Conference on Artificial Intelligence*, 2018.
- [27] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory-augmented networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6975–6984, 2018.
- [28] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *IEEE International Conference on Computer Vision*, pages 2623–2631, 2015.

- [29] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks. In *International Conference on Learning Representations*, 2015.
- [30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.
- [31] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.
- [32] Medhini Narasimhan and Alexander G Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *European Conference on Computer Vision*, pages 451–468, 2018.
- [33] Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.
- [34] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.
- [36] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.
- [37] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7736–7745, 2018.
- [38] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2440–2448, 2015.
- [39] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016.
- [40] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2017.
- [41] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7512–7520, 2018.
- [42] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.
- [43] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.
- [44] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. *arXiv preprint arXiv:1806.01810*, 2018.
- [45] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.
- [46] Jónatas Wehrmann and Rodrigo C Barros. Bidirectional retrieval made simple. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7718–7726, 2018.
- [47] J. Weston, S. Chopra, and A. Bordes. Memory networks. In *International Conference on Learning Representations*, pages 0–0, 2015.
- [48] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.
- [49] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pages 2397–2406, 2016.
- [50] Quanzeng You, Zhengyou Zhang, and Jiebo Luo. End-to-end convolutional semantic embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5735–5744, 2018.
- [51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [52] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.