The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Few-Shot Image and Sentence Matching via Gated Visual-Semantic Embedding

Yan Huang,[1,3] **Yang Long,**[4] **Liang Wang**[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
[4]Open Lab, School of Computing, Newcastle University
{yhuang, wangliang}@nlpr.ia.ac.cn yang.long@newcastle.ac.uk

Abstract

Although image and sentence matching has been widely studied, its intrinsic few-shot problem is commonly ignored, which has become a bottleneck for further performance improvement. In this work, we focus on this challenging problem of few-shot image and sentence matching, and propose a Gated Visual-Semantic Embedding (GVSE) model to deal with it. The model consists of three corporative modules in terms of uncommon VSE, common VSE, and gated metric fusion. The uncommon VSE exploits external auxiliary resources to extract generic features for representing uncommon instances and words in images and sentences, and then integrates them by modeling their semantic relation to obtain global representations for association analysis. To better model other common instances and words in rest content of images and sentences, the common VSE learns their discriminative representations directly from scratch. After obtaining two similarity metrics from the two VSE modules with different advantages, the gated metric fusion module adaptively fuses them by automatically balancing their relative importance. Based on the fused metric, we perform extensive experiments in terms of few-shot and conventional image and sentence matching, and demonstrate the effectiveness of the proposed model by achieving the state-of-the-art results on two public benchmark datasets.

Introduction

Image and sentence matching has drawn much attention, which aims to measure the visual-semantic similarity between an image and a sentence. It has been widely applied to the task of cross-modal retrieval, *e.g.*, given a query image to find similar sentences, namely image annotation, or given a query sentence to retrieve matched images, namely image search. Recently, much progress in this direction has been achieved, and many effective methods (Huang, Wu, and Wang 2018; Gu et al. 2017; Faghri et al. 2017) are proposed to deal with this task.

However, different from existing methods, here we focus on a more challenging problem in terms of few-shot image and sentence matching, which is commonly existed but rarely investigated. Particularly, most state-of-the-art methods learn discriminative data representations jointly with cross-modal association based on pairs of image and sentence, which makes them easily recognize and associate those pairs containing frequently appeared instances and words in a statistical manner. But they cannot well handle certain pairs having uncommon instances and words, because the number of similar pairwise data is quite limited during training.

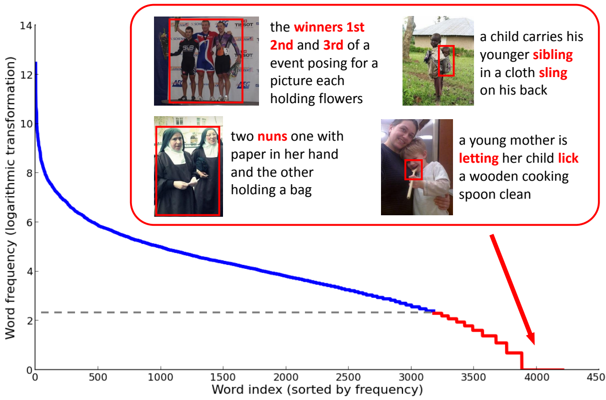


Figure 1: Logarithmic frequencies during training v.s. words from test set. (best viewed in colors).

We build a vocabulary consisting of all words extracted from test sentences in Flickr30k dataset (Young et al. 2014), and then count their appearing frequencies (after logarithmic transformation) during training in Figure 1. From the figure, we can see that about 24% words have very low frequencies of less than 10 (threshold is denoted by the dashed line), and even 8% words have frequencies of 1 or 0[1]. For these images and sentences containing uncommon instances and words (marked by red color), even current state-of-the-art methods cannot well associate them, and their performance drops heavily as demonstrated in the Tables 2 and 3. In fact, this few-shot problem has become a bottleneck for further performance improvement in image and sentence matching.

Although the traditional few-shot learning problem for

[1] Since the logarithmic transformation of 0 is negative infinity, for simple illustration, we plot the logarithmic frequencies as 0.

images and class labels has been widely studied, it is nontrivial to apply existing methods (Frome et al. 2013) to the new scenario of image and sentence matching directly. Rather than only one major instance or word in traditional few-shot settings, images and sentences in few-shot matching could have very complex content, which simultaneously includes multiple uncommon instances such as objects, actions and properties, and words such as nouns, verbs and adjectives. In addition, it has been demonstrated that the modeling of semantic relation among these instances or words plays an even more important role (Huang, Wu, and Wang 2018) in image and sentence matching. But how to suitably uncover and represent these uncommon instances and words, as well as model their semantic relation is unclear. Moreover, images and sentences usually contain both uncommon and common content in a hybrid manner, *i.e.*, their included uncommon instances or words only appear in small content, while in most of the rest content there are more common ones. So how to balance their relative importance when measuring the cross-modal similarity is another issue.

To deal with these issues, we propose a novel model named Gated Visual-Semantic Embedding (GVSE), which consists of two corporative VSE modules, *i.e.*, a uncommon VSE and a common VSE, and a gated metric fusion module. The uncommon VSE focuses on recognizing and associating uncommon instances and words that are rarely appeared in the training set. It first uses pretrained Faster-RCNN (Anderson et al. 2017) and Skip-Gram (Mikolov et al. 2013) based on external auxiliary resources to extract generic features for instances and words. Then it exploits attentional LSTM and Skip-Thought LSTM (Kiros et al. 2015) to model the semantic relation in images and sentences, respectively, and generates global representations for association analysis. Different but complementary to uncommon VSE, the common VSE focuses on learning discriminative features for images and sentences from scratch using advanced CNN and LSTM. In this way, the included common instances and words with higher appearing frequencies in the training set tend to be better associated.

After model learning, the two VSE modules can generate two different similarity metrics given same images and sentences. To take advantage of their complementary properties, the gated metric fusion module fuse them in a weighted sum manner, in which two weights can be automatically predicted to balance their relative importance. To demonstrate the effectiveness of the proposed model, we use the fused metric for several experiments in terms of few-shot and conventional image and sentence matching on two publicly available datasets, and achieve the state-of-the-art results.

Related Work

Visual-Semantic Embedding

Frome *et al.* (Frome et al. 2013) propose the Visual-Semantic Embedding (VSE) framework, which aligns image features with word features with ranking loss. Kiros *et al.* (Kiros, Salakhutdinov, and Zemel 2015) replace the word features with LSTM (Hochreiter and Schmidhuber 1997) for sentence representation learning, and extend this framework to image and sentence matching. Under similar framework, Vendrov *et al.* (Vendrov et al. 2016) achieve better performance by preserving the order structure of visual-semantic hierarchy. Wang *et al.* (Wang, Li, and Lazebnik 2016) add within-view constraints to ranking loss with the goal to learn structure-preserving representations.

Recently, more methods are proposed to improve the feature representability of VSE. Huang *et al.* (Huang, Wang, and Wang 2017) use different variants of attention modeling to compare local image regions with words. Gu *et al.* (Gu et al. 2017) attempt to incorporate generative processes into the VSE, which can learn more robust embeddings. Ren *et al.* (Ren et al. 2016) improve the VSE by modelling text concepts as Gaussian distributions in semantic space. Huang *et al.* (Huang, Wu, and Wang 2018) learn semantic concepts from images and organize them in a semantic order, which achieves the state-of-the-art results. Different from them, we aim to study the rarely studied problem of few-shot image and sentence matching by the proposed GVSE model.

Few-Shot Multimodal Learning

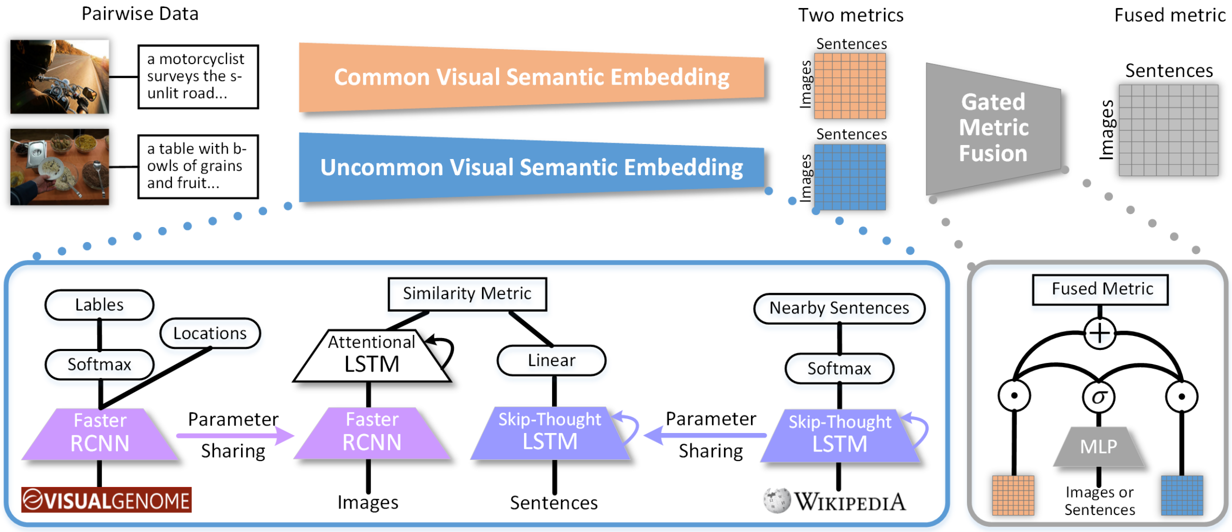
There is also some related work studying few-shot learning for multimodal data. Hendricks *et al.* (Anne Hendricks et al. 2016) propose the deep compositional captioner based on external images and text corpora for zero-shot image captioning. In contrast to the captioning methods, we focus on cross-modal matching rather than generation. Socher *et al.* (Socher et al. 2013) and Frome *et al.* (Frome et al. 2013) introduce VSE models that can recognize unseen objects in images by aligning image to semantic word space. Long *et al.* (Long et al. 2018) study the zero-shot problem in the image-attribute retrieval task. Wang *et al.* (Wang, Ye, and Gupta 2018) use graph convolutional networks for zero-shot classification. Rather than single word or multiple attributes (words), here we aim to deal with the few-shot matching for sentences, which have more complex semantic relation.

Few-Shot Image and Sentence Matching

We illustrate the proposed Gated Visual-Semantic Embedding (GVSE) model for few-shot image and sentence matching in Figure 2. Given pairwise training data, the model first learns two parallel VSE modules, namely uncommon VSE and common VSE, which focus on matching uncommon instances and words and common ones, respectively. Then two similarity metrics having these complementary properties can be predicted by two VSE modules, and then fused by the gated metric fusion module to finally produce the desired metric. In the next, we will explain these three modules in details.

Uncommon Visual-Semantic Embedding

1. Generic Features for Instances and Words For a pair of image and sentence, the image could contain various instances in terms of objects, their properties and actions, and the sentence might have diverse words in terms of nouns, adjectives, numbers and verbs. Therefore, few-shot image and sentence matching actually indicates a many-to-many



(a) Uncommon Visual Semantic Embedding

(b) Gated Metric Fusion

Figure 2: The proposed Gated Visual-Semantic Embedding (GVSE) model.

scenario in which multiple instances and words could be uncommon simultaneously. It is more challenging than the one-to-one setting in traditional few-shot learning (Frome et al. 2013; Socher et al. 2013), in which only one major instance and one class word are uncommon. To deal with this issue, we use pretrained models on large-scale external auxiliary images and sentences, to obtain generic features that can well detect and describe multiple uncommon instances and words.

In particular for an image, considering that not all regions contain the desired instances, we employ a Faster RCNN (Ren et al. 2015) jointly as an instance detector and a feature extractor. We follow (Anderson et al. 2017) to use the Faster RCNN that is pretrained on the large-scale Visual Genome dataset (Krishna et al. 2017), and modify its output to additionally predict attribute score vectors for detected instances, with the goal to enhance the generalization ability. When given an input image, the model can output I detected instances, each of which has a F-dimensional generic feature vector indicating the probabilities of belonging to a predefined set of objects and attributes. Although the target instance might not be included in the predefined set, the set is large enough so that it can be regarded as an attribute basis to comprehensively describe the target instance. While for a sentence, we follow (Frome et al. 2013) to use a Skip-Gram (Mikolov et al. 2013) pretrained on a large-scale corpus extracted from `wikipedia.org`. The model actually constructs a very large word vocabulary to encode arbitrary words into their generic features.

2. Semantic Relation Modeling for Association After obtaining the generic features for instances and words, we need to integrate them to represent the global image and sentence for cross-modal association analysis. A straightforward approach is to average all the features of instances or words as the unified representation. However, such an approach ignores the intrinsic semantic relation (Huang, Wu, and Wang 2018) among these instances and words, *i.e.*, the sequential order of words and spatial layout of instances. The relation actually plays an essential role during association analysis, since different organizations of instances or words could lead to diverse semantic meanings.

To model the sequential order of sentence, we use the word-level Skip-Gram features as pretrained vocabulary and extend it to a sentence-level Skip-Thought LSTM (Kiros et al. 2015) with multiple gating components for sequential information modeling. Instead of using a word to predict its surrounding words, the model encodes a sentence to predict its nearby sentences, as illustrated in Figure 2 (a). We regard the hidden state at the last timestep after a linear transformation layer as the desired sentence representation $\mathbf{s} \in \mathbb{R}^H$. Note that this representation is highly generic, because the Skip-Thought LSTM is trained on a large corpus of documents, which can encoder arbitrary sentences into their representations. While for an image, it is very difficult to directly model its spatial layout due to the lack of related annotations. But since we have already built a semantic relation-preserving space for sentences, we can similarly learn a sequential projection to integrate unorganized instances, and then align them to the sentence space to learn the semantic relation.

To sequentially integrate the instances, we resort to an attentional LSTM that can selectively attend to salient instances at each timestep and then fuse their features in a sequential order. Different from (Xu et al. 2016; Anderson et al. 2017), here we do not jointly perform image captioning. We denote the feature set of instances as $\{\mathbf{a}_i|\mathbf{a}_i \in \mathbb{R}^F\}_{i=1,\cdots,I}$, where $\mathbf{a}_i$ is the feature vector of the i-th detected instance. Based on this denotation, we can formulate the attention procedure as follows:

$$p_{t,i} = e^{\hat{p}_{t,i}} / \sum\nolimits_i e^{\hat{p}_{t,i}}, \ \hat{p}_{t,i} = f(\mathbf{a}_i, \mathbf{h}_{t-1}), \ \mathbf{a}'_t = \sum\nolimits_i p_{t,i}\mathbf{a}_i \quad (1)$$

where $p_{t,i}$ is the saliency value indicating the probability that the i-th instance will be attended to at the t-th timestep, $\mathbf{h}_{t-1}$ is the hidden state of attentional LSTM at the previous timestep, and $f(\cdot)$ is a two-way Multi-Layer Perceptron (MLP) to fuse $\mathbf{a}_i$ and $\mathbf{h}_{t-1}$. Note that we use the predicted saliency values of all instances as their weights to obtain the representation $\mathbf{a}'_t$ for attended instances in a soft manner, where the instances with higher saliency values contribute more to the fused representation. From the 1-st to T-th timestep, we can obtain a sequence of representations $\{\mathbf{a}'_t\}_{t=1,\cdots,T}$, where T is the total number of timesteps. Similar to Skip-Thought LSTM, we sequentially feed these representations into the attentional LSTM, and regard the hidden state at the last timestep as the desired image representation $\mathbf{v} \in \mathbb{R}^H$.

After obtaining the global representations for image and sentence, we use a structured objective to associate them, which encourages the cosine similarity score of matched image and sentence to be larger than a mismatched one with maximum violation:

$$\begin{aligned}&L(i,j|\Phi_{attlstm}, \Phi_{linear}) = \\ &\max_k[0, m - s_{ii} + s_{ik}]_+ + \max_k[0, m - s_{ii} + s_{ki}]_+\end{aligned} \tag{2}$$

where m is a margin parameter, $[x]_+ = max(x, 0)$, s_{ii} is the score of matched i-th image and i-th sentence, s_{ik} is the score of mismatched i-th image and k-th sentence, and vice-versa with s_{ki}. We empirically set the total number of mismatched pairs for each matched pair as 128 in our experiments. $\Phi_{attlstm}$ and Φ_{linear} are learning parameters of the attentional LSTM and linear transformation, respectively. Note that we fix the parameters of both Faster RCNN and Skip-Thought LSTM, with the goal to preserve their generalization ability from being disturbed by the majority of images and sentences with common content during learning.

Common Visual-Semantic Embedding

For a pair of image and sentence, only partial instances and words are uncommon, while most of the rest are common and have high appearing frequencies. The uncommon VSE mentioned above is especially designed to handle the uncommon content, and might be inferior to handle those common instances and words, since the fixed generic representations are not very discriminative for distinguishing them. Therefore, to better model the common content, the common VSE replaces the fixed generic representations, and learns discriminative representations for images and sentences directly from scratch. The representations are jointly learnt with cross-modal association based on pairwise training data, which tend to focus more on frequently appeared instances and words.

For simple implementation, we degenerate the previously used Faster RCNN and attention LSTM to a simple residual CNN, and replace the word vocabulary in Skip-Thought LSTM with a smaller one only consisting of the words from training sentences. To learn this model, we also use a structured objective similar to Equation 2, but optimize different learning parameters:

$$\begin{aligned}&L(i,j|\Phi_{cnn}, \Phi_{lstm}, \Phi_{linear}) = \\ &\max_k[0, m - s_{ii} + s_{ik}]_+ + \max_k[0, m - s_{ii} + s_{ki}]_+\end{aligned} \tag{3}$$

where Φ_{cnn}, Φ_{lstm} and Φ_{linear} are learning parameters of the CNN, LSTM and linear transformation, respectively. During learning, we follow (Faghri et al. 2017; Zheng et al. 2017) to exploit a two-step learning strategy to prevent the over-fitting problem. In the first step, we fix the parameters of pretrained residual CNN, and only optimize the parameters of LSTM and linear transformation. In the second step, we jointly optimize all the parameters and carefully fine-tune the CNN for more discriminative representations.

Gated Metric Fusion

After learning the two VSE modules, we can obtain two different metrics that are good at measuring similarities for uncommon and common content, respectively. To further exploit their complementary advantages, a simple method is to directly sum them as a new metric, in which importance weights of two metrics are equivalent. But in practice, their optimal importance weights are not always equivalent. For those images and sentences with uncommon content, the predicted metric by uncommon VSE should be more important than the metric predicted by common VSE. Based on this consideration, we want to adaptively learn the importance weights for two metrics rather than simply sum.

Although similar ideas of gated feature fusion have been previously explored (Arevalo et al. 2017), directly applying them to the fusion of metrics is infeasible. In the context of cross-modal retrieval, given a query and a gallery set with a size of N, the predicted two metrics can be denoted as $\mathbf{m}_0, \mathbf{m}_1 \in \mathbb{R}^N$. When varying the size of gallery set, the size of metrics changes accordingly. So it is inconsistent with the existing gated methods, which require the fused feature vectors must have a fixed size. In addition, they usually perform element-wise fusion by predicting importance weight for each feature dimension, while in our case we only need scalar importance weights for two metrics.

To handle this problem, we design a gated metric fusion module as illustrated in Figure 2 (b), which includes a gate to automatically control how much importance the two metrics contribute to their fused metric. To deal with the varying size problem, we use the representation of query instead of original metrics to predict the importance weights, which has a fixed size. Particularly, we concatenate two representations of the query in two VSE modules as $\mathbf{x} \in \mathbb{R}^{2H}$, and then feed it to a two-class classifier $f(\cdot)$ based on MLP to predict its probability t of containing uncommon content. Then the probability can be regarded as the importance weight for the metric predicted by uncommon VSE as follows:

$$t = \sigma(f(\mathbf{x})), \;\; \widehat{\mathbf{m}} = t \odot \mathbf{m}_0 + (1-t) \odot \mathbf{m}_1 \tag{4}$$

where sigmoid function σ is to rescale the probability value to $[0, 1]$, and $\widehat{\mathbf{m}}$ is the fused metric that can be used for image and sentence matching. Note that although all the three modules can constitute a very complex network, jointly training them in an end-to-end manner is infeasible, because the model could be easily over-fitted by the limited size of training data.

Table 1: Few-shot 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	
uncommon VSE:	54.8	82.1	89.3	34.8	57.8	70.6	64.9	31.2	61.9	73.1	22.3	50.2	62.6	50.2
+ relation	55.4	81.0	89.9	43.1	68.1	75.5	68.8	40.3	71.7	83.9	23.3	52.9	65.3	56.2
common VSE:	39.9	70.2	79.8	25.5	52.0	64.2	55.3	35.8	61.9	74.1	16.7	41.3	51.6	46.9
+ finetuned	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
Gated VSE:	60.1	**87.5**	**93.5**	44.6	72.5	80.9	73.2	46.3	76.4	87.8	31.1	61.0	70.1	62.1
+ gated	**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**

Experimental Results

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

Datasets and Protocols

The two evaluation datasets and their experimental protocols are described as follows. 1) Flickr30k (Young et al. 2014) consists of 31783 images collected from the Flickr website. Each image is accompanied with 5 human annotated sentences. 2) MSCOCO (Lin et al. 2014) consists of 82783 training and 40504 validation images, each of which is associated with 5 sentences.

For conventional image and sentence matching, we use the public training, validation and test splits on the two datasets. On the MSCOCO dataset, we perform 5-fold cross-validation and report the averaged results when using 1000 images for test. For few-shot image and sentence matching, we perform the K-shot matching ($K\in\{0,1,2,3\}$) on the two datasets. For each dataset, we select partial images and sentences from the standard test set to constitute a new few-shot test set, in which each sentence or image contains at least one word or instance whose appearing frequency in training set is less than or equals to K.

Implementation Details

The commonly used evaluation criterions for image annotation and retrieval are "R@1", "R@5" and "R@10", *i.e.*, recall rates at the top 1, 5 and 10 results. We also compute an additional criterion "mR" by averaging all the 6 recall rates, to evaluate the overall performance for both image annotation and retrieval.

For images, the number of detected instances is I=36, the dimension of predicted generic feature vectors is F=2002, and the number of timesteps in the attentional LSTM is T=3. For sentences, the dimension of embedded word is 300. We set the max length for all the sentences as 50, and use zero-padding when a sentence is not long enough. Other parameters are empirically set as follows: H=1024 and m=0.2.

When training the three modules, we use stochastic gradient descent with a learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, batch size of 128, and gradient clipping at 0.1. The VSE modules are trained for 30 epochs to guarantee its convergence. While for the gated metric fusion module, we use 100 epochs.

Evaluation of Ablation Models

To systematically demonstrate the effectiveness of the proposed modules, we study various ablation models as follows. 1) The basic "uncommon VSE" uses averaged features of instances or words as the global representations of images and sentences, while "+ relation" uses the attentional LSTM and Skip-Thought LSTM to model the semantic relation in images and sentences, respectively. 2) The basic "common VSE" uses the fixed pretrained CNN to extract image representations, while "+ fine-tuning" performs the two-step learning algorithm to additionally fine-tune the pretrained CNN. 3) The basic "Gated VSE" combines the two VSEs by directly summing their predicted similarity metrics, while "+ gated" replaces the averaged sum with the proposed gated metric fusion module.

We use these ablation models to perform the experiment of the few-shot image and sentence matching by setting K=0, and compare their results on the Flickr30k and MSCOCO datasets in Table 1. From the table, we can obtain the following observations. 1) Uncommon VSE using generic representations can better deal with the few-shot matching than common VSE based on self-learnt representations by a very large margin. 2) By modeling the semantic relation in images and sentences, uncommon VSE can further improve the mR performance by 3.9% and 5.0% on the two datasets, respectively. 3) Additionally performing fine-tuning in common VSE can improve the learned representations, but it still cannot outperform the powerful uncommon VSE + relation. 4) Due to the complementary properties of uncommon VSE and common VSE, even simply summing their predict metrics can greatly improve the performance. 5) By replacing the averaged sum with the proposed gated metric fusion, we can achieve better overall results. In the following experiments, we regard the best "Gated VSE + gated" as our default model.

Few-Shot Image and Sentence Matching

In this section, we will perform the experiment of few-shot image and sentence matching by varying the value of K from 0 to 3, and then make comparisons with two recent state-of-the-art methods in terms of VSE++ (Faghri et al. 2017) and SCO (Huang, Wu, and Wang 2018). For the two compared methods, we use their reported best model for conventional image and sentence matching, and then perform test on the K-shot test sets.

The comparison results are shown in Table 2, in which

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++	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	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
		Ours	**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**
1	321/ 754	VSE++	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	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
		Ours	**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**
2	437/ 973	VSE++	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	52.4	80.1	88.6	32.3	59.5	70.5	63.9	40.7	72.8	82.6	24.5	51.6	64.6	56.2
		Ours	**63.9**	**90.1**	**93.7**	**46.5**	**74.4**	**84.2**	**75.4**	**49.8**	**77.6**	**88.0**	**31.8**	**62.7**	**72.5**	**63.7**
3	513/ 1144	VSE++	51.7	79.3	88.5	31.2	61.4	73.1	64.2	42.4	72.2	82.0	23.3	50.3	63.0	55.5
		SCO	52.5	80.1	88.5	31.8	58.9	71.0	63.8	41.7	72.5	82.1	25.3	52.0	65.0	56.4
		Ours	**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**

Table 3: Conventional image and sentence matching on the Flickr30k and MSCOCO (1000 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	
DVSA	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	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	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	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	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
VQA	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	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
DSPE	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	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	49.8	67.5	-	36.0	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
RRF (Res)	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 (Res)	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-
CHAIN-VSE (Res)	-	-	-	-	-	-	-	59.4	88.0	94.2	43.5	79.8	90.2	75.9
DPCNN (Res)	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++ (Res)	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 (Res)	-	-	-	-	-	-	-	68.5	-	97.9	56.6	-	94.5	-
SCO (Res)	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
Ours	**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**

N indicates the numbers of uncommon words in K-shot test sets on the two datasets. We can see that in the most challenging 0-shot matching, our model can outperform the best method that we compare by 11.2% and 6.2% (in mR) on the two datasets, respectively. It indicates that our model can well understand and associate those unseen instances or words even they do not present in the training set. In addition, as K increases the N becomes much larger, but our model can consistently achieve much better performance, which demonstrates its good generalization ability on handling few-shot matching under different conditions.

Conventional Image and Sentence Matching

Although our model is especially proposed for dealing with the few-shot problem in image and sentence matching, it can also be applicable to conventional image and sentence matching by directly using standard test sets. We compare our model with recent published models on the Flickr30k and MSCOCO datasets in Table 3, including DVSA (Karpathy and Li 2015), MNLM (Kiros, Salakhutdinov, and Zemel 2015), m-CNN (Ma et al. 2015), RNN+FV (Lev et al. 2016), OEM (Vendrov et al. 2016), VQA (Lin and Parikh 2016), RTP (Plummer et al. 2015), DSPE (Wang, Li, and Lazebnik 2016), sm-LSTM (Huang, Wang, and Wang 2017), 2WayNet (Eisenschtat and Wolf 2017), RRF (Liu et al. 2017), DAN (Nam, Ha, and Kim 2017), CHAIN-VSE (Wehrmann and Barros 2018), DPCNN (Zheng et al. 2017), VSE++ (Faghri et al. 2017), LIM (Gu et al. 2017) and SCO (Huang, Wu, and Wang 2018). The methods marked by "(Res)" use the 152-layer ResNet (He et al. 2016), while the rest ones use the 19-layer VGGNet (Simonyan and Zisserman 2014).

We can see that our model outperforms the current state-of-the-art models on all 7 evaluation criterions on both the Flickr30k and MSCOCO datasets. It is mainly because our model can better associate those uncommon instances and words in the standard test sets, and thus greatly improve the overall performance. Note that our model obtains much larger improvements on the Flickr30k dataset than MSCOCO. It results from that the fewer training data of Flickr30k cannot guarantee the learnt models can well recognize instances and words in a learning-based way. But our

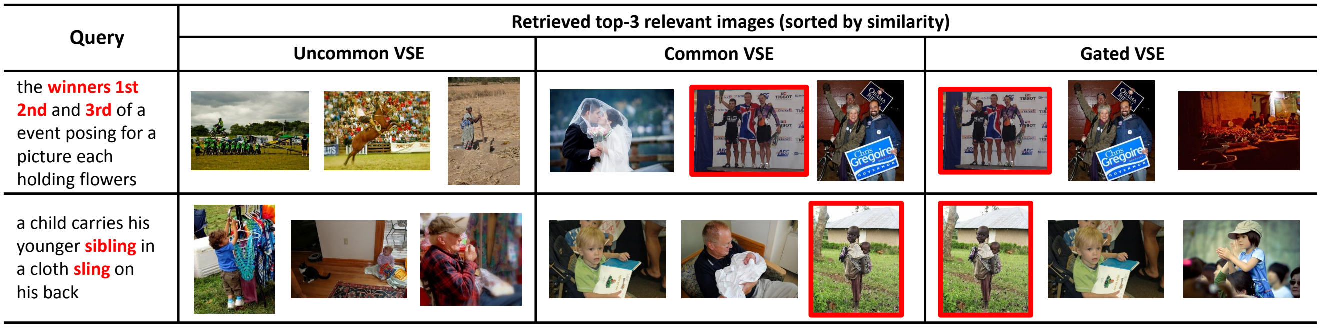

Query
Retrieved top-3 relevant images (sorted by similarity)
Uncommon VSE
Common VSE
Gated VSE
the winners 1st 2nd and 3rd of a event posing for a picture each holding flowers
a child carries his younger sibling in a cloth sling on his back

Figure 3: Results of few-shot image retrieval by 3 model variants. Groundtruth matched images and uncommon words are marked as red (best viewed in colors).

Table 4: Conventional image and sentence matching on the MSCOCO (5000 test) dataset.

Method	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA	11.8	32.5	45.4	8.9	24.9	36.3	26.6
FV	17.3	39.0	50.2	10.8	28.3	40.1	31.0
VQA	23.5	50.7	63.6	16.7	40.5	53.8	41.5
OEM	23.3	50.5	65.0	18.0	43.6	57.6	43.0
DPCNN (Res)	41.2	70.5	81.1	25.3	53.4	66.4	56.3
VSE++ (Res)	41.3	69.2	81.2	30.3	59.1	72.4	58.9
LIM (Res)	42.0	-	84.7	31.7	-	74.6	-
SCO (Res)	42.8	72.3	83.0	33.1	62.9	75.5	61.6
Ours	**49.9**	**77.4**	**87.6**	**38.4**	**68.5**	**79.7**	**66.9**

model can better describe them based on external auxiliary resources.

The above experiments on the MSCOCO dataset follow the first protocol (Karpathy and Li 2015), which uses 1000 images and their associated sentences for test. We also test the second protocol that uses all the 5000 images and their sentences for test, and present the comparison results in Table 4. We can observe that our model still achieves the best performance with a large performance gap, which again demonstrates its effectiveness. Note that our model has much larger improvements than those in the first protocol, which indicates the few-shot problem is more serious in larger gallery sets, and our model can well handle it to greatly improve the performance.

Analysis of Few-Shot Image Retrieval Results

To qualitatively validate the effectiveness of our proposed model, we analyze its results of few-shot image retrieval given sentence queries as follows. We select several representative sentences containing uncommon words (with appearing frequencies that are less than 10), and retrieve top-3 relevant images by 3 ablation models: “uncommon VSE”, “common VSE” and “Gated VSE” in Figure 3.

We can see that by learning representations only from provided pairwise data, common VSE cannot find matched images in top-ranked results. Because it cannot understand the meanings of uncommon words including “winners” and “sibling”, or recognize the corresponding instances in images. When using generic representations based on external auxiliary resources, uncommon VSE can well correlate those uncommon instances and words, and rank the matched images in top-3 results. But its performance is still not optimal, because the performance of few-shot cross-modal retrieval depends on not only the understanding of uncommon content, but also the discriminative representation of rest content containing common content. Therefore, Gated VSE incorporates the complementary advantages of common VSE and uncommon VSE, and is able to rank the matched images in top-1 results.

Conclusions and Future Work

In this work, we have proposed the Gated Visual-Semantic Embedding (GVSE) for the rarely investigated problem namely few-shot image and sentence matching. Our main contributions are: 1) greatly improving the representation and association of uncommon instances and words in images and sentences by the uncommon VSE module, and 2) adaptively fusing two similarity metrics by the gated metric fusion module. We have systematically studied the impact of different modules on the performance of few-shot image and sentence matching, and demonstrated the effectiveness of our model by achieving significant performance improvement.

In the future, we will consider to jointly train our model in an end-to-end manner and deal with the potential over-fitting problem. We will improve the modeling of semantic relation in the uncommon VSE module, and replace the attentional LSTM with more advanced implementations.

Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Beijing Natural Science Foundation (4162058), Capital Science and Technology Leading Talent Training Project (Z181100006318030), and Beijing Science and Technology Project (Z181100008918010). This work is also supported by MRC Innovation Fellowship with ref MR/S003916/1, and Chinese Academy Sciences Artificial Intelligence Research (CAS-AIR).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*.

Anne Hendricks, L.; Venugopalan, S.; Rohrbach, M.; Mooney, R.; Saenko, K.; and Darrell, T. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–10.

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Eisenschtat, A., and Wolf, L. 2017. Linking image and text with 2-way nets. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4601–4611.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2121–2129.

Gu, J.; Cai, J.; Joty, S.; Niu, L.; and Wang, G. 2017. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *arXiv preprint arXiv:1711.06420*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Huang, Y.; Wang, W.; and Wang, L. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2310–2318.

Huang, Y.; Wu, Q.; and Wang, L. 2018. Learning semantic concepts and order for image and sentence matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6163–6171.

Karpathy, A., and Li, F.-F. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, 3294–3302.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lev, G.; Sadeh, G.; Klein, B.; and Wolf, L. 2016. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, 833–850.

Lin, X., and Parikh, D. 2016. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, 261–277.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 740–755.

Liu, Y.; Guo, Y.; Bakker, E. M.; and Lew, M. S. 2017. Learning a recurrent residual fusion network for multimodal matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4107–4116.

Long, Y.; Liu, L.; Shen, Y.; Shao, L.; and Song, J. 2018. Towards affordable semantic searching: Zero-shot. retrieval via dominant attributes. In *AAAI Conference on Artificial Intelligence*.

Ma, L.; Lu, Z.; Shang, L.; and Li, H. 2015. Multimodal convolutional neural networks for matching image and sentence. In *IEEE International Conference on Computer Vision*, 2623–2631.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nam, H.; Ha, J.-W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 299–307.

Plummer, B.; Wang, L.; Cervantes, C.; Caicedo, J.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision*, 2641–2649.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 91–99.

Ren, Z.; Jin, H.; Lin, Z.; Fang, C.; and Yuille, A. 2016. Joint image-text representation by gaussian visual-semantic embedding. In *ACM Conference on Multimedia*, 207–211.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, 935–943.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2016. Order-embeddings of images and language. In *International Conference on Learning Representations*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5005–5013.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6857–6866.

Wehrmann, J., and Barros, R. C. 2018. Bidirectional retrieval made simple. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7718–7726.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2016. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. 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.

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