Knowledge Aware Semantic Concept Expansion for Image-Text Matching

Botian Shi[1*], **Lei Ji**[2,3†], **Pan Lu**[4], **Zhendong Niu**[1] and **Nan Duan**[3]
[1]Beijing Institute of Technology
[2]Institute of Computing Technology, Chinese Academy of Science, Beijing, China
[3]Natural Language Computing, Microsoft Research Asia, Beijing, China
[4]University of California, Los Angeles
botianshi@bit.edu.cn, leiji@microsoft.com, lupantech@gmail.com, zniu@bit.edu.cn, nanduan@microsoft.com

Abstract

Image-text matching is a vital cross-modality task in artificial intelligence and has attracted increasing attention in recent years. Existing works have shown that learning semantic concepts is useful to enhance image representation and can significantly improve the performance of both image-to-text and text-to-image retrieval. However, existing models simply detect semantic concepts from a given image, which are less likely to deal with long-tail and occlusion concepts. Frequently co-occurred concepts in the same scene, e.g. *bedroom* and *bed*, can provide common-sense knowledge to discover other semantic-related concepts. In this paper, we develop a Scene Concept Graph (SCG) by aggregating image scene graphs and extracting frequently co-occurred concept pairs as scene common-sense knowledge. Moreover, we propose a novel model to incorporate this knowledge to improve image-text matching. Specifically, semantic concepts are detected from images and then expanded by the SCG. After learning to select relevant contextual concepts, we fuse their representations with the image embedding feature to feed into the matching module. Extensive experiments are conducted on Flickr30K and MSCOCO datasets, and prove that our model achieves state-of-the-art results due to the effectiveness of incorporating the external SCG.

1 Introduction

Image and text matching is an important vision-language cross-modality task for many real-world applications including image retrieval [Xie *et al.*, 2016; Huang *et al.*, 2018] and caption [Karpathy and Fei-Fei, 2015; Lin and Parikh, 2016]. Before calculating the similarity between images and text, a matching model needs to obtain a rich representation of the images first. Inspired by the achievements of computer vision tasks, most of the current image-text matching models utilize pre-trained neural networks to extract mid-level feature embeddings as the representations of images. Although these representations can obtain global visual information of the images, they fail to extract high-level semantic information. So the semantic gap between images and language is not well addressed and it leads to the limited performance when matching an image with text.

Recently, some works have tried to learn semantic enhanced image representations for the image-text matching task. For example, [Karpathy and Fei-Fei, 2015] proposed to align image region with text words for image caption. [Huang *et al.*, 2018] exploited a multi-label CNN detection model to extract semantic concepts and then fused these concepts with global context of images, which has made significant improvement in image-text matching. However, these methods are limited to small size of detected concepts, since they depend on existing detection models normally dealing with common concept candidates. Therefore, these methods perform poorly if the concepts mentioned in the text cannot be detected via pre-trained detection models.

We argue that it is necessary to detect more accurate concepts from an image to help obtain a richer image representation. Take the example of Figure 1 as an example, current detection models (e.g. [Wang *et al.*, 2017]) could only detect related concepts *cat* and *laptop* from the given image. If the concept candidates are expanded to include more specific concepts like *calico* and *keyboard*, we can get higher semantic similarity score between the image and text. However, it is challenging to detect long-tail, even absent concepts based on existing detection models. These models are usually trained on a small-size concept vocabulary to avoid a sharp decrease of detection performance with a extended vocabulary. As shown in the Figure 1, the concept *calico* in the query text is a long-tail concept which is absent in the ldetection concept vocabulary and thus cannot be detected.

To address this issue, we incorporate external scene knowledge to privilege the imagination to the model. In this work, we mainly focus on utilizing co-occurrence common-sense knowledge which can be extracted from a large number of images. E.g. *laptop* and *keyboard*; *sky* and *cloud*; *boat* and *water*; *tree* and *leaves* are frequently appear in the same image. We define this common-sense knowledge as **Scene Concept Graph** (SCG). It provides rich prior scene information to expand more semantic concepts of images which are often

[*] This work was done during the first author's intership in MSR Asia
[†] Corresponding Author

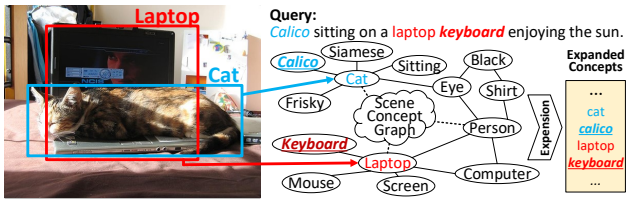

Figure 1: An example of image-retrieval by scene concept graph: semantic concepts {*cat, laptop*} are extracted, but the occluded concept *keyboard* and the long-tail concept *calico* cannot be detected. The scene concept graph can provide extra scene information which indicates that a laptop has a *keyboard* and *calico* is a breed of cat. These two concepts are the key for matching the query text and can be expanded.

noticed and described in query text. We directly use human annotated scene graphs from Visual Genome [Krishna *et al.*, 2017]. Scene graph of an image is a graph consisting of concepts and relationships between them. It can be represented as a set of triplets of <subject, relation, object>, where subject and object are concepts, and relations are the interactions and relationships between them. Then we aggregate co-occurred <subject, object> pairs from scene graphs of all images. After that, we clean the co-occurrence pairs by considering the appearance times and construct the SCG. After that, we can utilize this knowledge to infer more concepts. Figure 1 is an intuitive example of image-retrieval by SCG. In this case, the Scene Concept Graph provides common-sense information to expand the occluded concepts *keyboard* in the image. Since *laptop* and *keyboard* frequently appeared together, although the *keyboard* is occluded by the cat, it still can be inferred by co-occurrence pair in SCG. Furthermore, in order to avoid noisy concepts expanding, we used a neural network to select the semantic concepts related to the image. By introducing the SCG, the performance of our model outperforms state-of-the-art models.

Our contributions are summarized as below:

1. We build a Scene Concept Graph by considering co-occurrence pairs of semantic concepts in scene graph of images, which incorporate common-sense information.
2. We propose a novel model to expand more semantic concepts by the Scene Concept Graph and selectively fuse them to enhance image representation semantically.
3. We conduct extensive experiments and the results show that our model achieves state-of-the-art performance.

2 Related Work

Embedding based methods. DeViSE [Frome *et al.*, 2013] is the first embedding based method which projected image features and skip-gram word features by a linear mapping and calculated similarity accordingly. [Faghri *et al.*, 2017] introduced a simple modification on the loss function. [Kiros *et al.*, 2014] adopted a sequence to sequence pipeline to learn joint embedding space with a language model. [Wang *et al.*, 2016] proposed DSPE, a structured preserving network match image and text embeddings. [Zheng *et al.*, 2017] proposed using a dual task to embed image and text to a shared visual-textual space discriminatively. [Wang *et al.*, 2018] introduced two-branch embeddings and considered the novel neighborhood constraints. Cross-modality joint embedding is a baseline model but lack of semantic understanding.

Semantic knowledge based methods. [Karpathy and Fei-Fei, 2015] took advantage of R-CNN to detect local regions from an image and aligned them to each word. [Huang *et al.*, 2018] proposed learning semantic concepts and orders to improve image representation. This work is most similar to ours. The main difference is that we incorporate common-sense scene information to expand more contextual semantic concepts precisely. Many recent state-of-the-art models are attention based models. [Lee *et al.*, 2018b] proposed SCAN to use stacked cross attention to align image and text in a finer-grained model. Different from previous models, our work incorporated external common-sense scene information to enhance the representation of image and the experiment results showed a significant improvement.

Knowledge for image/text tasks. Combining a knowledge graph with deep learning for image-text tasks is related to our work. Knowledge can be used for image tasks like image classification [Marino *et al.*, 2017], zero shot classification [Lee *et al.*, 2018a] and for text tasks [Zhou *et al.*, 2018]. There are also several image and text tasks incorporating external knowledge like image captioning [Mogadala *et al.*, 2017]and VQA [Wu *et al.*, 2016]. For cross modality image and text match, [Wang *et al.*, 2006] used multi-modality ontology to retrieve images. [Belilovsky *et al.*, 2016] aligned scene graphs with images by the algorithms of bag-of-words, subpath representations and neural network. The difference is that we utilized the Scene Concept Graph which captures the co-occurred object pairs in the same scene as common-sense knowledge to expand more occluded and long tail concepts.

Our work starts from embedding based methods and utilized semantic concepts by incorporating Scene Concept Graph to enhance the representation of image/text embedding.

3 Scene Concept Graph Based Image-Text Matching

We formulate the image-text matching problem as a ranking model similar to [Wang *et al.*, 2018] and our model is depicted in Figure 2. Given the input image and the query text, the output is the similarity score of matching two modalities. We generate the pre-trained query text encoding and image encoding separately and jointly embed them into the same space by maximizing the margin of positive and negative image and text pairs. As shown in Figure 2, in order to incorporate contextual concepts to enhance image representation, we first adopt a *Concept Detection Module* to extract accurate concepts on a small vocabulary. Then we use the assembled *Scene concept Graph* to expand more contextual concepts by *Concept Expansion Module* and learn to predict more accurate semantic concepts by the *Concept Prediction Module*. Next we fuse whole-image encoding generated by *Vision Feature Module* with learned concepts to generate an enhanced image representation by the *Image-Concept Fusion Module*.

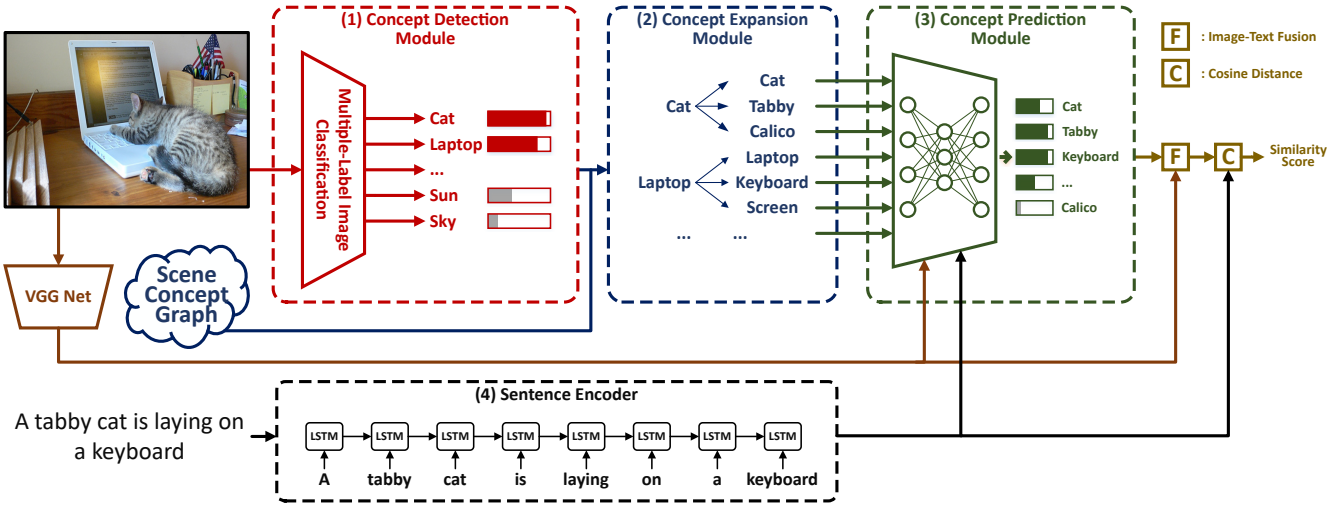

Figure 2: Main structure of our model: we extract semantic concepts first by *(1) Concept Detection Module* and then use *Scene Concept Graph* to expand more undetected concepts by *(2) Concept Expansion Module*. However, the expanded semantic concepts contains many unwanted noise, so we use *(3) Concept Prediction Module* to predict reliable concepts which are prone to efficiently enhance the image representation based on the visual feature (orange arrow). After the prediction, these filtered semantic concepts are fused with the vision feature and generate a enhanced representation of an image. Finally, we use cosine distance between enhanced image feature and text feature, extracted by *(4) Sentence Encoder*, as the similarity score and optimize parameters of the prediction model.

Finally we calculate the cosine similarity score with text encoding by *Text Encoding Module* and optimize the triplet loss function [Schroff *et al.*, 2015].

In the rest of this section, we first introduce the input text encoding. And then we describe the pre-trained whole-image encoding and the semantic-image feature encoding in details. The final loss function will be presented in the last part of this section.

3.1 Text Feature Encoding

We adopt LSTM [Hochreiter and Schmidhuber, 1997] as our text encoder to acquire the representation of text. For each word w_i in query text with N words, we train an embedding of x_i from scratch and input a token in text $S = \{x_i | x_i \in \mathbb{R}^k\}_{i=1...N}$ at each time step t to the LSTM model. We use the hidden state after feeding the last token as the representation of text f_t.

3.2 Image Feature Encoding

Visual Feature Module

For the whole-image representation (denoted as $\mathbf{f}_i \in \mathbb{R}^i$), we directly use a ImageNet [Deng *et al.*, 2009] pre-trained VGG19 [Simonyan and Zisserman, 2015] network and take the output of the last convolutional layer (16^{th} layer) as the image feature: we first resize each image to 256×256 and conduct a 5-cropping[1]. A left-right mirror flipping on each image is also applied to get 10 different pre-processed copies. Then, we feed this $10 \times 224 \times 224 \times 3$ (RGB 3-channel) image data matrix into the feature extractor and obtain a 10×4096-dimensional output of the model's first FC layer. After that, we average these feature vectors as the whole-image representation: $\mathbf{f}_i \in \mathbb{R}^{4096}$.

[1]Top-left, top-right, bottom-left, bottom-right and center.

Concept Detection Module

In order to bridge the semantic gap of text and image, we need to detect semantic concepts first. We use a multi-label images classification model [Wang *et al.*, 2017] to detect whether a concept appears in an image. The detection model aims to produce a multi-hot vector $\mathbf{g}_d \in [0,1]^{|\mathcal{V}_d|}$ as the detected concepts of the image, where $\mathcal{V}_d$ is the vocabulary of detectable concepts and $\mathcal{V}_c$ is the vocabulary of all semantic concepts; $\mathcal{V}_d \subseteq \mathcal{V}_c$ (mostly, $|\mathcal{V}_d| \ll |\mathcal{V}_c|$). For each i-th concept, we used 1 or 0 on i-th of vector to indicate existence or absence respectively. E.g. if the $\mathcal{V}_d = \{\text{people, cat, laptop, cake}\}$, and $\mathbf{g}_d$ of Figure 1 would be $[0,1,1,0]$. After that, we can directly fuse $\mathbf{g}_d$ with its whole-image feature encoding $\mathbf{f}_i$ to enhance image representation.

According to our experiment, although the detection model using the small concept vocabulary can guarantee the accuracy of concept detection, it will neglect long-tail but informative concepts. However, the performance of the detection model will drop significantly when using a large concept vocabulary. This motivates us to come up with a method to discover as many concepts as possible without accuracy loss. Therefore we introduce the *Scene Concept Graph* to expand more frequently queried concepts based on images and their description.

Concept Expansion Module

To learn to expand the concepts, we first build a commonsense Scene Concept Graph using Visual Genome [Krishna *et al.*, 2017], which is a dataset consisting of human-labeled scene graphs. We aggregate all co-occurring concept pairs in scene graphs of images to build a *Scene Concept Graph*: $\mathcal{K}_{SCG} = \{t | t =< c_s, c_o >; c_s, c_o \in \mathcal{V}_c\}$ where $\mathcal{V}_c$ is the concept vocabulary including all concepts. Figure 3 illustrates

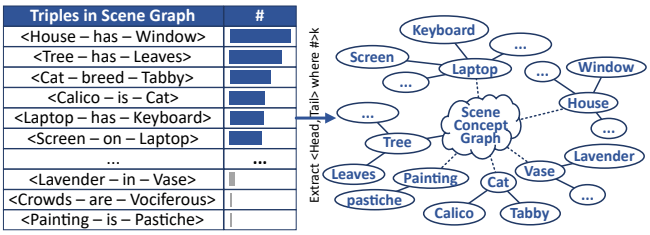

Figure 3: Scene Concept Graph Construction. We aggregate all Scene Graph triples of Visual Genome and # stands for the number of times triple appearance. We extracted <*head*, *tail*> pairs of triples which $\# > k$ to construct the Scene Concept Graph.

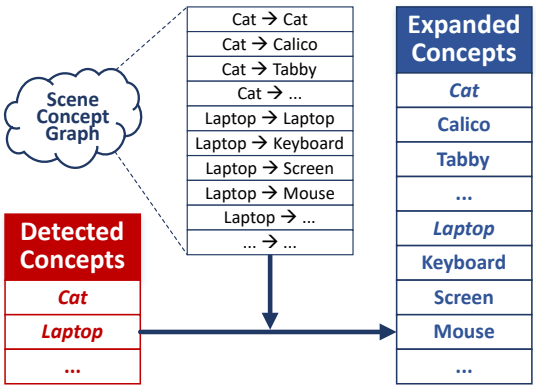

Figure 4: Procedure of Concept Expansion. We iterate semantic concepts detected in previous stage and use Scene Concept Graph to expand more concepts to which directly connected.

the construction of SCG. The SCG can provide strong contextual signals for semantic image understanding. For example, *House* can be found frequently in the pair <House, Window> and *Tree* always accompanies with *Leaves* in the pair <Tree, Leaves>. After we detect *House* and *Tree* from an image, expansion model will predict neglected concepts *Window* and *Leaves* by the common-sense scene information.

We formulate the concept expansion model as the algorithm 1. In this algorithm, we iterate all concepts $\mathbf{g}_d$ detected by the concept detection module as a seed to retrieve from the SCG. For each concept c, we extract all co-occurred concepts and aggregate all these concepts as expanded semantic concepts $\mathbf{g}_a \in [0,1]^{|\mathcal{V}_c|}$. This procedure is depicted in figure 4.

Concept Prediction Module

Although we can expand a lot of undetected semantic concepts by SCG, it is obvious that some noisy concepts irrelevant to the image will also be expanded, leading to even worse performance. How to selectively learn relevant concepts of the image is a challenge to us. For example, with a high probability, *laptop* co-occurred with *mouse*, which is a noisy for the example in Figure 1. So we propose a mechanism to selectively learn and predict whether a concept is relevant to the image or not.

To learn relevant concepts, we construct a prediction model which takes the whole-image encoding $\mathbf{f}_i$; the detected concept vector $\mathbf{g}_d$ and the expanded concept vector $\mathbf{g}_a$ as input and output the relevant concept vector $\hat{\mathbf{g}}_c \in [0,1]^{|\mathcal{V}_c|}$, where 1 indicates the corresponding concept is in the image and 0 otherwise.

Algorithm 1 Concept Expansion

Require: detected concepts of an image $L_d \subseteq \{c|c \in \mathcal{V}_d\}$
Require: Scene Concept Graph $\mathcal{K}_{SCG}$
Ensure: expanded concept vector $\mathbf{g}_a \in [0,1]^{|\mathcal{V}_c|}$
1: $L_a \leftarrow L_d$
2: **for** c in L_d **do**
3: $\quad L_{subject} \leftarrow \{c_s | < c_s, c_o > \in \mathcal{K}_{SCG}, c_o = c\}$
4: $\quad L_{object} \leftarrow \{c_o | < c_s, c_o > \in \mathcal{K}_{SCG}, c_s = c\}$
5: $\quad L_a \leftarrow L_a \cup L_{subject} \cup L_{objects}$
6: **end for**
7: convert concept list L_a to a multi-hop concept vector $\mathbf{g}_a \in [0,1]^{|\mathcal{V}_c|}$.

In order to predict whether concepts are relevant or not, we first fuse expanded concepts $\mathbf{g}_a$ and whole-image representation $\mathbf{f}_i$ by:

$$\mathbf{g}_f = \text{ELU}\left(\mathbf{f}_i \times \mathbf{M}_i + \mathbf{b}_i\right) \times \text{ELU}\left(\mathbf{g}_a \times \mathbf{M}_g + \mathbf{b}_g\right) \quad (1)$$

Then, we construct a model to predict g_c directly:

$$\hat{\mathbf{g}}_c = \sigma\left(\mathbf{g}_f \times \mathbf{M}_c + \mathbf{b}_c\right) \quad (2)$$

where $\mathbf{M}_i \in \mathbb{R}^{i \times k}$; $\mathbf{M}_g \in \mathbb{R}^{|\mathcal{V}_c| \times k}$; $\mathbf{M}_c \in \mathbb{R}^{k \times |\mathcal{V}_c|}$; $\mathbf{b}_i, \mathbf{b}_g \in \mathbb{R}^k$; $\mathbf{b}_c, \mathbf{b}_m \in \mathbb{R}^{|\mathcal{V}_c|}$, σ is the sigmoid function and $\circ$ is the element-wise product. $\hat{\mathbf{g}}_c \in \mathbb{R}^{|\mathcal{V}_c|}$ is the prediction of ground-truth $\mathbf{g}_c \in [0,1]^{|\mathcal{V}_c|}$. ELU is Exponential Linear Units [Clevert *et al.*, 2016].

To train this prediction model, we use description text of images to build a pseudo labels by matching the concepts with text literally. We heuristically label concept vector using 0, 1 as $\mathbf{g}_c \in [0,1]^{|\mathcal{V}_c|}$, where 1 means the concept appeared in the description text literally and 0 means the concept is in the text.

During training procedure, First, we use the text to label the expanded concepts $\mathbf{g}_a$ to generate the expected concepts $\mathbf{g}_c$. Then we train the prediction model to make the learned concept vector $\hat{\mathbf{g}}_c$ similar to expected concept vector $\mathbf{g}_c$. We use the log-exponential objective function to train this prediction model:

$$L_{cp} = \frac{1}{|\mathcal{V}_c|} \sum_{k=1}^{|\mathcal{V}_c|} \log\left(1 + \exp(-\mathbf{g}_{c,k} \times \hat{\mathbf{g}}_{c,k})\right) \quad (3)$$

Our overall model is trained by an end-to-end pipeline, and we add this prediction loss L_{cp} to the final loss function. After training, we fix the prediction model and directly run the model on inference phase.

Image-Concept Fusion Module

In this section, we will describe how we conduct image-concept fusion.

We fuse the predicted concept vector $\hat{\mathbf{g}}_c \in \mathbb{R}^{|\mathcal{V}_c|}$ with whole-image encoding $\mathbf{f}_i \in \mathbb{R}^{4096}$ to get a final image representation $\mathbf{f}_{ci} \in \mathbb{R}^e$ where e is the dimension of embedding vector. In this paper, we use two different fusion methods.

The baseline fusion method named *element-wise product*. First we map each of these two modality inputs into the same embedding space by projecting with $\mathbf{W}_g \in \mathbb{R}^{|\mathcal{V}_c| \times e}$ and $\mathbf{W}_f \in \mathbb{R}^{i \times e}$ respectively. Then, we normalize these two mapped embeddings by L2-normalization and combine them by element-wise product and $\mathbf{f}_{ci}^{(mul)} \in \mathbb{R}^e$ is the enhanced image representation.

The second fusion method is called *gated fusion* [Huang *et al.*, 2018], which selectively fuses concepts and efficiently parameterizes interactions between two modalities. We use a gated mechanism to learn to fuse those two vectors by the following equations:

$$\tilde{\mathbf{g}}_c = ||\hat{\mathbf{g}}_c \times \mathbf{W}_g||_2 \tag{4}$$

$$\tilde{\mathbf{f}}_i = ||\mathbf{f}_i \times \mathbf{W}_f||_2 \tag{5}$$

$$t = \sigma\left(\hat{\mathbf{g}}_c \times \mathbf{U}_g + \mathbf{f}_i \times \mathbf{U}_f\right) \tag{6}$$

$$\mathbf{f}_{ci}^{(sco)} = ||t \circ \tilde{\mathbf{g}}_c + (1-t) \circ \tilde{\mathbf{f}}_i||_2 \tag{7}$$

where $\mathbf{W}_g, \mathbf{U}_g \in \mathbb{R}^{|\mathcal{V}_c| \times e}$; $\mathbf{W}_f, \mathbf{U}_f \in \mathbb{R}^{i \times e}$ and the $\mathbf{f}_{ci}^{(sco)} \in \mathbb{R}^e$ is the concept-enhanced image feature.

3.3 Loss Function of Joint Learning

To learn image and text matching as well as image-relevant semantic concepts jointly, our loss function consists of two parts: prediction loss L_{cp} and image-text matching loss L_m so as to train the model in an end-to-end style. As for the loss L_m, we use cosine similarity to calculate the score. After generating the representation of image and text, we embed them jointly into the same space by maximizing the margin of positive and negative samples. We use triplet loss [Schroff *et al.*, 2015] for two modality encodings:

$$f_{rank}(x, y) = \sum_{y' \in \mathcal{N}_x} [\gamma + d(x, y) - d(x, y')]_+ \tag{8}$$

where x and y are encodings of two modality, $\mathcal{N}(x)$ is the set of negative samples of x, $d(\cdot)$ is the similarity function and $[x]_+ = max(0, x)$.

In order to train a discriminative model, we also adopt a neighborhood constraint [Wang *et al.*, 2018; Kiela *et al.*, 2018]:

$$\hat{f}_{rank}(x) = \sum_{\substack{\forall x_j \in \mathcal{P}(x_i) \\ \forall x_k \notin \mathcal{P}(x_i)}} [\gamma + d(x_i, x_j) - d(x_i, x_k)]_+ \tag{9}$$

where $\mathcal{P}(x_i)$ is the set of positive samples for x_i (e.g. all descriptions of the image that described by x_i).

Finally, we merge these ranking constraints into one loss function with the semantic concept prediction loss L_{cp} as we discussed before:

$$L = \lambda_1 L_{cp} + \lambda_2 \sum_{I,D} f_{rank}(I, D) + \lambda_3 \sum_{I,D} f_{rank}(D, I) + \lambda_4 \sum_{D} \hat{f}_{rank}(D) \tag{10}$$

where I is the image set and D is the descriptions of all images.

4 Experiments

We conducted extensive experiments to evaluate the performance of our model and compared different settings of our method to state-of-the-art algorithms. We also conducted ablation studies to analyze the effectiveness of incorporating different vocabulary size of semantic concepts for the final image-text matching task.

4.1 Datasets

Visual Genome [Krishna *et al.*, 2017] is a structured dataset that contains hundreds of thousands of dense annotated images. It has 2.3M human annotated scene graphs for 108,077 images, which is represented as a set of triplets including concepts and relationships. After the normalization by synsets of each object and the clearing of tailed triplets, we collected **7,699** different concepts as our vocabulary to construct a dataset for training the concept detection module. We aggregated the triplets, which appear at least 1000 times, to build the Scene Concept Graph which includes 121,307 concept pairs.

MSCOCO [Lin *et al.*, 2014] is a large-scale dataset which contains 123,287 images (the combination of train2014 and val2014), each accompanied with 5 captions which are the text used to query the image. We follow [Karpathy and Fei-Fei, 2015] to prepare the training, validation and test dataset by splitting all images to 113,287 (for training), 5,000 (for validation) and 5,000 (for test). For the evaluation on *MSCOCO 5K* setting, we used all these 5K testing images and their captions (25K). We used 1/5 of these testing dataset (1K images, 5K captions) for *MSCOCO 1K* evaluation in order to compare with some algorithm which report their result only on MSCOCO 1K dataset.

Flickr30K [Young *et al.*, 2014] is a dataset that consists of 31,783 images of events, activities and scenes and 158,915 captions which are used to query the image. We followed the split in [Karpathy and Fei-Fei, 2015] and [Faghri *et al.*, 2017] that used 1,000 images for testing and 1,000 images for validation and the rest of them (28,783 images) for training.

4.2 Evaluation Metrics

For the evaluation metric, we adopt the widely used measurement recall at K (R@K) for both sentence retrieval and image retrieval task. To be specific, each image in MSCOCO and Flickr30K dataset has 5 sentences as ground-truth, and each sentence has 1 corresponding image. Take image retrieval as an example, we rank the similarity scores of all images and select top K candidates. We regard the query as a "successful query" if the target image presents in these candidates. And the R@K is the proportion of success in all queries.

4.3 Implementation Details

In our two-modality joint embedding network, we used LSTM with 1024 hidden units to encode text of images and a VGG19 [Simonyan and Zisserman, 2015] pre-trained by ImageNet as our image feature extractor. The dimension of

Methods	MSCOCO 1K						MSCOCO 5K					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	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 [Karpathy and Fei-Fei, 2015]	38.4	69.9	80.5	27.4	60.2	74.8	11.8	32.5	45.4	8.9	24.9	36.3
VQA-ICR [Lin and Parikh, 2016]	50.5	80.1	89.7	37.0	70.9	82.9	23.5	50.7	63.6	16.7	40.5	53.8
DSPE [Wang *et al.*, 2016]	50.1	79.7	89.2	39.6	75.2	86.9	-	-	-	-	-	-
VSE++ [Faghri *et al.*, 2017]	64.6	90.0	95.7	52.0	84.3	92.0	41.3	71.1	81.2	30.3	59.4	72.4
TBNN[Wang *et al.*, 2018]	54.0	84.0	91.2	43.3	76.8	87.6	-	-	-	-	-	-
DPC [Zheng *et al.*, 2017]	65.6	89.8	95.5	47.1	79.9	90.0	41.2	70.5	81.1	25.3	53.4	66.4
DXN [Gu *et al.*, 2018]	68.5	-	97.9	56.6	-	94.5	42.0	-	84.7	31.7	-	74.6
SCO [Huang *et al.*, 2018]	69.9	92.9	97.5	56.7	87.5	94.8	42.8	72.3	83.0	33.1	62.9	75.5
SCAN [Lee *et al.*, 2018b]	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	**69.3**	80.4
Ours (Prod)	73.4	94.8	97.6	56.3	85.6	93.5	49.9	78.9	88.1	33.2	62.4	74.7
Ours (Gated)	**76.6**	**96.3**	**99.2**	**61.4**	**88.9**	**95.1**	**56.6**	**84.5**	**92.0**	**39.2**	68.0	**81.3**

Table 1: Experimental results on MSCOCO 1K and 5K. The sentence retrieval is to retrieve the correct sentence given an input image as a query. And the image retrieval is to search the specific image given a sentence as a query.

Methods	Flickr30K					
	Sentence Retrieval			Image Retrieval		
	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
VQA-ICR	33.9	62.5	74.5	24.9	52.6	64.8
DSPE	40.3	68.9	79.9	29.7	60.1	72.1
VSE++	41.3	69.0	77.9	31.4	59.7	71.2
TBNN	37.5	64.7	75.0	28.4	56.3	67.4
DPC	55.6	81.9	89.5	39.1	69.2	80.9
DXN	56.8	-	89.6	41.5	-	80.1
SCO	55.5	82.0	89.3	41.1	70.5	80.1
SCAN	67.4	90.3	**95.8**	48.6	**77.7**	85.2
Ours (Prod)	57.2	85.1	92.1	40.1	69.5	79.5
Ours (Gated)	**71.8**	**90.8**	94.8	**49.3**	76.4	**85.6**

Table 2: Experimental results on Flickr30K

whole-image representation is $i = 4096$ ($\mathbf{f}_i \in \mathbb{R}^{4096}$). The dimension of concept-enhanced image representation and text representation is $e = 512$ ($\mathbf{f}_t, \mathbf{f}_{ci} \in \mathbb{R}^{512}$).

In order to study the effectiveness of the concept detection model, we tried several different size of vocabularies by picking up the most frequent concepts in all images. The vocabulary sizes various from 256, 512 to 1024. Experiments indicated that we can get the best results by setting vocabulary size to $|\mathcal{V}_d| = 512$. Extensive experiments were conducted and will be explained in the ablation study section later. For the expansion and prediction module, we used all 7,699 concepts as $\mathcal{V}_c$, i.e. $|\mathcal{V}_c| = 7699$.

We used $\lambda_1 = 5.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.5$ and $\lambda_4 = 0.05$ as the hyper-parameters of loss function. An Adam Optimizer was adopted to optimize model's parameters. We also utilized batch normalization before non-linear activation and the dropout mechanism after non-linear activation in order to promote speed of training.

4.4 Comparison with State-of-the-Art Models

Table 1 lists the experiment results on MSCOCO dataset and a comparison with other methods. We tried two different evaluation datasets that using 1,000 and 5,000 test images respectively. Prod (Element-wise production) and Gated Fusion are the two different fusion methods. From the table we can see that our algorithm outperforms all of the previous state-of-the-art methods on both 1K and 5K splits and Gated setting except R@5 result on image retrieval for the 5K dataset. SCO [Huang *et al.*, 2018] is our baseline which also learned the semantic concepts from an image to facilitate this task. From the result, we can see our method significantly improved the matching performance on all datasets and settings over SCO which demonstrates Scene Concept Graph provides effective semantic information to understand the image. SCAN [Lee *et al.*, 2018b] is the most recent report on this task which is using stacked co-attention on a finer-grained level. Although our work and SCAN follow different research directions, our model can still get better results in most cases. Another conclusion is that Gated Fusion consistently outperforms the Production Fusion, which proved that the carefully designed mechanisms fuse better.

Table 2 shows the experiment results on the Flickr30K dataset. Our model with Gated Fusion can achieved the best results on most cases which is similar to the experiment conclusion of MSCOCO.

Ablation Methods (All used SCO fusion)	MSCOCO 1K					
	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
128 Concepts	48.5	78.4	88.1	37.0	71.9	83.8
256 Concepts	49.2	79.6	89.0	39.1	72.2	83.6
512 Concepts	**50.1**	79.1	89.0	**39.4**	**72.9**	**84.4**
768 Concepts	50.0	**80.9**	**90.4**	39.3	**72.9**	84.3
1024 Concepts	49.1	79.0	88.4	36.4	71.2	82.9
Expansion Only	49.6	78.9	88.5	38.7	72.6	84.6
Our Best (Gated Fusion)	**76.6**	**96.3**	**99.2**	**61.4**	**88.9**	**95.1**
Ground-Truth	*92.4*	*99.6*	*100*	*75.9*	*95.2*	*97.8*

Table 3: Ablation experiment results

4.5 Ablation Studies

We also investigated the effectiveness of detection, expansion, and prediction models by several ablation experiments. Table 3 shows the results of matching. First, we tried to explore the impact of different vocabulary sizes of the detection model. The vocabulary sizes are set to 129, 256, 512, 768 and 1024 separately. From the result, we can see that either small the size or the large size got worse results than the middle size 512, which demonstrated the existing shortcoming of the detection model. We set the vocabulary size to 512 in the rest of the experiments.

We conducted another experiment on semantic concepts generated using different methods. Since Gated Fusion always achieved the best results, we used all Gated Fusion mechanisms in the rest of the experiments. First, SCO [Huang *et al.*, 2018] in Table 1 shows our baseline result without expansion. Other results are shown in Table 3. Next we used the detected concepts $\mathbf{g}_d \in \mathbb{R}^{512}$ to expand more concepts $\mathbf{g}_a \in \mathbb{R}^{7699}$ by Scene Concept Graph without prediction

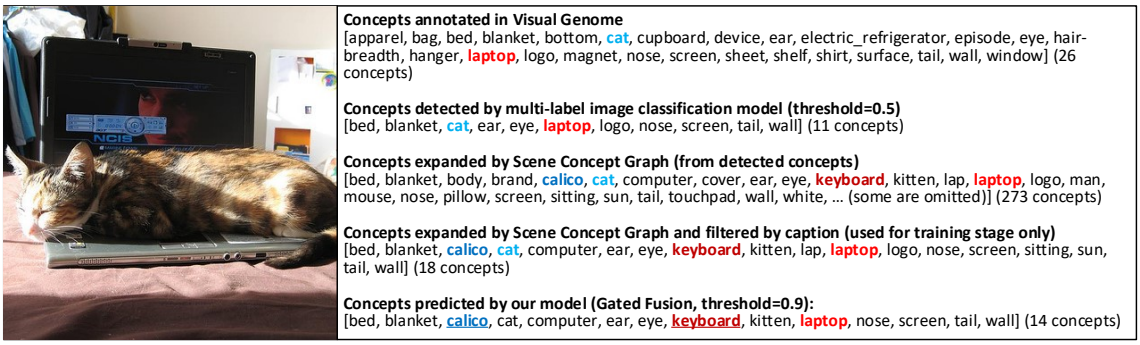

Figure 5: A real showcase of intermediate result of our expansion model: the image is annotated by human in Visual Genome with 26 concepts, and the detection model only extracted 11 concepts using vocabulary size 512. The expansion model discovers more concepts including occluded concept *keyboard* and long-tail concept *calico* etc. Finally, our model predicted 14 concepts for further process.

(*Expansion Only*) which got a worse result than our baseline. This is caused by introducing many noisy concepts. Then our best model (Gated Fusion) achieved a much better result. We also conducted an experiment on a ground-truth dataset, which we used to train our prediction model. This is the upper bound of our algorithm and shown that there is still a gap to predict the semantic concept well.

4.6 Statistics of Concept Expansion

	# Min	# Max	# Average
Detected Concepts	1	23	10.66
Expanded Concepts	0	392	107.56
Ground-truth Prediction Concepts	1	34	17.71

Table 4: Statistics of intermediate results of semantic concepts

Table 4 presents a statistics of concepts of the dataset. *Detected Concepts* indicates the numbers of semantic concepts that extracted by multi-label image classification model. The second row is the numbers of concepts that expanded by Scene Concept Graph. The third row is the numbers of semantic concepts expanded by SCG and filtered by caption tokens, which is the ground-truth label for training prediction model.

4.7 Case Study and Analysis

We ran our model on the real case in Figure 1 and inspected all intermediate results of semantic concept detection, expansion, and prediction models. The results are shown in Figure 5. Although some noisy concepts were introduced, our model can discover the concepts mentioned in text. Besides, we also presented several real showcases of retrieval using different methods. Because of the limited space, please check out this link: https://goo.gl/izcSN9.

5 Conclusion and Future Work

In this paper, we proposed learning semantic concepts using co-occurred common-sense knowledge for image-text matching. Our main contribution is to utilize the Scene Concept Graph we extracted by Scene Graph of Visual Genome to expand more semantic concepts, which can deal with the challenges of detecting occluded and long-tailed concepts. Experiment results demonstrated that our method can achieve state-of-the-art results. We also conducted intensive ablation studies which showed the effectiveness of incorporating the Scene Concept Graph.

In the future, we will consider using triplets in scene graphs by taking the relationship between concepts into account. We would also like to extend our scene knowledge aware model for more cross-modality tasks.

Acknowledgements

We thank the reviewers for their carefully reading and suggestions. This work was supported by the National Natural Science Foundation of China (No. 61370137), the National Basic Research Program of China (No.2012CB7207002), the Ministry of Education - China Mobile Research Foundation Project (2016/2-7)

References

[Belilovsky *et al.*, 2016] Eugene Belilovsky, Matthew Blaschko, Jamie Kiros, Raquel Urtasun, and Richard Zemel. Joint embeddings of scene graphs and images. In *International Conference On Learning Representations-Workshop*, 2016.

[Clevert *et al.*, 2016] Djorkarne Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *international conference on learning representations*, 2016.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *british machine vision conference*, page 12, 2017.

[Frome *et al.*, 2013] 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.

[Gu *et al.*, 2018] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Huang *et al.*, 2018] Yan Huang, Qi Wu, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Computer Vision and Pattern Recognition, CVPR*, 2018.

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

[Kiela *et al.*, 2018] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. *North American Chapter of the Association for Computational Linguistics*, 1:408–418, 2018.

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

[Krishna *et al.*, 2017] 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.

[Lee *et al.*, 2018a] Chungwei Lee, Wei Fang, Chihkuan Yeh, and Yuchiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. *Computer Vision and Pattern Recognition*, pages 1576–1585, 2018.

[Lee *et al.*, 2018b] 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 (ECCV)*, 2018.

[Lin and Parikh, 2016] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer, 2016.

[Lin *et al.*, 2014] 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. Springer, 2014.

[Marino *et al.*, 2017] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. *Computer Vision and Pattern Recognition*, pages 20–28, 2017.

[Mogadala *et al.*, 2017] Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. Describing natural images containing novel objects with knowledge guided assitance. *arXiv preprint arXiv:1710.06303*, 2017.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[Wang *et al.*, 2006] Huan Wang, Song Liu, and Liang-Tien Chia. Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 109–112. ACM, 2006.

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

[Wang *et al.*, 2017] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 464–472, 2017.

[Wang *et al.*, 2018] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony R Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. *Computer Vision and Pattern Recognition*, pages 4622–4630, 2016.

[Xie *et al.*, 2016] Liang Xie, Jialie Shen, Lei Zhu, et al. Online cross-modal hashing for web image retrieval. In *AAAI*, pages 294–300, 2016.

[Young *et al.*, 2014] 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.

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

[Zhou *et al.*, 2018] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.