Similarity Reasoning and Filtration for Image-Text Matching

Haiwen Diao,[1] Ying Zhang,[2] Lin Ma,[3] Huchuan Lu[1]*

[1] Dalian University of Technology, Dalian, China
[2] Tencent AI Lab, Shenzhen, China
[3] Meituan, Beijing, China
r1228240468@mail.dlut.edu.cn, yinggzhang@tencent.com,
forest.linma@gmail.com, lhchuan@dlut.edu.cn

Abstract

Image-text matching plays a critical role in bridging the vision and language, and great progress has been made by exploiting the global alignment between image and sentence, or local alignments between regions and words. However, how to make the most of these alignments to infer more accurate matching scores is still underexplored. In this paper, we propose a novel Similarity Graph Reasoning and Attention Filtration (SGRAF) network for image-text matching. Specifically, the vector-based similarity representations are firstly learned to characterize the local and global alignments in a more comprehensive manner, and then the Similarity Graph Reasoning (SGR) module relying on one graph convolutional neural network is introduced to infer relation-aware similarities with both the local and global alignments. The Similarity Attention Filtration (SAF) module is further developed to integrate these alignments effectively by selectively attending on the significant and representative alignments and meanwhile casting aside the interferences of non-meaningful alignments. We demonstrate the superiority of the proposed method with achieving state-of-the-art performances on the Flickr30K and MSCOCO datasets, and the good interpretability of SGR and SAF modules with extensive qualitative experiments and analyses.

Introduction

Image-text matching refers to measuring the visual-semantic similarity between image and text, which is becoming increasingly significant for various vision-and-language tasks, such as cross-modal retrieval (Wang et al. 2020), image captioning (Anderson et al. 2018), text-to-image synthesis (Xu et al. 2018), and multimodal neural machine translation (Toyama et al. 2017). Although great progress has been made in recent years, image-text matching remains a challenging problem due to complex matching patterns and large semantic discrepancies between image and text.

To accurately establish the association between the visual and textual observations, a large proportion of methods (Liu et al. 2017; Nam, Ha, and Kim 2017; Lee et al. 2018; Song and Soleymani 2019; Wang et al. 2019c; Li et al. 2019; Wang et al. 2020) utilize deep neural networks to firstly encode image and text into compact representations, and then learn to measure their similarity under the guidance of a matching criterion. For example, Wang *et al.* (Wang, Li, and Lazebnik 2016) and Faghri *et al.* (Faghri et al. 2017) map the whole image and the full sentence into a common vector space, and compute the cosine similarity between the global representations. To improve the discriminative ability of the unified embeddings, many strategies such as semantic concept learning (Huang et al. 2018; Shi et al. 2019) and region relationship reasoning (Li et al. 2019) are developed to enhance visual features by incorporating local region semantics. However, these approaches fail to capture the local interactions between image regions and sentence fragments, leading to limited interpretability and performance gains. To address this problem, Karpathy *et al.* (Karpathy and Li 2015) and Lee *et al.* (Lee et al. 2018) propose to discover all the possible alignments between image regions and sentence fragments, which produce impressive retrieval results and inspire a surge of works (Wang et al. 2019c; Hu et al. 2019; Zhang et al. 2020; Chen et al. 2020; Wehrmann, Kolling, and Barros 2020) to explore more accurate fine-grained correspondence. Although noticeable improvements have been made by designing various mechanisms to encode more powerful features or capture more accurate alignments, these approaches neglect the importance of similarity computation, which is the key to explore the complex matching patterns between image and text.

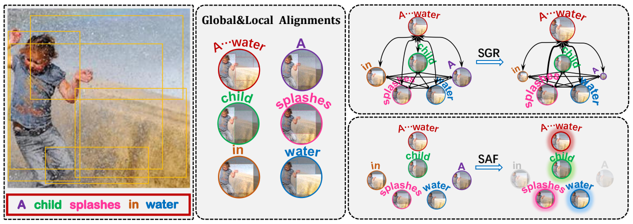


Figure 1: Illustration of the SGRAF. Nodes of red and other colors encode image-text and region-word alignments respectively. SGR module captures their relationships to achieve comprehensive similarity reasoning and SAF module reduces the interferences of less-meaningful alignments

To be more specific, there are three defects in previous approaches. Firstly, these methods compute scalar-based co-

*Corresponding author

sine similarities between local features, which may not be powerful enough to characterize the association patterns between regions and words. Secondly, most of them aggregate all the latent alignments between regions and words simply with max pooling (Karpathy and Li 2015) or average pooling (Lee et al. 2018; Chen et al. 2020), which hinders the information communication between local and global alignments, and thirdly, fails to consider the distractions of less-meaningful alignments, such as the alignments built with `"a"` and `"in"`, as shown in Figure 1.

To address these problems, in this paper we propose a novel Similarity Graph Reasoning and Attention Filtration (SGRAF) network for image-text matching. Specifically, we start with capturing the global alignments between the whole image and the full sentence, as well as the local alignments between image regions and sentence fragments. Instead of characterizing these alignments with scalar-based cosine similarity, we propose to learn the vector-based similarity representations to model the cross-modal associations more effectively. Then we introduce the Similarity Graph Reasoning (SGR) module, which relies on a Graph Convolution Neural Network (GCNN) to reason more accurate image-text similarity via capturing the relationship between local and global alignments. Furthermore, we develop the Similarity Attention Filtration (SAF) module to aggregate all the alignments attended by different significance scores, which reduces the interferences of non-meaningful alignments and achieves more accurate cross-modal matching results. Our main contributions are summarized as follows:

- We propose to learn the vector-based similarity representations for image-text matching, which enables greater capacity in characterizing the global alignments between images and sentences, as well as the local alignments between regions and words.
- We propose the Similarity Graph Reasoning (SGR) module to infer the image-text similarity with graph reasoning, which can identify more complex matching patterns and achieve more accurate predictions via capturing the relationship between local and global alignments.
- We attempt to consider the interferences of non-meaningful words in similarity aggregation, and propose an effective Similarity Attention Filtration (SAF) module to suppress the irrelevant interactions for further improving the matching accuracy.

Related Work

Image-Text Matching

Feature Encoding Many prior Approaches (Karpathy and Li 2015; Song and Soleymani 2019; Liu et al. 2017; Nam, Ha, and Kim 2017; Lee et al. 2018; Wang et al. 2019c; Li et al. 2019; Wang et al. 2020) focused on feature extraction and optimization for cross-modal retrieval. For textual features, Frome *et al.* (Frome et al. 2013) employed Skip-Gram (Mikolov et al. 2013) to extract word representations. Klein *et al.* (Klein et al. 2015) explored Fisher Vectors (FV) (Perronnin and Dance 2007) for text representation. Kiros *et al.* (Kiros, Salakhutdinov, and Zemel 2014) adopted a GRU as the text encoder. For visual features, Liu *et al.* (Liu et al. 2017) adapted Recurrent Residual networks to refine global embeddings. (Song and Soleymani 2019; Wei et al. 2020) employed multi-head self-attention to combine global context with locally-guided features. Besides, Some works (Nam, Ha, and Kim 2017; Ji et al. 2019) exploited block-based visual attention to gather semantics on feature maps, while (Lee et al. 2018; Wang et al. 2019c,b; Li et al. 2019; Wang et al. 2020; Chen and Luo 2020) followed (Anderson et al. 2018) to obtain region-based features of visual objects with the pre-trained model on Visual Genomes (Krishna et al. 2017). Especially, (Chen and Luo 2020) explored Bi-GRU to gain high-level object features, while (Li et al. 2019; Wang et al. 2020) proposed GCN-based networks to generate relationship-enhanced object features. We employ self-attention (Vaswani et al. 2017) on region or word features to get image or text representation. We concentrate on the similarity encoding mechanism that models global image-text and local region-word alignments comprehensively and fully encodes fine-grained relations between image and text.

Similarity Prediction Most existing works (Faghri et al. 2017; Wang, Li, and Lazebnik 2016; Zheng et al. 2017; Vendrov et al. 2016; Gu et al. 2018) for image-text matching learned the joint embedding and the similarity measures for cross-modal matching. For global alignments, some works (Faghri et al. 2017; Wang, Li, and Lazebnik 2016; Liu et al. 2017; Song and Soleymani 2019; Nam, Ha, and Kim 2017; Li et al. 2019) explored a joint space and calculated the inner product (e.g. cosine distance) for similarity computation. Others (Vendrov et al. 2016; Gu et al. 2018) introduced an ordered representations to measure antisymmetric visual-semantic hierarchy. For local alignments, most networks (Karpathy and Li 2015; Lee et al. 2018; Hu et al. 2019; Wang et al. 2019b; Chen et al. 2020) computed scalar-based alignments and adopted simple operation (e.g. sum and average) to fuse local alignments. For example, Lee *et al.* (Lee et al. 2018) studied the latent semantic alignments among region-words pairs and integrated local cosine alignments by average or LogSumExp. Differently, our network aggregates similarities by exploring global-local relationships among vector-based alignments and reducing the distraction from less-meaningful ones.

Graph Convolution Network

The researches based on Graph modeled the dependencies between concepts and facilitated graph reasoning such as GCNN (Duvenaud et al. 2015; Kipf and Welling 2017), and Gated Graph Neural Network (GGNN) (Li et al. 2016). These graph neural networks have been widely employed in various visual semantic tasks, such as image captioning (Yang et al. 2019), VQA (Teney, Liu, and van den Hengel 2017), and grounding referring expressions (Wang et al. 2019a). In recent years, there are several approaches to utilize graph structures to enhance single visual or textual features referring to image-text matching. Shi *et al.* (Shi et al. 2019) adopted Scene Concept Graph (SCG) by using image scene graphs and frequently co-occurred concept pairs as scene common-sense knowledge. Li *et al.* (Li et al. 2019) proposed Visual Semantic Reasoning to build up connections between image regions and generate visual representations with semantic relationships. Wang *et al.* (Wang et al. 2020) employed visual scene graph and textual scene graph, each of which separately refines visual and textual features including objects and relationships. They all focus on ”feature encoding” by learning single-modality contextualized representations, while our SGR targets at ”similarity reasoning” and explores more complex matching patterns with global and local cross-modal alignments.

Attention Mechanism

The attention mechanism has been applied to adaptively filter and aggregate information in natural language processing. When it comes to image-text matching, it has been intended to attend to certain parts of visual and textual data. (Lee et al. 2018; Wang et al. 2019b) developed Stacked Cross Attention to match latent alignments using both image regions and textual words as context. (Liu et al. 2019; Hu et al. 2019; Wang et al. 2019c) designed more complicated Cross Attentions to improve image-text matching. Chen *et al.* (Chen et al. 2020) proposed an Iterative Matching with Recurrent Attention Memory to explore fine-grained region-word correspondence progressively. We adopt textual-to-visual attention (Lee et al. 2018) with region-word pairs and calculate textual-attended alignments. In this paper, our SAF aims to discard less-semantic alignments instead of exploiting precise cross-modal attention.

Method

In this section, we focus on improving the visual-semantic similarity learning via capturing the relationship between local and global alignments, and suppressing the disturbance of less-meaningful alignments. As illustrated in Figure 2, we begin with introducing how to encode the visual and textual observations, and then compute the similarity representations of all local and global representation pairs. Afterwards, we elaborate on the proposed Similarity Graph Reasoning (SGR) module for relation-aware similarity reasoning and Similarity Attention Filtration (SAF) module for representative similarity aggregation. Finally, we present the detailed implementations of training objectives and inference strategies with both the SGR and SAF modules.

Generic Representation Extraction

Visual Representations. For each input image, we follow (Anderson et al. 2018) to extract K region-level visual features, with the Faster R-CNN (Ren et al. 2015) model pretrained on Visual Genomes (Krishna et al. 2017). We add a fully-connect layer to transform them into d-dimensional vectors as local region representations $\boldsymbol{V} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_K\}$, with $\boldsymbol{v}_i \in \mathbb{R}^d$. Afterwards, we perform self-attention mechanism (Vaswani et al. 2017) over the local regions, which adopts average feature $\bar{\boldsymbol{q}}_v = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{v}_i$ as the query and aggregates all the regions to obtain global representation $\bar{\boldsymbol{v}}$.

Textual Representations. Given a sentence, we split it into L words with tokenization technique, and sequentially feed the word embeddings into a bi-directional GRU (Schuster and Paliwal 1997). The representation of each word is then obtained by averaging the forward and backward hidden state at each time step, with $\boldsymbol{T} = \{\boldsymbol{t}_1, ..., \boldsymbol{t}_L\}$, and $\boldsymbol{t}_j \in \mathbb{R}^d$ denoting the representation of j-th word. Similarly, the global text representation $\bar{\boldsymbol{t}}$ is computed by the self-attention method over all the word features.

Similarity Representation Learning

Vector Similarity Function. Most previous methods utilize the cosine or Euclidean distance to represent the similarity between two feature vectors, which can capture the relevance to a certain degree while lacks the detailed correspondence. In this paper, we compute a similarity representation, which is a similarity vector instead of a similarity scalar, to capture more detailed associations between feature representations from different modalities. The similarity function between vector $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^d$ is defined as

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{W}) = \frac{\boldsymbol{W}|\boldsymbol{x} - \boldsymbol{y}|^2}{\left\| \boldsymbol{W}|\boldsymbol{x} - \boldsymbol{y}|^2 \right\|_2} \tag{1}$$

where $|\cdot|^2$ and $\|\cdot\|_2$ indicate element-wise square and ℓ_2-norm respectively, and $\boldsymbol{W} \in \mathbb{R}^{m \times d}$ is a learnable parameter matrix to obtain the m-dimensional similarity vector.

Global Similarity Representation. We compute the similarity representation between the global image feature $\bar{\boldsymbol{v}}$ and sentence features $\bar{\boldsymbol{t}}$ with Eq. (1),

$$\boldsymbol{s}^g = \boldsymbol{s}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{t}}; \boldsymbol{W_g}) \tag{2}$$

where $\boldsymbol{W}_g \in \mathbb{R}^{m \times d}$ aims to learn the global similarity representation.

Local Similarity Representation. To exploit local similarity representations between local features of visual and textual observations, we apply textual-to-visual attention (Lee et al. 2018) to attend on each region with respect to each word. Attention weight for each region is computed by

$$\alpha_{ij} = \frac{exp(\lambda \hat{c}_{ij})}{\sum_{i=1}^{K} exp(\lambda \hat{c}_{ij})} \tag{3}$$

Here the weight α_{ij} is calculated by the softmax function with a temperature parameter λ. c_{ij} indicates the cosine similarity between region feature $\boldsymbol{v}_i$ and word feature $\boldsymbol{t}_j$, $\hat{c}_{ij} = [c_{ij}]_+ / \sqrt{\sum_{j=1}^{L} [c_{ij}]_+^2}$ aims to normalize the cosine similarity matrix, and $[x]_+ = max(x, 0)$.

Then we generate the attended visual features $\boldsymbol{a}_j^v$ with respect to j-th word by

$$\boldsymbol{a}_j^v = \sum_{i=1}^{K} \alpha_{ij} \boldsymbol{v}_i, \tag{4}$$

and finally we compute the local similarity representation between $\boldsymbol{a}_j^v$ and $\boldsymbol{t}_j$ as

$$\boldsymbol{s}_j^l = \boldsymbol{s}(\boldsymbol{a}_j^v, \boldsymbol{t}_j; \boldsymbol{W}_l) \tag{5}$$

where $\boldsymbol{W}_l \in \mathbb{R}^{m \times d}$ is also a learnable parameter matrix. The local similarity representations capture the associations between a specific word and its corresponding image regions, which exploit more fine-grained visual-semantic alignments to boost the similarity prediction.

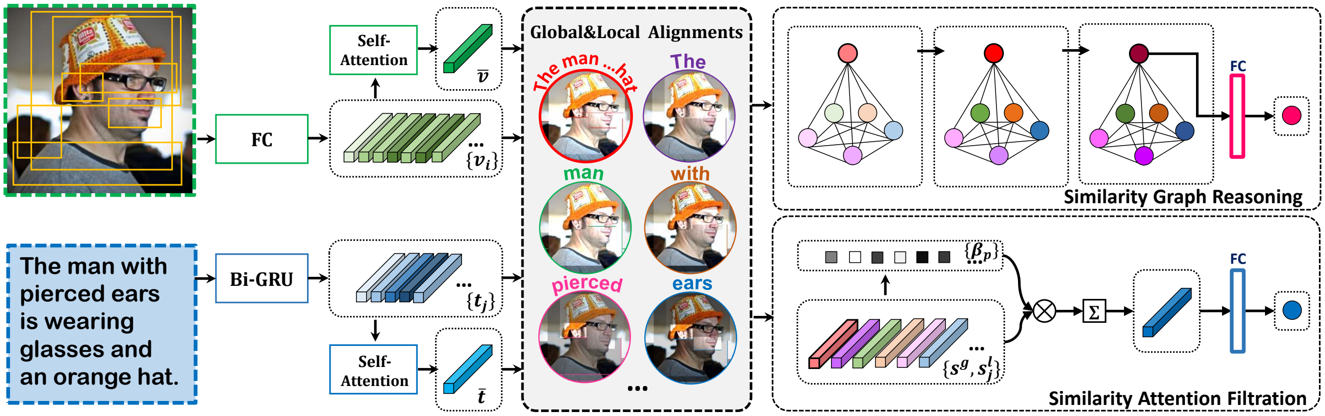


Figure 2: The proposed SGRAF network for image-text matching. The image and sentence are firstly encoded into local and global feature representations, followed by a similarity representation computation module to capture the correspondence between all local and global cross-modal pairs. The Similarity Graph Reasoning (SGR) module reasons the similarity with giving consideration to the relationship between all the alignments, and the Similarity Attention Filtration (SAF) module attends to more informative alignments for more accurate similarity prediction

Similarity Graph Reasoning

Graph Building. To achieve more comprehensive similarity reasoning, we build a similarity graph to propagate similarity messages among the possible alignments at both local and global levels. More specifically, we take all the word-attended similarity representations and the global similarity representation as graph nodes, i.e. $\mathcal{N} = \{\boldsymbol{s}_1^l, ..., \boldsymbol{s}_L^l, \boldsymbol{s}^g\}$, and follow (Kuang et al. 2019) to compute the edge from node $\boldsymbol{s}_q \in \mathcal{N}$ to $\boldsymbol{s}_p \in \mathcal{N}$ as

$$e(\boldsymbol{s}_p, \boldsymbol{s}_q; \boldsymbol{W}_{in}, \boldsymbol{W}_{out}) = \frac{exp((\boldsymbol{W}_{in}\boldsymbol{s}_p)(\boldsymbol{W}_{out}\boldsymbol{s}_q))}{\sum_q exp((\boldsymbol{W}_{in}\boldsymbol{s}_p)(\boldsymbol{W}_{out}\boldsymbol{s}_q))}, \tag{6}$$

where $\boldsymbol{W}_{in} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{W}_{out} \in \mathbb{R}^{m\times m}$ are the linear transformations for incoming and outgoing nodes, respectively. Note that the edges between node $\boldsymbol{s}_p$ and $\boldsymbol{s}_q$ are directed, which allow efficient and complex information propagation for similarity reasoning.

Graph Reasoning. With the constructed graph nodes and edges, we perform similarity graph reasoning by updating the nodes and edges with

$$\hat{\boldsymbol{s}}_p^n = \sum_q e(\boldsymbol{s}_p^n, \boldsymbol{s}_q^n; \boldsymbol{W}_{in}^n, \boldsymbol{W}_{out}^n) \cdot \boldsymbol{s}_q^n \tag{7}$$

$$\boldsymbol{s}_p^{n+1} = ReLU(\boldsymbol{W}_r^n \hat{\boldsymbol{s}}_p^n) \tag{8}$$

with $\boldsymbol{s}_p^0$ and $\boldsymbol{s}_q^0$ taken from $\mathcal{N}$ at step $n = 0$, and $\boldsymbol{W}_r^n$, $\boldsymbol{W}_{in}^n$, $\boldsymbol{W}_{out}^n$ are learnable parameters in each step. After current step of graph reasoning, the node $\boldsymbol{s}_p^n$ is replaced with $\boldsymbol{s}_p^{n+1}$.

We iteratively reason the similarity for N steps, and take the output of the global node at the last step as the reasoned similarity representation, and then feed it into a fully-connect layer to infer the final similarity score. The SGR module enables the information propagation between local and global alignments, which can capture more comprehensive interactions to facilitate the similarity prediction.

Similarity Attention Filtration

Although the exploitation of local alignments can boost the matching performance via discovering more fine-grained correspondence between image regions and sentence fragments, we notice that the less-meaningful alignments hinder the distinguishing ability when aggregating all the possible alignments in an undifferentiated way. Therefore we propose a Similarity Attention Filtration (SAF) module to enhance important alignments, as well as suppress ineffectual alignments, such as the alignments with `"the"`, `"be"` and etc.

Given the local and global similarity representations, we calculate an aggregation weight β_p for each similarity representation $\boldsymbol{s}_p \in \mathcal{N}$ by

$$\beta_p = \frac{\delta(BN(\boldsymbol{W}_f \boldsymbol{s}_p))}{\sum_{\boldsymbol{s}_q \in \mathcal{N}} \delta(BN(\boldsymbol{W}_f \boldsymbol{s}_q))} \tag{9}$$

where $\delta(\cdot)$ is the Sigmoid function, BN indicates the batch normalization, and $\boldsymbol{W}_f \in \mathbb{R}^{m\times 1}$ is a linear transformation.

Then we aggregate the similarity representations with $\boldsymbol{s}_f = \sum_{\boldsymbol{s}_p \in \mathcal{N}} \beta_p \boldsymbol{s}_p$, and feed $\boldsymbol{s}_f$ into a fully-connect layer to predict the final similarity between the input image and sentence. The SAF module learns the significance scores to increase the contribution of more-informative similarity representations and meanwhile reduce the disturbance of less-meaningful alignments.

Training Objectives and Inference Strategies

We utilize the bidirectional ranking loss (Faghri et al. 2017) to train both the SGR and SAF modules. Given a matched image-text pair $(\boldsymbol{v}, \boldsymbol{t})$, and the corresponding hardest negative image $\boldsymbol{v}^-$ and the hardest negative text $\boldsymbol{t}^-$ within a minibatch, we compute the bidirectional ranking loss with

$$\begin{aligned}\mathcal{L}_r(\boldsymbol{v}, \boldsymbol{t}) = &[\gamma - \mathcal{S}_r(\boldsymbol{v}, \boldsymbol{t}) + \mathcal{S}_r(\boldsymbol{v}, \boldsymbol{t}^-)]_+ \\ &+[\gamma - \mathcal{S}_r(\boldsymbol{v}, \boldsymbol{t}) + \mathcal{S}_r(\boldsymbol{v}^-, \boldsymbol{t})]_+\end{aligned} \tag{10}$$

Methods	MSCOCO dataset						Flickr30K dataset					
	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
CAMP (Wang et al. 2019c)	72.3	94.8	98.3	58.5	87.9	95.0	68.1	89.7	95.2	51.5	77.1	85.3
SCAN (Lee et al. 2018)	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
SGM (Wang et al. 2020)	73.4	93.8	97.8	57.5	87.3	94.3	71.8	91.7	95.5	53.5	79.6	86.5
VSRN* (Li et al. 2019)	74.0	94.3	97.8	60.8	88.4	94.1	70.4	89.2	93.7	53.0	77.9	85.7
RDAN (Hu et al. 2019)	74.6	96.2	98.7	61.6	89.2	94.7	68.1	91.0	95.9	54.1	80.9	87.2
MMCA (Wei et al. 2020)	74.8	95.6	97.7	61.6	89.8	95.2	74.2	92.8	96.4	54.8	81.4	87.8
BFAN (Liu et al. 2019)	74.9	95.2	-	59.4	88.4	-	68.1	91.4	-	50.8	78.4	-
CAAN (Zhang et al. 2020)	75.5	95.4	98.5	61.3	89.7	95.2	70.1	91.6	97.2	52.8	79.0	87.9
DPRNN (Chen and Luo 2020)	75.3	95.8	98.6	62.5	89.7	95.1	70.2	91.6	95.8	55.5	81.3	88.2
PFAN (Wang et al. 2019b)	76.5	**96.3**	**99.0**	61.6	89.6	95.2	70.0	91.8	95.0	50.4	78.7	86.1
VSRN (Li et al. 2019)	76.2	94.8	98.2	62.8	89.7	95.1	71.3	90.6	96.0	54.7	81.8	88.2
IMRAM (Chen et al. 2020)	76.7	95.6	98.5	61.7	89.1	95.0	74.1	93.0	96.6	53.9	79.4	87.2
Ours(SAF)	76.1	95.4	98.3	61.8	89.4	95.3	73.7	93.3	96.3	56.1	81.5	88.0
Ours(SGR)	78.0	95.8	98.2	61.4	89.3	95.4	75.2	93.3	96.6	56.2	81.0	86.5
Ours(SGRAF)	**79.6**	96.2	98.5	**63.2**	**90.7**	**96.1**	**77.8**	**94.1**	**97.4**	**58.5**	**83.0**	**88.8**

Table 1: Comparison of bi-directional retrieval results (R@K(%)) on MSCOCO 1K test set and Flickr30K test set. VSRN* denotes a single model for a fair comparison with SGR. SGRAF denotes the whole framework with independent training

where γ is the margin parameter and $\mathcal{S}_r(\cdot,\cdot)$ indicates similarity prediction function implemented with SGR. Similarly, we define the training objectives on SAF module as $\mathcal{L}_f$.

In this paper, we explore different training and inference strategies with the proposed SGR and SAF modules: joint training and independent training. For joint training, we combine $\mathcal{L}_r$ and $\mathcal{L}_f$ to train SGR and SAF modules simultaneously, where the similarity representations are shared for the proposed two modules. For independent training, we train the SGR and SAF modules separately. At the inference stage, we average the similarities predicted by SGR and SAF modules for the retrieval evaluation.

Methods	Sen. Ret.		Ima. Ret.	
	R@1	R@10	R@1	R@10
SGM (Wang et al. 2020)	50.0	87.9	35.3	76.5
CAMP (Wang et al. 2019c)	50.1	89.7	39.0	80.2
VSRN* (Li et al. 2019)	50.3	87.9	37.9	79.4
SCAN (Lee et al. 2018)	50.4	90.0	38.6	80.4
CAAN (Zhang et al. 2020)	52.5	90.9	41.2	**82.9**
VSRN (Li et al. 2019)	53.0	89.4	40.5	81.1
IMRAM (Chen et al. 2020)	53.7	91.0	39.7	79.8
MMCA (Wei et al. 2020)	54.0	90.7	38.7	80.8
Ours(SAF)	53.3	90.1	39.8	80.2
Ours(SGR)	56.9	90.5	40.2	79.8
Ours(SGRAF)	**57.8**	**91.6**	**41.9**	81.3

Table 2: Comparison of bi-directional retrieval results (R@K(%)) on MSCOCO 5K test set

Experiments

To verify the effectiveness of the our model, in this section we demonstrate extensive experiments on two benchmark datasets. We also introduce detailed implementations and training strategy of the proposed SGRAF model.

Datasets and Settings

Datasets. We evaluate our model on the MSCOCO (Lin et al. 2014) and Flickr30K (Young et al. 2014) datasets. The MSCOCO dataset contains 123,287 images, and each image is annotated with 5 annotated captions. The dataset is split into 113,287 images for training, 5000 images for validation and 5000 images for testing. We report results by averaging over 5 folds of 1K test images and testing on the full 5K images. The Flickr30K dataset contains 31,783 images with 5 corresponding captions each. Following the split in (Frome et al. 2013), we use 1,000 images for validation, 1,000 images for testing and the rest for training.

Protocols. For image-text retrieval, we measure the performance by Recall at K (R@K) defined as the proportion of queries whose ground-truth is ranked within the top K. We adopt R@1, R@5 and R@10 as our evaluation metrics.

Implementation Details. For each image, we take the Faster-RCNN (Ren et al. 2015) detector with ResNet-101 provided by (Anderson et al. 2018) to extract the top $K = 36$ region proposals and obtain a 2048-dimensional feature for each region. For each sentence, we set the word embedding size as 300, and the number of hidden states as 1024. The dimension of similarity representation m is 256, with smooth temperature $\lambda = 9$, reasoning steps $N = 3$, and margin $\gamma = 0.2$. Our model employs the Adam optimizer (Kingma and Ba 2015) to train the SGRAF network with the mini-batch size of 128. The learning rate is set to be 0.0002 for the first 10 epochs and 0.00002 for the next 10 epochs on

model	GLO	LOC	Step				Sen. Ret.		Ima. Ret.	
			1	2	3	4	R@1	R@10	R@1	R@10
1	✓						62.4	92.6	46.0	83.1
2		✓	✓				71.8	95.6	52.1	82.3
3		✓			✓		73.6	96.1	54.3	85.1
4	✓	✓	✓				74.2	96.3	55.5	86.0
5	✓	✓		✓			75.3	**96.7**	56.0	85.9
6	✓	✓			✓		75.2	96.6	**56.2**	**86.5**
7	✓	✓				✓	**76.2**	96.3	55.0	86.1

Table 3: The impact of SGR configurations. GLO and LOC respectively indicates the employment of global and local alignments, and Step denotes the graph reasoning steps

model	I2T	T2I	SS	SV	AA	SGR	SAF	Sen. Ret.		Ima. Ret.	
								R@1	R@10	R@1	R@10
1	✓		✓		✓			66.7	94.1	43.2	82.3
2	✓			✓	✓			67.2	94.8	47.6	83.1
3	✓			✓		✓		66.1	94.1	45.6	81.6
4	✓			✓			✓	**68.2**	**95.1**	**49.8**	**85.1**
5		✓	✓		✓			62.6	93.6	45.3	82.4
6		✓		✓	✓			65.2	95.1	49.5	83.5
7		✓		✓		✓		**73.6**	96.1	54.3	85.1
8		✓		✓			✓	72.9	**96.3**	**55.7**	**87.8**

Table 4: The impact of Similarity configurations. I2T and T2I denotes the visual-to-textual and textual-to-visual attention to generate local similarity representations separately. SS denotes the scalar-based cosine similarity and SV indicates the vector-based similarity, and AA represents the average aggregation of all alignments

Dataset	SAF	SGR	Joint	Split	Sen. Ret.		Ima. Ret.	
					R@1	R@10	R@1	R@10
MSCOCO	✓				76.1	98.3	61.8	95.3
		✓			78.0	98.2	61.4	95.4
	✓	✓	✓		77.8	98.2	61.6	95.2
	✓	✓		✓	**79.6**	**98.5**	**63.2**	**96.1**
Flickr30K	✓				73.7	96.3	56.1	88.0
		✓			75.2	96.6	56.2	86.5
	✓	✓	✓		75.1	96.1	56.2	85.8
	✓	✓		✓	**77.8**	**97.4**	**58.5**	**88.8**

Table 5: The impact of Training configurations on MSCOCO 1K test set and Flickr30K test set. Split and Joint denotes independent and joint training of two modules

MSCOCO. For Flickr30K, we start training the SGR (SAF) module with learning rate 0.0002 for 30 (20) epochs and decay it by 0.1 for the next 10 epochs. We select the snapshot with the best performance on the validation set for testing.

Quatitative Results and Analysis

In this section, we present the retrieval results on the MSCOCO and Flickr30K datasets, aiming to demonstrate the effectiveness and superiority of the proposed approach.

Comparisons on MSCOCO. Table 1 and 2 report the experimental results on MSCOCO dataset with 1K and 5K test images, separately. We can see that our proposed SGRAF model outperforms the existing methods, with the best R@1=79.6% for sentence retrieval and R@1=63.2% for image retrieval with 1K test images. For 5K test images, the proposed approach maintains the superiority with an improvement of more than 3% on the R@1 results. It should be noted that competitive retrieval performance can be also achieved with the SGR/SAF module alone, demonstrating the effectiveness and complementarity of our modules.

Comparisons on Flickr30K. Table 1 compares the bidirectional retrieval results on Flickr30K dataset with the latest algorithms. We can observe that the SAF module alone produces comparable retrieval results and the SGR module achieves state-of-the-art performance with R@1 of 75.2% and 56.2% for sentence and image retrieval, separately. This verifies the effectiveness of exploiting the relationship between alignments to boost similarity reasoning. When we combine the SAF and SGR module, the performance is further improved to achieve the best R@1 of 77.8% and 58.5%.

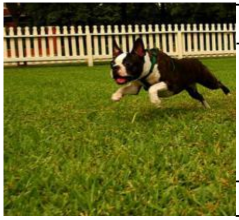
Ablation Studies

In this section, we carry a series of ablation studies to explore the impact of different configurations for the SGR module, the similarity representation learning module and the process of training. We also compare different strategies of similarity prediction to demonstrate the superiority of SGR and SAF modules. All the comparative experiments are conducted on the Flickr30K dataset.

Configurations of SGR module. In Table 3 we investigate the effectiveness of each component in the SGR module. 1) Graph reasoning. We employ a framework without graph reasoning as the baseline(#1), which adopts a fully-connected layer and sigmoid function on the global alignment to obtain the final similarity. Comparing #1 and #6 based on R@1, the SGR module achieves 12.8% improvement for sentence retrieval and 10.2% for image retrieval. 2) Reasoning steps setting. Comparing #4, #5, #6 and #7, we set the step of the SGR module to 3 for maximum performance. 3) Global and local alignments. #2 and #3 only utilize local alignments for graph reasoning and adopt a mean-pooling operation on them after reasoning. Comparing #2, #4 and #3, #6, we discover that global similarity is beneficial for aggregating local similarities and exploring their relations which improves at least 1.6% for sentence retrieval and 1.9% for image retrieval on R@1.

Configurations for Similarity Computation. Table 4 illustrates the impact of different strategies in similarity representation computation and the similarity score prediction. We test the results on local alignments and set the reasoning step of the SGR module to 3. we following(Lee et al. 2018) to explore two types of the cross-attention modes, i.e. I2T and T2I. Comparing #1, #2, #5 and #6, we find that averaging the local alignments calculated by a fully-connected layer and sigmoid function leads to better performance than

Query:



Positive	Local alignments												Global
Caption	A	dog	runs	on	the	green	grass	near	a	wooden	fence	.	---
SAF β	0.0	0.2	0.04	0.01	0.01	0.07	0.06	0.01	0.0	0.2	0.14	0.0	0.2
SGR α	0.37	0.37	0.39	0.37	0.46	0.48	0.53	0.46	0.3	0.29	0.33	0.3	0.18
cosine	0.0	0.9	0.8	0.3	0.4	0.7	0.7	0.3	0.0	0.8	0.98	0.0	0.2
Final sim			AVE score:0.54				SAF score:0.89				SGR score: 0.92		

Negative	Local alignments															Global
Caption	A	brown	dog	with	white	paws	is	trotting	through	a	field	of	green	grass	.	---
SAF β	0.0	0.12	0.3	0.01	0.05	0.04	0	0.1	.01	0	.05	0	0.09	0.07	0	0.2
SGR α	0.0	0.35	0.31	0.03	0.30	0.22	0.13	0.45	0.29	0.29	0.46	0.46	0.46	0.46	0.3	0.0
cosine	0.1	0.1	0.6	0.4	0.8	0.7	0.2	0.8	0.7	0.8	0.8	0.6	0.8	0.8	0.3	0.2
Final sim			AVE score:0.56						SAF score:0.54					SGR score:0.38		

Figure 3: The visualization of SAF and SGR module. Positive and Negative denotes ground-truth and hard negative examples respectively. SAF β denotes attention weight distribution of SAF module. SGR α denotes the cosine distance between final alignment and raw alignments. Final sim denotes similarity calculated by AVE (average), SAF or SGR module

averaging local cosine distance. Comparing #3 and #7, it is more reasonable for the SGR module to count on the local alignments attended by word features (T2I) than the ones by region features (I2T). Besides, the SGR module fails to achieve significant improvement on I2T which indicates that the region features are redundant, relatively independent and irregular in order. Therefore, it is difficult for the SGR module to exploit semantic connections compared with word features. In terms of #4 and #8, the SAF module achieves impressive progress both in I2T and T2I modes that demonstrates that the SAF module filters and aggregates plenty of discriminative local alignments steadily to improve the precision of image-text matching.

Configurations for Training Process. In table 5, we report the results of different training strategies: joint learning and independent learning. Compared with the SGR/SAF module alone, joint learning can help the SAF module improve the performance of sentence retrieval, and also help the SGR module enhance the ability of image retrieval. In terms of independent learning, the SGRAF network gains an exact and impressive promotion. We assume that the SGR module frequently captures several crucial cues by propagating information between local and global alignments and throws out some relatively unimportant interactions. Moreover, the SAF module attempts to gather all the meaningful alignments and eliminates completely irrelevant interactions. Therefore, the global and local alignments for the SAF and SGR modules are seemingly not incompatible resulting in the unobvious improvement. It is worth noting that the SAF module tends to be more susceptible to the hard negative samples than the SGR module because of the high correlation. On the other hand, it is more challenging for the SGR module to resolve the transmission and integration of numerous semantic alignments. As a result, they can cooperate with each other and further achieve more accurate similarity prediction through independent training.

Qualitative Results and Analysis

As it is shown in Figure 3, we illustrate the distribution of attention weights learned by the SAF module. Given an image query, the SAF module captures the key cues (`"dog runs"`, `"green grass"`, `"wooden fence"`) for positive image-text pairs, and also highlights the meaningful instances (`"brown dog"`, `"white paws"`, `"trotting"`, `"green grass"`) for negative pairs. Note that there exists a crucial discrepancy (`"brown"`) which is submerged by AVE operation between negative text and image that depicts a black and white dog. Compared with the wrong matching of AVE, SAF module can stress on all the useful alignments including unmatched instance (`"brown"`) and suppress irrelevant interactions (`"of"`, `"with"`, `"is"`, and etc). On the other hand, the process of SGR module reinforces the role of the alignment (`"brown"`), which leads to lower similarity between hard negative and query image. Our implementation of this paper is publicly available on GitHub at: https://github.com/Paranioar/SGRAF.

Conclusion

In this work, we present a SGRAF network consisting of similarity graph reasoning (SGR) and similarity attention filtration (SAF) module. The SGR module performs multi-step reasoning based on global and local similarity nodes and captures their relations through information propagation, while the SAF module attends more to discriminative and meaningful alignments for similarity aggregation. We demonstrate that it is important to exploit the relationship between local and global alignments, and suppress the disturbances of less-meaningful alignments. Extensive experiments on benchmark datasets show that both SGR and SAF modules can effectively discover the associations between image and text and achieve further improvements when cooperating with each other.

Acknowledgments

The paper is supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102001 and National Natural Science Foundation of China under Grant No. 61725202, U1903215, 61829102, 91538201, 61771088, 61751212 and the Fundamental Research Funds for the Central Universities under Grant No. DUT19GJ201 and Dalian Innovation Leader's Support Plan under Grant No. 2018RD07.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. IMRAM: Iterative Matching with Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *CVPR*, 12655–12663.

Chen, T.; and Luo, J. 2020. Expressing Objects Just Like Words: Recurrent Visual Embedding for Image-Text Matching. In *AAAI*, 10583–10590.

Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *NIPS*, 2224–2232.

Faghri, F.; Fleet, D. J.; Kiros, R.; and Fidler, S. 2017. VSE++: Improved Visual-Semantic Embeddings. *arXiv: 1707.05612* .

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*, 2121–2129.

Gu, J.; Cai, J.; Joty, S. R.; Niu, L.; and Wang, G. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models. In *CVPR*, 7181–7189.

Hu, Z.; Luo, Y.; Lin, J.; Yan, Y.; and Chen, J. 2019. Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching. In *IJCAI*, 789–795.

Huang, Y.; Wu, Q.; Song, C.; and Wang, L. 2018. Learning Semantic Concepts and Order for Image and Sentence Matching. In *CVPR*, 6163–6171.

Ji, Z.; Wang, H.; Han, J.; and Pang, Y. 2019. Saliency-Guided Attention Network for Image-Sentence Matching. In *ICCV*, 5753–5762.

Karpathy, A.; and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv: 1411.2539* .

Klein, B.; Lev, G.; Sadeh, G.; and Wolf, L. 2015. Associating neural word embeddings with deep image representations using Fisher Vectors. In *CVPR*, 4437–4446.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV* 123(1): 32–73.

Kuang, Z.; Gao, Y.; Li, G.; Luo, P.; Chen, Y.; Lin, L.; and Zhang, W. 2019. Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid. In *ICCV*.

Lee, K.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked Cross Attention for Image-Text Matching. In *ECCV*, 212–228.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual Semantic Reasoning for Image-Text Matching. In *ICCV*, 4653–4661.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. S. 2016. Gated Graph Sequence Neural Networks. In *ICLR*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Liu, C.; Mao, Z.; Liu, A.; Zhang, T.; Wang, B.; and Zhang, Y. 2019. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In *ACMMM*, 3–11.

Liu, Y.; Guo, Y.; Bakker, E. M.; and Lew, M. S. 2017. Learning a Recurrent Residual Fusion Network for Multimodal Matching. In *ICCV*, 4127–4136.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Nam, H.; Ha, J.; and Kim, J. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *CVPR*, 2156–2164.

Perronnin, F.; and Dance, C. R. 2007. Fisher Kernels on Visual Vocabularies for Image Categorization. In *CVPR*.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 91–99.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *TSP* 45(11): 2673–2681.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, 5182–5189.

Song, Y.; and Soleymani, M. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *CVPR*, 1979–1988.

Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-Structured Representations for Visual Question Answering. In *CVPR*, 3233–3241.

Toyama, J.; Misono, M.; Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2017. Neural Machine Translation with Latent Semantic of Image and Text. In *ICLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2016. Order-Embeddings of Images and Language. In *ICLR*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*, 5005–5013.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2019a. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. In *CVPR*, 1960–1968.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. In *WACV*, 1497–1506.

Wang, Y.; Yang, H.; Qian, X.; Ma, L.; Lu, J.; Li, B.; and Fan, X. 2019b. Position Focused Attention Network for Image-Text Matching. In *IJCAI*, 3792–3798.

Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; and Shao, J. 2019c. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *ICCV*, 5763–5772.

Wehrmann, J.; Kolling, C.; and Barros, R. C. 2020. Adaptive Cross-Modal Embeddings for Image-Text Alignment. In *AAAI*, 12313–12320.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *CVPR*, 10941–10950.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *CVPR*, 1316–1324.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 10685–10694.

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. *TACL* 2: 67–78.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020. Context-Aware Attention Network for Image-Text Retrieval. In *CVPR*, 3536–3545.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; and Shen, Y. 2017. Dual-Path Convolutional Image-Text Embedding. *arXiv: 1711.05535* .