Adapted Graph Reasoning and Filtration for Description-Image Retrieval

Shiqian Chen[1], Zhiling Luo[1,*], Yingqi Gao[1], Wei Zhou[1], Chenliang Li[2], Haiqing Chen[1]
[1]Alibaba Group, China
{shiqian.csq,godot.lzl,gaoyingqi.gyq,fayi.zw,haiqing.chenhq}@alibaba-inc.com
[2]School of Cyber Science and Engineering, Wuhan University, China
cllee@whu.edu.cn

ABSTRACT

Due to the significant cognition reduction, multi-media content has become an increasingly important information type nowadays. More and more descriptions are coupled with images to make them more attractive and persuasive. Currently, several text-image retrieval methods have been developed to improve the efficiency of the time-consuming and professional process. However, in practical retrieval applications, it is the vivid and terse descriptions that are widely used, instead of the shallow captions that describe what is contained. Therefore, the most existing methods designed for the caption-style text can not achieve this purpose. To eliminate the mismatch, we introduce a novel problem about description-image retrieval and propose the specially designed method, named Adapted Graph Reasoning and Filtration (AGRF). In AGRF, we firstly leverage an adapted graph reasoning network to discover the combination of visual objects in the image. Then, a cross-modal gate mechanism is proposed to cast aside those description-independent combinations. Experiment results on the real-world dataset demonstrate the advantages of the AGRF over the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Specialized information retrieval**; **Multimedia and multimodal retrieval**;

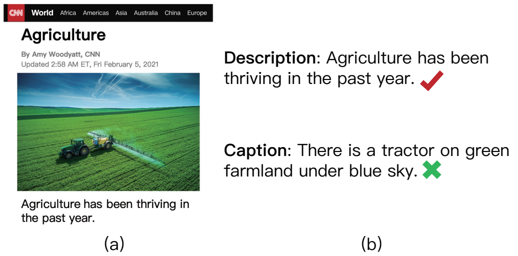
KEYWORDS

Multimedia Retrieval, Graph Neural Network, Gate Mechanism

ACM Reference Format:
Shiqian Chen[1], Zhiling Luo[1,*], Yingqi Gao[1], Wei Zhou[1], Chenliang Li[2], Haiqing Chen[1]. 2021. Adapted Graph Reasoning and Filtration for Description-Image Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463047

1 INTRODUCTION

In either social network (e.g., Facebook.com), and traditional public media (e.g., CNN), the multi-media content is more easily and widely propagated, compared with plain text. As illustrated in Fig.1 (a), A snippet of agricultural news with a well-written description and a well-selected image, is more attractive and easier to understand. As depicted in the Statistical Report on China’s Internet development, users prefer to browse content with visual elements, such as icons, images, and videos. Due to this increasing trend and the accessibility of fruitful image resources, more and more content creators tend to pair descriptions with images to grab the user’s attention and facilitate understanding. However, the process of selecting the perfect match image for a given description is time-consuming and requires lots of domain knowledge, since content creators have to browse a lot of images and take the complex object, scene, imagery into consideration.



Agriculture has been thriving in the past year.

Description: Agriculture has been thriving in the past year. ✓

Caption: There is a tractor on green farmland under blue sky. ✗

Figure 1: (a). A snippet of agricultural news. (b). the comparison between caption and description.

To alleviate the problem of image selection, several approaches have been developed on *text-image retrieval* [2, 5, 11]. Nevertheless, these approaches focus on the text type of image caption, as shown in Fig.1 (b), with low utilities in real application scenarios. Compared with the description, the image caption merely describes the shallow and apparent objects and appearances (e.g., "green farmland"), which are sometimes meaningless. Different from image caption, the description has the following two obvious characteristics: (1) A description often convey a moral or a profound meaning reflected by an image as a whole. For example, description uses "agricultural" to imply the tractors, the farmland and their combination together. (2) Each specific object contained in the image is not explicitly mentioned by a description. Considering these characteristics, there are two main challenges for description-image retrieval. Firstly, reasoning about the relationship of multiple image objects and their combinations. Secondly, casting aside the interference of description-independent detailed information in the image.

To address these challenges, which are merely discussed in previous studies on text-image retrieval, we propose a novel *Adapted Graph Reasoning and Filtration (AGRF)* network. It contains two major components for above two challenges respectively. At first, the *Adapted Graph Reasoning network (AGR)* , constructs an object

SIGIR ’21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463047

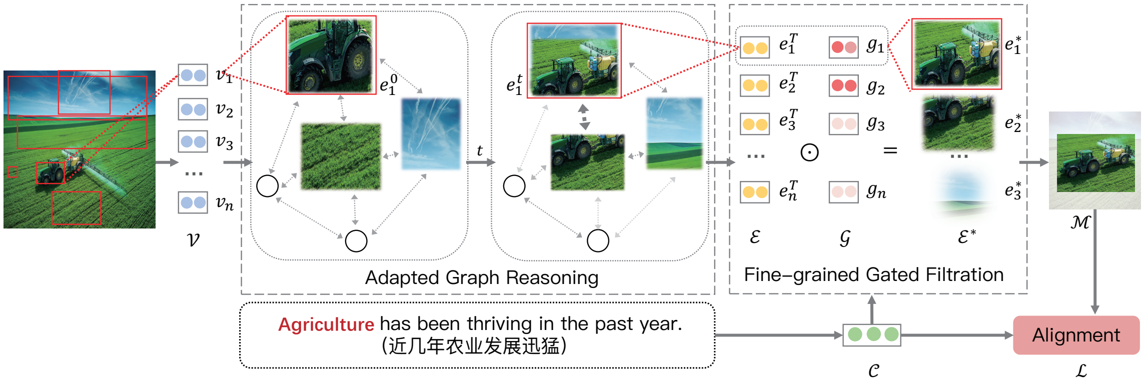


Figure 2: The overview of the major components and workflow of AGRF. Detected from the image, v_1 represents the *tractor*, v_2 the *farmland* and v_3 the *sky*. In AGR, the graph is initialized by $e^0 = v$ for $v_{1,2,3}$, and updated by t-steps reasoning. After that, e_1^t contains not only the feature of *tractor* but also the joint representation of object combination (*tractor,farmland*). Intuitively it represents *tractor on the farmland*. Also e_3^t represents *farmland under the sky*. In GF, gate g_1 is close to 1 due to the semantic similarity between *agriculture* from C and tractor on the farmland from e_1^t. However e_3^t cannot be aligned to C and results in g_3 closing to 0. In this way, most features from e_1 and few from e_3 are kept in M. Thus in alignment part, the similarity score between *tractor on the farmland* and *agriculture* is high. In other words, with the help of AGR and GF, we implicitly associate the concept *agriculture* to the combination (*tractor,farmland*).

relationship graph on the image and takes a step-by-step reasoning on both nodes and edges. The nodes represent the objects by the region features and the edges represent the correlations. The whole image representation is derived by reasoning. Then, a *Cross-Modal Fine-grained Gated Filtration (GF)* is introduced to utilize an element-level cross-modal gate mechanism to filter out the description-independent combinations and the fine-grained noises in each combinations, in order to suppress the ineffectual alignments.

In summary, the contribution of this paper is two-fold: (1) We introduce a novel problem of description-image retrieval and collect a new real-world dataset, VCG-DI. (2) We propose an Adapted Graph Reasoning and Filtration (AGRF) method which simultaneously extract the underlying semantics composited by different objects in an image jointly and perform description-independent information pruning for description-image retrieval.

2 THE PROPOSED SOLUTION

Description-Image Retrieval Problem. Compared with caption, description prefers to use one or few words to represent the key concepts, instead of the long and detailed phrases. Consider the example in Fig. 1, the previous method gives a high score to the sentence caption (C1) but a low score to description (D1). Because the objects, "tractor",and "farmland", are easily detected from the image and aligned to these corresponding tokens in C1. But these objects fail to be aligned with D1 due to the missing tokens. In this way, the problem of description-image retrieval is deduced by the proving on following necessary condition:

Definition 2.1 (Object Combination). Given an image $\mathcal{M}$ and a description C, there is a group of combination $comb \subset \mathbb{P}(\mathbf{obj})$ that can be aligned to C, where **obj** denotes the visual objects detected from $\mathcal{M}$ and $\mathbb{P}$ is the power set on **obj**.

Once above condition is proved, or satisfied for any pair $< \mathcal{M}, C >$, it comes a high similarity score. Note that the upbound of the volume of the combination is $O(2^{|\mathbf{obj}|})$ which is usually too big for calculation. Therefore the supervised sign-driven prune strategy is required. To summarize, we specify what we need to add in the normal text-image retrieval method such that it is more suit for this task: **combination reasoning** to find out *comb* and **gated filtration** to prune unnecessary combinations.

Method overview. The overall structure of AGRF is illustrated in Fig.2. Given an image, we extract the visual representation $\mathcal{V}$ on visual regions $\mathcal{R}$. And then, AGR takes step-by-step reasoning on the graph whose nodes represent $\mathcal{R}$ and edges represent their correlation $\mathcal{A}$. It derives the aggregated node representation $\mathcal{E}$. After that, GF computes the gate $\mathcal{G}$ concerning description representation C and $\mathcal{E}$. And the whole image representation $\mathcal{M}$ is conducted by filtering $\mathcal{G}$ on $\mathcal{E}$. At last, we derive the similarity score on $\mathcal{M}$ and C.

2.1 Generic Representation Extraction

Visual Representation. Taking advantage of the bottom-up attention[1], we represent each image by a set of features $\mathcal{V} = \{v_1, ..., v_n\}, v_i \in \mathbb{R}^K$, each of that feature v_i encodes a region r_i in this image. And K is the features' dimension, set up as a hyper-parameter. Following the previous methods[1, 10], we implement the bottom-up attention with a Faster R-CNN[13] model pre-trained on Visual Genomes[9]. The model uses greedy non-maximum suppression with an IoU threshold to select the top-n ($n = 36$) regions with the highest detection confidence scores, and extract the features f_i of selected regions by ResNet-101[7]. We further add a fully-connected layer to transform f_i into k-dimensional region representation v_i.

Textual Representation. About description processing, we use BERT[4] to map descriptions into deep continuous representations.

A given description is first tokenized into a token sequence according to WordPieces, and then the functional relationships of tokens are captured through the self-attention mechanism. Finally, the description representation $C \in \mathbb{R}^K$ is obtained after an additional fully-connected layer.

2.2 Adapted Graph Reasoning and Aggregation

As previously mentioned, the match between image and description comes from either the visual objects or their combination. Revisiting Fig.1, we witness that the keyword *agriculture*, is corresponds to the combination of *tractor* and *farmland*. Which can be found in the image. Intuitively, the alignment of image region and description is a not one-to-one mapping. Related image regions are often corresponding to the same description semantic concept. To discover the relationship of regions and aggregate the related semantic information from each region, we apply AGR to mimic human's step-by-step reasoning and aggregating behavior.

For a given image, we build up a complete graph G by taking all the detected regions $\mathcal{R}$ in this image as graph nodes. We consider that the information of every region to disseminate to neighbors should be different from step to step. Therefore, we dynamic measure the pairwise relevant coefficients between image regions at each reasoning step instead of fixing it as previous works. The relevant coefficient $a_{i,j}^t$ for region r_i and r_j in reasoning step t is computed by a shared attentional mechanism[14] as in following way:

$$z_i^t = W_a^t e_i^{t-1} + b^t \tag{1}$$

$$\theta_{i,j}^t = \sigma(\phi^T [z_i^t || z_j^t]) \tag{2}$$

where e_i^0 is setting as v_i, z_i^t is the candidate representation, $W_a^t \in \mathbb{R}^{K \times K}$ and $b^t \in \mathbb{R}^K$ are linear transformations and bias parameters shared by every region for reasoning step t. And $\phi \in \mathbb{R}^{2K}$ is a weight vector, $\sigma(\cdot)$ is the LeakyReLU nonlinearity, $||$ is the concatenation operation.

To make coefficients easier to compare, we use softmax function to normalize them.

$$a_{i,j}^t = softmax(\theta_{i,j}^t) \tag{3}$$

Here, the relevant coefficient $a_{i,j}^t$ indicate the proportion of semantic information that will be propagated from region i to j at current step t. Once obtained, the relevant coefficients $\mathcal{A}$ are used to aggregate the semantic information received from neighbors of each region as follow:

$$e_i^t = \sigma(\sum_{i=k}^{N} a_{k,i}^t z_k^t) + e_i^{t-1} \tag{4}$$

The residual connection is added to facilitate the origin information propagation in the multi-hop reasoning and aggregating process. At the last reasoning step, we obtained the relationship enhanced and aggregated representation $\mathcal{E} = \{e_1^T, e_2^T, ..., e_n^T\}$, $e_i^T \in R^K$ for each region nodes at the last step.

2.3 Cross-Model Fine-grained Gated Filtration

We notice that not all semantic information contained by region representation is useful for alignments. Aggregating all the possible alignments in an undifferentiated way may hinder the interactive ability of model. Therefore, we propose a cross-modal fine-grained gated filtration mechanism (GF) applying on each image region and description to explore the fine-grained correspondences and suppress ineffectual alignments, as Fig.2 shows. Given the region representations $\mathcal{E}$ of the input image and the textual representation C of the input description, we calculate the fine-grained corresponding and filtering gate g_i for each region as:

$$g_i = \delta(W_g(e_i^T + C)) \tag{5}$$

And then we have the filtered representation $\mathcal{E}^*$ by $e_i^* = g_i \odot e_i^T$ where $W_g \in \mathbb{R}^{K \times K}$ is a linear transformation, $\delta(\cdot)$ is the Sigmoid function, $\odot$ denotes the element-wise product. Afterwards, a mean pooling is applied to obtain the whole image representation $\mathcal{M}$.

2.4 Alignment Objective

For the alignment part, we train AGRF with a hinge-based triplet loss, which is a common ranking objective for text-image retrieval[8, 10]:

$$\mathcal{L} = \sum_{\hat{C} \in B} [\alpha - F(\mathcal{M}, C) + F(\mathcal{M}, \hat{C})]_+ + \sum_{\hat{\mathcal{M}} \in B} [\alpha - F(\mathcal{M}, C) + F(\hat{\mathcal{M}}, C)]_+ \tag{6}$$

where $[x]_+ = max(0, x)$. α is a margin parameter. $F(\cdot)$ is a similarity function in the jointing embedding space. Here, we use the inner product in AGRF. $\hat{C}$ and $\hat{\mathcal{M}}$ is the negative sample of description and image in a mini-batch B of stochastic gradient descent. Rather than considering the hard negative sample, we sum over all negative samples to make AGRF more robust.

3 EXPERIMENTS

To evaluate the effectiveness of the proposed AGRF, we conduct experiments on a real-world description-image dataset and compare it with the recent state-of-the-art methods. We also carry out ablation studies to investigate each component of our model. As is common in information-retrieval, we measure the performance of image retrieval (description query) and description retrieval (image query) by recall at top-K (R@K) .

3.1 Dataset and Settings.

Dataset. We construct a new real-world Chinese description-image dataset, VCG-DI, with 3788 images collected from *Visual China Group*[1] website. The images are collected from different domain such as people, animal and scenery. The number of image objects contained in VCG-DI is 4241. Further, we recruit 38 annotators to write pithy and accurate descriptions for each image. The average length of descriptions is 14.7, and the number of noun words is 2885. For description-image retrieval task, we use 406 images for testing and 3382 images for validation and training.

Implementation Details. Following the setting in [1], the image region features f_i is 2048-dimensional. For textual representation, we use the pre-trained 12-layers BERT-base[2] with 768 hidden units. The weight of BERT-base is fine-tuned during the training stage. The dimension of joining embedding space $K = 2048$. The margin is set to 0.2. The Adam optimizer is used to train AGRF for 30 epochs

[1] https://www.vcg.com/

[2] https://pypi.org/project/transformers/

Table 1: Comparisons of the SOTA methods on VCG-DI

	Image Retrieval			Description Retrieval		
Method	R@1	R@5	R@10	R@1	R@5	R@10
SCAN(2018)	13.1	47.0	68.7	14.0	42.9	57.6
VSRN(2019)	7.1	28.3	40.9	9.1	27.8	41.9
IMRAM(2020)	21.4	56.7	76.1	22.9	59.1	75.9
SGRAF(2021)	17.2	51.0	66.7	14.0	46.8	66.0
AGRF(Ours)	**33.3**	**73.4**	**85.2**	**34.9**	**75.6**	**86.7**

Table 2: Results of ablation studies

	Image Retrieval			Description Retrieval		
Method	R@1	R@5	R@10	R@1	R@5	R@10
MP	25.6	67.5	82.8	30.3	67.9	81.5
MP+GCN	31.8	68.5	84.5	30.2	69.5	84.5
MP+AGR	32.0	71.7	84.5	30.3	70.4	86.5
MP+AGR+GF	**33.3**	**73.4**	**85.2**	**34.9**	**75.6**	**86.7**

with mini-batch size of 32. The learning rate $\gamma = 0.00002$. In term of baselines, to better adapt to Chinese description task, we perform *Jieba* toolkit as tokenizer for all the compared models requiring Chinese word segmentation. The other parameters setting are kept following the previous works. For experiments, we select the best snapshot based on the sum of the recalls on the validation set to avoid over-fitting. we run each experiment three times, then report the average performances.

Baseline. Since the proposed AGRF is the first model for description-image retrieval, there are no previous works that can be directly applied to the task. We choose several state-of-the-art methods of text-image retrieval for performance comparison: (1) SCAN[10], a cross attention model with latent semantic alignment between objects the corresponding words. (2) IMRAM[2], which proposed a Recurrent Attention Memory(RAM) with a memory distillation unit to combine and refine alignment knowledge. (3) SGRAF[5], a reasoning and filtration network, which infers the relation with a sharing weight graph (i.e., GCN[12]) and suppress the irrelevant alignment with an attention filter.(4) VSRN, a local and global reasoning network with GCN and a single modality GRU[3]. There are many other state-of-the-art methods, such as VSE++[6], CAMP[15], MMCA[16]. These works have been outperformed by one or several baselines compared here. Hence, we omit further comparison for space saving.

3.2 Performance Comparison

A summary of the results is shown in Tab.1. Here, we make the following observations:

Effectiveness of aggregation. Compare SCAN with SGRAF, IMRAM and AGRF, we find the methods with reasoning and aggregation achieve substantially better performance than the method without them. This is expected since most alignments are related to multiple regions, exploiting the relationship between regions would boost the complex similarity measure. This observation is consistent with prior works[11].

Effectiveness of adapted aggregation. From the results of VSRN, SGRAF, and AGRF, we observe that the method with adapted aggregate strategy such as AGR achieves relatively higher R@K. This suggests that adapted learning the weight coefficients of image regions in different propagation step is a promising avenue towards the better selection of the discriminating semantic information.

Effectiveness of cross-modal filtration. Among the methods with information filter (i.e., SGRAF, IMRAM, AGRF), we observe that the cross-modal with interactive filter performs better than the single modality filter. This is reasonable because the interactive filter can help to filter trivial information in the image, and enable the representation learning of each image region to focus more on the shared semantics with description.

As Tab.1 shows, benefit from the adapted graph reasoning mechanism and cross-modal fine-grained gate mechanism, AGRF significantly outperforms the state-of-the-art methods. Compared with the best baseline model (i.e., IMRAM), AGRF obtains 11.9% (R@1), 16.7% (R@5), 9.1% (R@10) improvement on image retrieval, and obtains 12.0% (R@1) , 16.5% (R@5) , 10.8% (R@10) improvement on description retrieval. Overall, the experimental results demonstrate that the AGRF is effective for description-image retrieval.

3.3 Ablation Study

In this section, we would like to incrementally validate each component in AGRF by starting from a very basic baseline model MP, which does not perform any aggregation and filtration. This baseline model adapts a mean-pooling operation on the region representation $\mathcal{V}$ to obtain the final representation for the whole image $\mathcal{M}$. The other parts are kept the same as AGRF.

We then respectively add two type of reasoning and aggregation module before the mean-pooling operation, the one is a sharing weight graph neural network (i.e., GCN), the other is the adapted graph reasoning mechanism (AGR). From Tab.2, we can find that these two reasoning and aggregation modules both help to obtain better image representation and improve the matching performance effectively. Further, we also observe that the module with the adapted aggregate strategy outperforms that with the sharing weight strategy. This observation is consistent with what has been made in performance comparison part.

Finally, we combine AGR and GF to get the AGRF. As shown in Tab.2, the performance gain by adding GF validates the effectiveness of interactive filter to boost meaningful alignments.

4 CONCLUSION

we propose a novel *Adapted Graph Reasoning and Filtration (AGRF)* network to overcome the two major challenges of description-image retrieval: (1) reasoning about the visual relationship of multiple objects. (2) casting aside the interference of description-independent detailed information in the image. Our experiments on the real-world dataset demonstrate that AGRF significantly outperform the state-of-the-art methods.

5 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61802340). Zhiling Luo is the corresponding author.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12655–12663.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

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

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity Reasoning and Filtration for Image-Text Matching. *arXiv preprint arXiv:2101.01368* (2021).

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

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

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

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

[12] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*. PMLR, 2014–2023.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[14] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[15] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5764–5773.

[16] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multimodality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10941–10950.