Cross-Graph Attention Enhanced Multi-Modal Correlation Learning for Fine-Grained Image-Text Retrieval

Yi He
Dep. of CS, Huaqiao University
Provincial Key Lab. for Comput. Inf.
Process. Technol., Soochow Univ.
Xiamen and Suzhou, China
yhe@stu.hqu.edu.cn

Xin Liu*
Dep. of CS, Huaqiao University
Fujian Key Lab. of Big Data
Intelligence and Security
Xiamen, China
xliu@hqu.edu.cn

Yiu-ming Cheung
Department of Computer Science,
Hong Kong Baptist University
Hong Kong SAR, China
ymc@comp.hkbu.edu.hk

Shu-Juan Peng*
Dep. of CS, Huaqiao University
Xiamen Key Lab. of Computer Vision
and Pattern Recognition
Xiamen, China
pshujuan@hqu.edu.cn

Jinhan Yi
Dep. of CS, Huaqiao University
Fujian Key Lab. of Big Data
Intelligence and Security
Xiamen, China
jhyi@stu.hqu.edu.cn

Wentao Fan
Dep. of CS, Huaqiao University
Xiamen Key Lab. of Computer Vision
and Pattern Recognition
Xiamen, China
fwt@hqu.edu.cn

ABSTRACT

Fine-grained Image-text retrieval is challenging but vital technology in the field of multimedia analysis. Existing methods mainly focus on learning the common embedding space of images (or patches) and sentences (or words), whereby their mapping features in such embedding space can be directly measured. Nevertheless, most existing image-text retrieval works rarely consider the shared semantic concepts that potentially correlated the heterogeneous modalities, which can enhance the discriminative power of learning such embedding space. Toward this end, we propose a Cross-Graph Attention model (CGAM) to explicitly learn the shared semantic concepts, which can be well utilized to guide the feature learning process of each modality and promote the common embedding learning. More specifically, we build semantic-embedded graph for each modality, and smooth the discrepancy between two modalities via cross-graph attention model to obtain shared semantic-enhanced features. Meanwhile, we reconstruct image and text features via the shared semantic concepts and original embedding representations, and leverage multi-head mechanism for similarity calculation. Accordingly, the semantic-enhanced cross-modal embedding between image and text is discriminatively obtained to benefit the fine-grained retrieval with high retrieval performance. Extensive experiments evaluated on benchmark datasets show the performance improvements in comparison with state-of-the-arts.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Novelty in information retrieval**.

KEYWORDS

Image-text retrieval, cross-graph attention, shared cemantic concept, multi-head mechanism

ACM Reference Format:
Yi He, Xin Liu*, Yiu-ming Cheung, Shu-Juan Peng*, Jinhan Yi, and Wentao Fan. 2021. Cross-Graph Attention Enhanced Multi-Modal Correlation Learning for Fine-Grained Image-Text 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/10.1145/3404835.3463031

1 INTRODUCTION

With the fast development of multimedia technology, multimedia data, such as image and text, has been emerging rapidly and accumulated explosively on the Internet. In order to maximally benefit from the richness of multimedia data, image-text retrieval has become an essential technique for searching engine, featuring on providing flexible retrieval experience to index semantically relevant instance from one modality to another modality [4, 5, 10].

In recent years, image-text retrieval has been extensively studied, and exising works can be roughly categoried into global correspondence learning methods, local correspondence learning methods and high-order semantic concept learning methods. The global correspondence learning methods aim to jointly project the entire image and text data into a common latent space for heterogeneity minimization, whereby the mapping features of image and text in this latent space can be directly measured [3]. Local correspondence learning is designed to capture the fine-grained interplay and semantic correlations between local image patches and partical text data [6, 13, 14]. Along this line, salient image patches are detected and image-text similarity scores are aggregated by all or salient region-word pairs, which have gained significant improvements over traditional global correspondence matching works. To capture more valuable concept information, some recent works focus on investigating the high-order semantic concepts of image and text to reason their higher-level relationships for better peformance. Alone this way, Liu et al. [9] explicitly model object, relation and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR ’21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463031

attribute as a structured phrase, which benefits to learn fine-grained correspondence of structured phrase. Wang et al. [12] utilize the graph model to model the image and text, and jointly characterize the objects and relationships for efficient image-text matching.

To the best of our knowledge, high-order semantic concepts are very useful to correlate the heterogeneous modalities, and different modalities may have shared semantic concepts as well as the modality-specific semantic concept. For cross-modal retrieval, the shared semantic concepts contribute significantly to the learning of commond embedding space, and most existing image-text retrieval works rarely consider the the difference between the shared semantic concepts and modality-specific semantic concepts. enhance the discriminative power of learning such embedding space. Towards this end, this paper presents an efficient cross-graph attention model to explore the consensus information between image and text, which can well enhance the discriminative power of learning the common embedding space. To be specific, we build fully connected weighted graph for both image and text modality, in which the nodes in each graph will be updated via graph attention network. Accordingly, the shared semantic concepts within two modalities are enhanced while the discrepancy between the modality-specific concepts are smoothed. To summarize, the main contributions of our methods are three-fold:

- A novel cross-graph attention model is efficiently designed to explore the shared semantic concepts while smoothing the discrepancy between modality-specific semantic concepts.
- A multi-head mechanism is leveraged to calculate the image-text similarity, which can maximally benefit the fine-grained cross-modal similarity matching task.
- Extensive experiments verify the advantages of the proposed approach under various image-text retrieval tasks.

2 METHOD

Give an image-text pair, the proposed framework aims to learn the semantically consistent feature representations for efficient image-text retrieval. The overall framework is illustrated in Fig. 1, we build semantic-embedded graph for each modality, and propose a cross-graph attention model to obtain shared semantic-enhanced concept features while smoothing the discrepancy between different modalities. Meanwhile, we reconstruct image and text features via the shared semantic concept features and original embedding representations, and leverage multi-head mechanism for similarity calculation. Accordingly, the semantic-enhanced cross-modal embedding between image and text is discriminatively obtained to benefit the fine-grained retrieval.

2.1 Modality Encoder

Image Representation. For an input image $\boldsymbol{I}$, the bottom-up attention model [1] is utilized to discriminate region features. Accordingly, the category of instance and the object attribute can be well obtained, simiply denoted as $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_{n_o}\}$. Further, a fully-connect layer is then applied to transform these object features with more discriminative power, and the transformed features are denoted as $\boldsymbol{H}^{\mathcal{I}} = \{\boldsymbol{h}_1^{\mathcal{I}}, \boldsymbol{h}_2^{\mathcal{I}}, \cdots, \boldsymbol{h}_{n_o}^{\mathcal{I}}\}$, with $\boldsymbol{h}_i^{\mathcal{I}}$ corresponding to the transformed feature of $\boldsymbol{o}_i$.

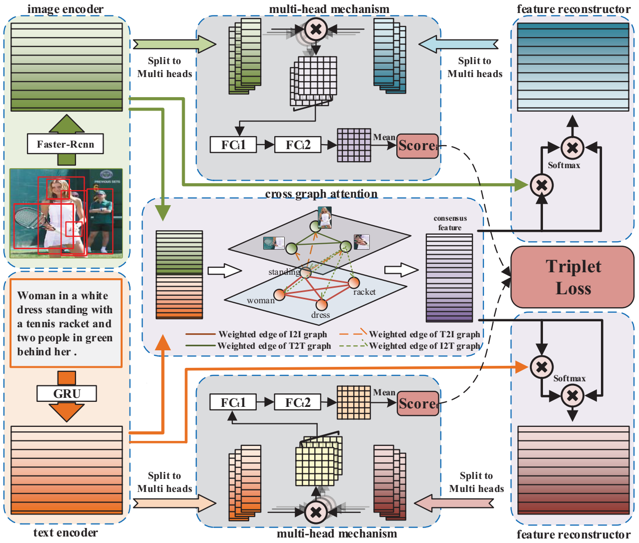


Figure 1: Schematic architecture of the proposed model.

Text Representation. For a text $\boldsymbol{T}$, we first represent each word as a one-hot vector, and then embed it into d-dimensional feature space using a Bidirectional Gated Recurrent Unit. Accordingly, the representation of i-th word is obtained by averaging the hidden state of forward and backward GRU at i-th time step. Therefore, the representation of text is defined as $\boldsymbol{H}^{\mathcal{T}} = \{\boldsymbol{h}_1^{\mathcal{T}}, \boldsymbol{h}_2^{\mathcal{T}}, \cdots, \boldsymbol{h}_{n_w}^{\mathcal{T}}\}$.

2.2 Cross-Graph Attention Model

Motivated by graph attention network (GATs) [11], we design a cross graph attention model to smooth the semantic discrepancy between image and text, and simultaneously enhance the shared shared semantic concepts to refine the feature represenations. First, we combine the feature representation of image and text as Eq. (1).

$$Z = \begin{bmatrix} \boldsymbol{H}^{\mathcal{I}} \\ \boldsymbol{H}^{\mathcal{T}} \end{bmatrix} \tag{1}$$

Further, we build semantic connections in both inter modalities and intra modalities, whereby four fully connected weighted graphs are obtained: $G_{I\to I} = (V_1, E_1)$, $G_{I\to T} = (V_2, E_2)$, $G_{T\to I} = (V_3, E_3)$ and $G_{T\to T} = (V_4, E_4)$, where V denotes the node sets of the graph, E represents the associated edge between each node pairs. For graph attention learning, it is noted that the semantic connection of node pairs is bi-directed, which can explicitly characterize the destination and source information for cross-modal matching task.

$$Q_i{=}FC_q(\boldsymbol{H}^{\mathcal{I}}), K_i{=}FC_k(\boldsymbol{H}^{\mathcal{I}}), Q_t{=}FC_q(\boldsymbol{H}^{\mathcal{T}}), K_t{=}FC_k(\boldsymbol{H}^{\mathcal{T}}) \tag{2}$$

where Q and K denote the source and destination information in graph attention model, and $FC(\cdot)$ denotes fully-connect layer. Then, the affinity matrix of graph nodes can be formulated as:

$$A_{I\to I}{=}Q_i{K_i}^T, A_{I\to T}{=}Q_i{K_t}^T, A_{T\to T}{=}Q_t{K_t}^T, A_{T\to I}{=}Q_t{K_i}^T \tag{3}$$

$$A = \begin{bmatrix} A_{I\to I}/N & A_{I\to T}/N \\ A_{T\to I}/N & A_{T\to T}/N \end{bmatrix} \tag{4}$$

where N is the sum of n_o and n_w that utilized to regularize the weight, and A denotes the adjacent matrix of graph. Note that, the shared semantic concepts contribute significantly to the learning of commond embedding space, and therefore we combine the four sub-matrix to aggregate the attention features across different modalities. Accordingly, the shared concept feature embedding C of image and text is calculated as:

$$C = FC(AZW + Z) \tag{5}$$

where W is the trainable parameter in the network model.

2.3 Feature Reconstructor

In order to judge whether an image-text pair matches or not, we select to reconstruct the modality-specific feature represenations by using the shared concept representation. First, we calculate the weight between original and shared feature representation:

$$W^{\mathcal{I}} = \boldsymbol{H}^{\mathcal{I}} C^T, W^{\mathcal{T}} = \boldsymbol{H}^{\mathcal{T}} C^T \tag{6}$$

$$W_{ij}^{\mathcal{I}} = \frac{exp(\lambda W_{ij}^{\mathcal{I}})}{\sum_{j=1}^{N}(\lambda exp(W_{ij}^{\mathcal{I}}))}, W_{ij}^{\mathcal{T}} = \frac{exp(\lambda W_{ij}^{\mathcal{T}})}{\sum_{j=1}^{N}(\lambda exp(W_{ij}^{\mathcal{T}}))} \tag{7}$$

where $W^{\mathcal{I}}$ and $W^{\mathcal{T}}$ are respectively the weights between original feature and shared common feature of image and text, λ is a scaling factor. Accordingly, we implement the feature reconstruction process by directly using the weighting information:

$$R^{\mathcal{I}} = W^{\mathcal{I}} C, \quad R^{\mathcal{T}} = W^{\mathcal{T}} C \tag{8}$$

where $R^{\mathcal{I}}$ and $R^{\mathcal{T}}$ are the reconstructed feature of image and text.

2.4 Multi-Head Similarity Calculating

After obtaining the modality-specific features, we calculate the similarity between the reconstructed feature representation and the shared concept feature representations, with large similarity indicating the matched image-text pair and small similarity indicating the unmatched pair. Specifically, the multi-head mechanism is leveraged for calculating the similarity, by splitting the reconstructed features and the original features into k-heads:

$$H^{\mathcal{I}} = Concat(H_1^{\mathcal{I}}, \cdots, H_K^{\mathcal{I}}), H_i^{\mathcal{T}} = Concat(H_1^{\mathcal{T}}, \cdots, H_K^{\mathcal{T}}) \tag{9}$$

$$R^{\mathcal{I}} = Concat(R_1^{\mathcal{I}}, \cdots, R_K^{\mathcal{I}}), R_i^{\mathcal{T}} = Concat(R_1^{\mathcal{T}}, \cdots, R_K^{\mathcal{T}}) \tag{10}$$

Specifically, we calculate the cosine similarity between the reconstructed features and original features for each head:

$$S_k^{\mathcal{I}} = \frac{H_k^{\mathcal{I}} R_k^{\mathcal{I}\,T}}{\left\|H_k^{\mathcal{I}}\right\| \left\|R_k^{\mathcal{I}}\right\|}, \quad S_k^{\mathcal{T}} = \frac{H_k^{\mathcal{T}} R_k^{\mathcal{T}\,T}}{\left\|H_k^{\mathcal{T}}\right\| \left\|R_k^{\mathcal{T}}\right\|} \tag{11}$$

where $\|\cdot\|$ denotes ℓ_2 norm regularization. Then, two fully-connect layers are applied to calculate similarity.

$$S^{\mathcal{I}} = W_2^{\mathcal{I}}(tanh(W_1^{\mathcal{I}}(concat(S_1^{\mathcal{I}}, \cdots, S_k^{\mathcal{I}})) + b_1^{\mathcal{I}})) + b_2^{\mathcal{I}} \tag{12}$$

$$S^{\mathcal{T}} = W_2^{\mathcal{T}}(tanh(W_1^{\mathcal{T}}(concat(S_1^{\mathcal{T}}, \cdots, S_k^{\mathcal{T}})) + b_1^{\mathcal{T}})) + b_2^{\mathcal{T}} \tag{13}$$

where $W_2^{\mathcal{I}}, b_2^{\mathcal{I}}, W_1^{\mathcal{I}}, b_1^{\mathcal{I}}, W_2^{\mathcal{T}}, b_2^{\mathcal{T}}, W_1^{\mathcal{T}}, b_1^{\mathcal{T}}$ are trainable parameters. Thus, the final similarity is the average of $S^{\mathcal{I}}$ and $S^{\mathcal{T}}$.

2.5 Loss Function

Following [3], the triplet loss is utilized to optimize the hard negative samples, and the similarity stated in Subsection 2.4 is employed for regularization, totally resulting the following loss:

$$\mathcal{L} = \sum_{(I,T)} \left([\alpha - S_{IT} + S_{I\hat{T}}]_+ + [\alpha - S_{IT} + S_{\hat{I}T}]_+\right) \tag{14}$$

where $[\cdot]_+ = max(\cdot, 0)$, α denotes the margin between the positive pairs and negative pairs that should be converged in training process, $\hat{I}$ and $\hat{T}$ are the hard negative samples, denoted as $\hat{I} = argmax_{\tilde{I}} S_{\tilde{I}T}$ and $\hat{T} = argmax_{\tilde{T}} S_{I\tilde{T}}$, $\tilde{I}, \tilde{T}$ are negative samples.

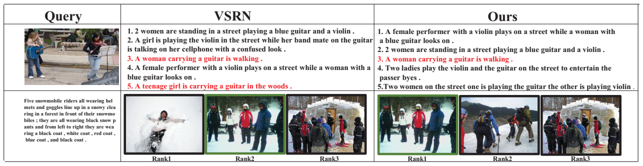
3 EXPERIMENTS

3.1 Dataset and Evaluation

Two public available multi-modal datasets, i.e., MSCOCO [15] and Flickr30K [8], are chosen in the experiments. MSCOCO contains 123,287 images, and each image is annotated with five captions. The widely used splitting scheme contains 113,287 images for training, 5000 images for validation and 5000 images for testing. Flickr30K contains 31,000 images collected from Flickr website with five captions. Following the splitting scheme in [3], we select 1,000 images for validation and 1,000 images for testing and the rest for training. To quantitatively evaluate the retrieval performance, we report the score of Recall@K and 'mR' scores for overall evaluation [2].

3.2 Implementation Details

The proposed framework is implemented in pytorch platfom. For image representation learning, the pre-trained visual features with 36 patches provided by SCAN [6] is selected for training, and each patch is characterized with 2048-dimension vector. The dimension of the joint embedding space is 1024. For training, we utilize Adam optimizer with 25 epochs and the initial learning rate is set at 0.0002, with decaying 10% every 8 epochs for both Flickr30k and MSCOCO datasets. In the multi-head similarity calculating, we set the heads k to 16 and the scaling factor λ in Eq. (7) is set to 4. For the regularization parameters, the margin α is set at 0.2.



Query	VSRN	Ours
	1. 2 women are standing in a street playing a blue guitar and a violin . 2. A girl is playing the violin in the street while her band mate on the guitar is talking on her cellphone with a confused look . 3. A woman carrying a guitar is walking . 4. A female performer with a violin plays on a street while a woman with a blue guitar looks on . 5. A teenage girl is carrying a guitar in the woods .	1. A female performer with a violin plays on a street while a woman with a blue guitar looks on . 2. 2 women are standing in a street playing a blue guitar and a violin . 3. A woman carrying a guitar is walking . 4. Two ladies play the violin and the guitar on the street to entertain the passer byes . 5.Two women on the street one is playing the guitar the other is playing violin .
Five snowmobile riders all wearing hel mets and goggles line up in a snowy clea ring in a forest in front of their snowmo biles ; they are all wearing black snow p ants and from left to right they are wea ring a black coat , white coat , red coat , blue coat , and black coat .	Rank1 Rank2 Rank3	Rank1 Rank2 Rank3

Figure 2: Representative results obtained by our method.

3.3 Performance Analysis and Comparison

The cross-modal retrieval results tested on MSCOCO and Flick30K are summarized in Table 2, it can be observed that the proposed framework always delivers the best Recall@K and 'mR' performance, and generally performs better than all the baselines. For instance, comparing with the competing GSMN (dense) [9], our propsed framework gains 2.3% and 0.8% improvements on R@1 scores, respectively, for text retrieval and image retrieval on Flickr30k. This indicates that the proposed framework is capable of indexing

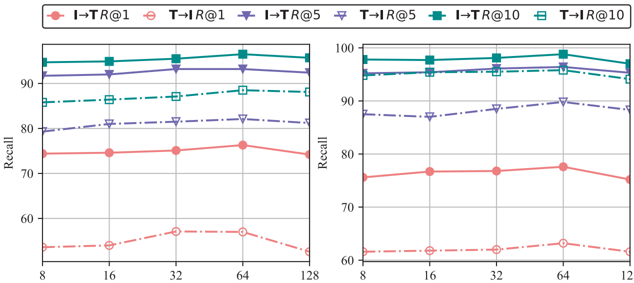
Table 1: Quantitative results of text retrieval and image retrieval tested on MSCOCO and Flickr30k test set in terms of R@K. ’*’ denotes that ensemble results obtained from two trained models achieved by the same method.

Method	Flick30K							MSCOCO						
	Text Retrieval			Image Retrieval			mR	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++[3]	52.9	80.5	87.2	39.6	70.1	79.5	68.3	64.6	90.0	95.7	52.0	84.3	92.0	79.8
SCAN* [6]	67.4	90.3	95.8	48.6	77.7	85.2	77.5	72.7	94.8	98.4	58.8	88.4	94.8	84.7
CAMP [13]	68.1	89.7	95.2	51.5	77.1	85.3	77.8	72.3	94.8	98.3	58.5	87.9	95.0	84.5
CASC* [14]	68.5	90.6	95.9	50.2	78.3	86.3	78.3	72.3	96.0	**99.0**	58.9	89.8	96.0	85.3
SGM [12]	71.8	91.7	95.5	53.5	79.6	86.5	79.8	73.4	93.8	97.8	57.5	87.3	94.3	84.0
VRAN* [7]	71.3	90.6	96.0	54.7	81.8	88.2	80.5	76.2	94.8	98.2	62.8	89.7	95.1	86.1
GSMN (sparce)	71.4	92.0	96.1	53.9	79.7	87.1	80.0	76.1	95.6	98.3	60.4	88.7	95.0	85.7
GSMN (dense)	72.6	93.5	96.8	53.7	80.0	87.0	80.6	74.7	95.3	98.2	60.3	88.5	94.6	85.3
GSMN*[9]	76.4	94.3	97.3	57.4	82.3	89.0	82.8	78.4	96.4	98.6	63.3	90.1	95.7	87.1
Ours	76.3	93.2	96.5	57.0	82.1	88.5	82.3	77.6	96.4	98.8	62.2	89.8	95.8	86.8
Ours*	**78.7**	**94.5**	**97.9**	**58.2**	**83.6**	**89.6**	**83.8**	**78.9**	**97.5**	98.8	**65.7**	**90.2**	**96.6**	**87.9**

much more similar samples in fine-grained cross-modal matching results. Meanwhile, some representative retrieval examples obtained by the proposed approach are shown in Fig. 2, it can be clearly observed that the proposed model is able to distinguish the similar queries well and have successfully indexed the most semantically matched counterparts.

Table 2: Ablation studies tested on Flickr30k test set.

Method	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
w/o-i	71.5	89.3	94.6	51.6	78.7	86.4
w/o-t	75.1	92.8	97.8	55.6	80.1	86.9
w/o cross graph	72.6	91.6	94.6	52.5	79.4	87.4
w/o multi-head	74.3	93.5	95.2	56.8	81.7	87.7
baseline	76.3	93.2	96.5	57.0	82.1	88.5



(a) num of heads on Flickr30K (b) num of heads on MSCOCO

Figure 3: Evaluation of heads number k.

3.4 Ablation Studies and Discussions

We further evaluate the effectiveness of each learning module and experiment with four forms of the proposed model: 1)**w/o-i** is the model without the visual similarity calculation and feature reconstruction; 2)**w/o-t** is the model without the textual similarity calculation and feature reconstruction; 3) **w/o cross graph** is the model that combines the image and text feature directly; 4) **w/o multi-head** is the model without multi-head mechanism. As shown in Table 2, it can be found that the proposed framework embedded with cross-graph attention, similarity calculation and feature reconstruction is able to yield the discriminative shared representation and theirfore significantly promote the retrieval performance. Further, we explore the number effect of heads of multi-head mechanism in Eqs. (9) and (10), and representative results are shown in Fig. 3, it can be observed that different k values just induce a minor fluctuation on the retrieval performance, and yield very stable retrieval performance on different retrieval tasks.

4 CONCLUSION

This paper has proposed an efficient cross-graph attention model to explicitly learn the shared semantic concepts between image and text, which can be well utilized to guide the feature learning process for efficient fine-grained image-text retrieval. Meanwhile, the leveraged multi-head mechanism is able to well correlate the similarity between heterogeneous modalities. Accordingly, the semantic-enhanced cross-modal embedding can be discriminatively obtained to benefit the fine-grained retrieval. Extensive experiments have shown its outstanding performances.

ACKNOWLEDGMENTS

The work was supported in part by the National Science Foundation of China (Nos. 61673185, 61672444 and 61876068), in part by the National Science Foundation of Fujian Province (Nos. 2020J01083 and 2020J01084), in part by Quanzhou City Science & Technology Program of China (No. 2018C107R), in part by Hong Kong Baptist University (HKBU) under Grants: RC-FNRA-IG/18-19/SCI/03 and RC-IRCMs/18-19/01, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong under Project ITS/339/18, in part by the SZSTC (SGDX20190816230207535), and in part by Program (KJS2047) of the Provincial Key Laboratory for Computer Information Processing Technology (Soochow University).

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 CVPR*. 6077–6086.

[2] Yewang Chen, Xiaoliang Hu, Wentao Fan, Lianlian Shen, Zheng Zhang, Xin Liu, Jixiang Du, Haibo Li, Yi Chen, and Hailin Li. 2020. Fast density peak clustering for large scale data based on kNN. *Knowledge Based Systems* 187 (2020). Article No. 104824.

[3] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of BMVC*. 1–14.

[4] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-Modal Retrieval. In *Proceedings of ACM SIGIR*. 2251–2260.

[5] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *Proceedings of ACM SIGIR*. 635–644.

[6] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of ECCV*. 201–216.

[7] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *Proceedings of IEEE ICCV*. 4654–4662.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of ECCV*. 740–755.

[9] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of IEEE CVPR*. 10921–10930.

[10] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu ming Cheung. 2021. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (2021), 964–981.

[11] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of ICRL*.

[12] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 1497–1506.

[13] 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 ICCV*. 5764–5773.

[14] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-Modal Attention With Semantic Consistence for Image-Text Matching. *IEEE Transactions on Neural Networks and Learning Systems* 31, 12 (2020), 5412–5425.

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.