Self-supervised Visual-Semantic Embedding Network Based on Local Label Optimization

Zhukai Jiang and Zhichao Lian(✉)

Nanjing University of Science and Technology, Nanjing, China
lzcts@163.com

Abstract. Image-text retrieval has always been an important direction in the field of vision-language understanding, which is dedicated to bridging the semantic gap between two modalities. The existing methods are mainly divided into global visual-semantic embedding and local region-word alignment. Although the local region-word alignment method has achieved remarkable results, this method based on fine-grained features often leads to low retrieval efficiency. At the same time, the method based on global embedding lacks extra semantic information, resulting in insufficient accuracy. In this paper, we propose a novel self-supervised visual-semantic embedding network based on local label optimization. Specifically, we generate a label for the entire image-text pair from the local information and use this label to optimize our embedding network, which can not only affect the retrieval efficiency but also significantly improve the retrieval accuracy. Experimental results on two benchmark datasets validate the effectiveness of our method.

Keywords: Visual-semantic embedding · Self-supervised · Local label · Image-text retrieval

1 Introduction

With the rapid growth of various modal data in the Internet, information processing between cross-modal data is urgently needed. In all modalities of data, images and text dominate, so vision-language understanding has naturally become a hot research direction. With the progress of deep learning technology in recent years, many vision-and-language tasks have achieved remarkable results, such as image caption [1], text-to-image synthesis [19] and image-text retrieval [5]. In this paper, we focus on the image-text retrieval task, which refers to returning a similarity ranking of data from another modal in a database based on input image or text modal data. Obviously, bidirectional image-text retrieval has a wide range of applications. People often need to enter a text description to search for the most relevant image or enter an image to search for news corresponding to the image. The current popular image-text retrieval approaches

Y. Xu et al. (Eds.): ML4CS 2022, LNCS 13657, pp. 400–412, 2023.
https://doi.org/10.1007/978-3-031-20102-8_31

can be roughly divided into two classes, global visual-semantic embedding and local region-word alignment.

Global visual-semantic embedding refers to mapping the data of two different modalities into the common representation space to obtain the global representation, and calculate the image-text similarity in this space. Traditional methods use statistical correlation analysis to learn linear projections by optimizing target statistics. One of the most representative works is Canonical Correlation Analysis [6], which learns common spaces by maximizing associations between cross-modal data. Thanks to advances in deep learning technology, many methods have emerged to learn public spaces through deep neural networks [5,11,13,15,22].For example, Faghri et al. [5] use classical feature extraction networks such as ResNet [7] and GRU [3] to extract global features for image-text pairs, and fully exploit the potential information between image and text pairs through the triple loss function based on the most negative sample. This loss function has also become the basis of subsequent work. For labeled data, DSCMR [22] designed a new loss function to minimize the discriminative loss in two representation spaces so as to preserve the semantic difference and intra-modal invariance.

The visual-semantic embedding methods mentioned above all input the image as a whole into the network, and the use of the overall feature has indeed achieved certain results. But gradually people found that if you want to further improve the accuracy of retrieval, the use of overall features is far from enough. In fact, when people describe what they see, they often describe objects or other key areas in the image, and the words in the text correspond to a specific area in the image. Therefore, people began to study fine-grained local region-word alignment methods. In these methods, the overall similarity is computed by the correlation between image local regions and text words. SCAN [10] proposed a cross-modal stack attention module (Stack Cross Attention) for region-text alignment, and then used the aligned image region features and text word features for similarity

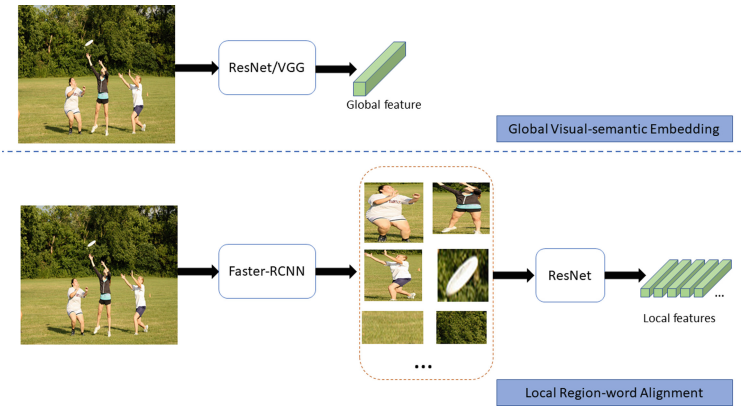

Fig. 1. Two methods of image feature extraction.

calculation. Later, people carried out further research [12,17,21] on this basis, and the local region-word alignment method achieved remarkable success.

But the drawbacks of the local region-word alignment method are equally obvious. As shown in Fig. 1, to extract the region features, it is necessary to first detect the salient regions of the image through the Faster-RCNN [14] model, and then use the ResNet101 [7] network to extract the deep semantic features of these regions. In addition, the process of calculating similarity by local region-word alignment method during retrieval is more complicated, and it will consume a lot of time when the retrieval data set is too large. When performing large-scale retrieval in practical applications, the visual-semantic embedding method with high efficiency but insufficient accuracy is the mainstream choice. Based on the above discussion, in order to improve retrieval accuracy while maintaining retrieval efficiency, this paper proposes a self-supervised visual-semantic embedding network (S-VSE) based on local label optimization. During network training, we simultaneously extract global and local features of images and text. Besides the key main global embedding branch, we augment the pseudo-label generation branch with image local information to provide additional supervision information. Using additional supervision information, the network can learn a better common embedding space, and only need to extract global features and use the main branch to obtain the global representation for similarity calculation during testing, so that the retrieval time will not be increased. Our main contributions can be summarized as follows:

(1) We propose a novel self-supervised visual-semantic embedding network based on local label optimization(S-VSE), which uses the local information of image text to obtain the labels of image-text pairs, and uses this label information to optimize the embedding space.

(2) We use the self-attention module to obtain the text global representation, and the self-attention module can effectively select important parts of the entire text.

(3) We conduct experiments on two benchmark image-text retrieval datasets, and the experimental results demonstrate the effectiveness of our method.

2 Related Work

2.1 Global Visual-Semantic Embedding

The essence of global visual semantic embedding is to embed cross-modal data into a latent space in which the similarity of heterogeneous data can be directly calculated. The main challenge of this type of method is to bridge the "heterogeneous gap" between the cross-modal data. Initially, people tried to use some traditional methods to study. For example, CCA [6] uses linear projection to encode cross-modal data into a highly correlated common subspace. Later, people began to apply deep neural networks to projection. DCCA [2] builds multiple stacked non-linear transformation layers and learns to maximize the correlation of visual and textual representations.

Driven by the wave of artificial intelligence, the fields of natural language processing and computer vision have made tremendous progress. People have begun to use classical feature extraction networks and pre-trained models on large-scale datasets for cross-modal retrieval research. VSE++ [5] uses the pre-trained model of ResNet network on ImageNet [4] and the GRU [3] network with strong ability in the direction of natural language processing to extract the image-text pair global features, and firstly proposes a triplet based on the most negative sample loss. This loss function has also become the basis for subsequent research. DAN [13] uses a two-stream convolutional neural network to extract image and text features, treats each image-text pair as a different instance, and uses additional instance loss to optimize the common space, which can not only mine the intra-modality The subtle differences can also maintain the differences between modalities. DSCMR [22] uses the weight sharing strategy to eliminate the cross-modal differences in the public representation space and learn the modality invariant features. SSAH [11] incorporates adversarial learning into cross-modal hashing research in a self-supervised manner. The main contribution of this work employs several adversarial networks to maximize semantic relevance and representation consistency between different modalities. In addition, the self-supervised semantic network is used to discover high-level semantic information in the form of multi-label annotations. PRDH [20] considers the similarity of different instances within the same modality and utilizes a matrix to constrain the generated hash codes.

All the methods mentioned above can be summarized as global visual-semantic embedding methods, because these methods embed the whole image and text into a common representation space, using a vector to represent the whole image or text. However, there are more complex correspondences between image and text. Text often selectively describes the content of an image, and the same text or area also has complex semantic information in different image-text pairs. To mine complex relationships, coarse-grained global features are not enough, and people gradually begin to use fine-grained local features for related research.

2.2 Local Region-Word Alignment

Karpathy and Fei-Fei [8] first extract local features for each image and text, and compute image-text similarity by aggregating local similarities. SCAN [10] proposes a stacked cross-attention mechanism that aligns each region with all words and aligns each word with all regions, and accurate local similarity can be obtained through the aligned local features. On the basis of SCAN, PFAN [16] further considers the location information of the region, and designs a location information aggregation strategy for accurate retrieval. Considering contextual information, CAAN [21] exploits both global inter-modal alignment and intra-modal correlation to find underlying correlation. Although local region-word alignment methods have achieved remarkable results, local alignment methods consume more time in both the feature extraction stage and retrieval stage.

Therefore, more efficient global visual-semantic embedding methods still have research value.

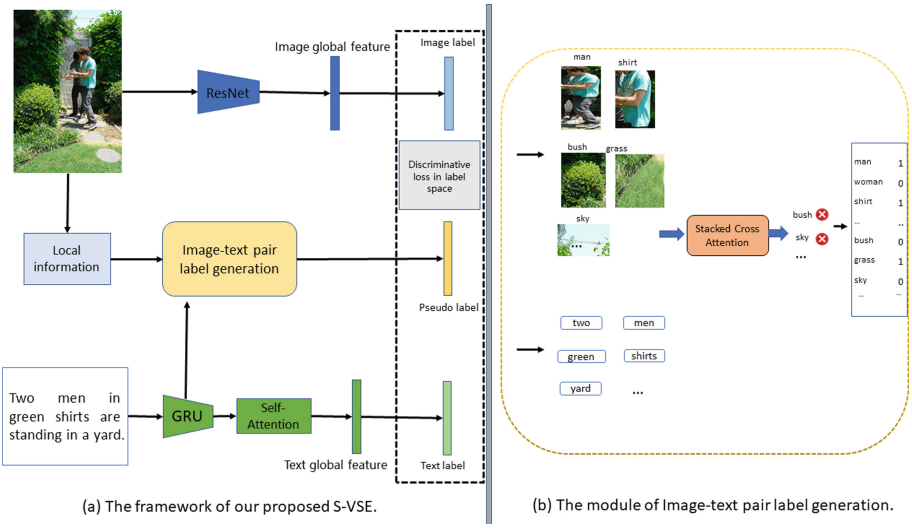

Fig. 2. (a) is the framework of S-VSE, which mainly consists of the main visual-semantic embedding branch and the image-text label generation module. (b) describes the flow of image-text label generation module, which is to eliminate irrelevant regions through the alignment of local information.

3 Method

The overall network framework of our proposed method is shown in Fig. 2(a). The network is mainly divided into a visual-semantic embedding branch and a image-text pair label generation module. Regarding the visual-semantic embedding branch, we add a self-attention module based on VSE++ to obtain a better global representation of the text. The general framework of the image-text label generation branch is shown in Fig. 2(b). We use Stack Cross Attention to calculate the correlation of image regions with a given text, so as to select important regions for label generation. We will detail our approach from the following five sections.

3.1 Feature Extraction

Image Representation. For each input image I, we follow VSE++ [5] to extract image global feature $V_0 \in \mathbb{R}^{4096}$ using a pretrained ResNet152 model. Regarding image local information, we follow DSRAN [18] to first detect K

salient regions of the input image using the Faster-RCNN model pre-trained on the Visual Genomes [9] dataset, and then extract the feature vectors and label probability vectors of these regions. For regional feature vectors, we need to go through a fully connected layer for dimensionality reduction, and finally we get the local feature set $V = \{v_1, v_2, ..., v_K\} \in \mathbb{R}^{1024\ timesK}$ and the local label probability vector set $C = \{c_1, c_2, ..., c_K\} \in \mathbb{R}^{1600\times K}$.

For each input text T, we follow VSE++ [5] to divide the text into M words, and then use a bidirectional GRU network to extract the feature representation of each word. In this way, we get the local feature set $T = \{t_1, t_2, ..., t_M\} \in \mathbb{R}^{1024\times M}$. Since the text itself consists of several words, the mean feature of the local feature set of the text is often used as the global representation of the text. In my approach, we use a self-attention part to mine inter-word relations (we will introduce this module in detail in Sect. 3.3) to suppress unimportant information in the text and obtain a more accurate global text representation T_g.

3.2 Image-Text Pair Label Generation

First we convert the obtained image local label probability set C into one-hot labels. For each local region of the image, we have obtained a 1600-dimensional probability vector, i.e. the total number of categories is 1600. We set the class with the largest probability value as 1 to indicate the class the region belongs to, and the rest of the values are 0. In this way, we get the local label set $L = \{l_1, l_2, ..., l_K\} \in \mathbb{R}^{1600\times K}$, but such a simple image local label obviously cannot meet our needs. In fact, the text description corresponding to the image is often a description of some specific content of the image, and not all important areas will appear in the text description. Also, different regions have different semantics in different texts. Therefore, when generating the label of the entire image-text pair, we need enough interaction between the image and text to mine complex correspondences.

We use Stack Cross Attention [10] for local region-word alignment. For K regions and M words, first calculate the cosine similarity between image regions and text words:

$$s_{ij} = \frac{v_i^T t_j}{\|v_i\| \, \|t_i\|}, i \in [1, K], j \in [1, M], \tag{1}$$

where s_{ij} represents the similarity between the i-th region and the j-th word. The local similarity is then normalized using the following formula:

$$\overline{s}_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{i=1}^{K} [s_{ij}]_+^2}}, \tag{2}$$

where $[x]_+ \equiv max(x, 0)$.

Then, to attend on words with respect to each image region, we use a weighted combination of word representations. For example for the j-th word, the defini-

tion is as follows:

$$a_i^t = \sum_{j=1}^{M} \alpha_{ij} t_j, \tag{3}$$

where

$$\alpha_{ij} = \frac{exp(4\overline{s}_{ij})}{\sum_{j=1}^{M} exp(4\overline{s}_{ij})}. \tag{4}$$

We still use cosine similarity to calculate the correlation between each region and the aligned words, i.e.

$$R_i = \frac{v_i^T a_i^t}{\|v_i\| \, \|a_i^t\|}, i \in [1, K]. \tag{5}$$

The label $\hat{L}$ of the entire image-text pair is defined as:

$$\hat{L} = \sum_{i=1}^{K} R_i^* l_i, \tag{6}$$

where

$$R_i^* = \begin{cases} 1, R_i \geq \beta \\ 0, R_i < \beta \end{cases} \tag{7}$$

. β represents the threshold of correlation, when the correlation of the region is less than β, we consider the region to be irrelevant to the content of the text.

3.3 Self-attention Module

For the obtained text local features T, we calculate the mean value of text local features $q = \frac{1}{M} \sum_{j=1}^{M} t_j$, and use q as the query in the self-attention module. That is to calculate the attention score of q about all word features:

$$w = (W_1 q)^T (W_2 T), \tag{8}$$

where W_1 and W_2 are learnable parameter matrices. Then use the Softmax function to normalize the score w to get the attention weight of each word $\bar{w} = \{\bar{w}_1, \bar{w}_2, .., \bar{w}_M\}$.

Finally we use this weight to get the text global representation T_g:

$$T_g = \sum_{j=1}^{M} \bar{w}_j t_j. \tag{9}$$

3.4 Visual-Semantic Embedding

For the image global feature V_0, we normalize it and let it into a 1024-dimensional common representation space to get the global representation V_g:

$$V_g = W^T V_0, \tag{10}$$

where W is a learnable embedding matrix. In this way, for the input image-text pair (I, T), we use the cosine similarity function to calculate their similarity $S(I, T)$, i.e.

$$S(I,T) = \frac{V_g^T T_g}{\|V_g\| \, \|T_g\|}. \tag{11}$$

3.5 Loss Function

We adopt a hinge-based triplet ranking loss with emphasis on hard negatives [5] as our main loss. This loss function is defined as:

$$L_{embed} = [\mu + S(I,\widehat{T}) - S(I,T)]_+ + [\mu + S(\widehat{I},T) - S(I,T)]_+, \tag{12}$$

where $[x]_+ \equiv max(x, 0)$, μ is a margin parameter, $\widehat{T} = argmax_{t \neq T} S(I, t)$ and $\widehat{I} = argmax_{i \neq I} S(i, T)$ stand for hardest negatives in a mini-batch.

In addition, drawing on the idea of DSCMR, we also add a fully connected layer as the label space after the embedding layer of the image and text, so that the discriminative loss in the label space can be obtained:

$$L_{label} = \left\| P_1 V_g - \widehat{L} \right\|_F + \left\| P_2 T_g - \widehat{L} \right\|_F, \tag{13}$$

where P_1 and P_2 are the parameter matrices of the fully connected layer, $\| * \|_F$ represents the F-norm. The total loss is defined as follows:

$$Loss = L_{embed} + \lambda * L_{label}, \tag{14}$$

where λ is a constant value that balances the impact of two loss terms.

4 Experiments

4.1 Dataset and Evaluation Metric

We evaluated our model on MS-COCO and Flickr30K datasets. MS-COCO contains 123,287 images, each with five text annotations. According to VSE++ [5], we divide the dataset into 25000 image-text pairs for validation, 25000 image-text pairs for testing, and the remaining image-text pairs for training. Flickr30K contains 31,000 images, each with five text annotations. We also follow VSE++ [5] to divide the dataset into 5000 image-text pairs for validation, 5000 image-text pairs for testing, and the remaining image-text pairs for training. We follow DSRAN [18] to adopt R@1, R@5, R@10 and Rsum as our evaluation metrics.

4.2 Implementation Details

We follow the VSE++ approach and divide the model into two types: finetune (ft) and no finetune. No fine-tuning refers to freezing the parameters in the pre-trained ResNet152 model used to extract the global features of the image during

training, and only training the parameters of the embedding layer. Fine-tuning is to train all parameters used to extract global image features without fine-tuning the model. In addition, in the label generation part of the image and text, the parameters in the SCAN model that we have trained are used in the extraction of local features of images and texts. On both datasets, we set the number of image regions K to 36, the parameter $\mu to 0.2$, the parameter β to 0.4 and λ to 10. The maximum number of epochs for model training without fine-tuning is 25, the initial learning rate is 0.0002, the learning rate is reduced to 0.00002 after 15 epochs, and the batch size is 128. The maximum number of epochs for fine-tuning model training is 25, the initial learning rate is 0.00002, and the learning rate is reduced to 0.000002 after 15 epochs, and the batch size is 64. All experiments are performed on an NVIDIA 3090 GPU, and the Adam optimizer is used for training.

Table 1. Results on Flickr30k

Method	Image-to-text			Text-to-image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
DSPE	40.3	68.9	79.9	29.7	60.1	72.1	351.0
VSE++	43.7	71.9	82.1	32.3	60.9	72.1	363.0
VSE++(ft)	52.9	80.5	87.2	39.6	70.1	79.5	409.8
DAN	55.6	81.9	89.5	39.1	69.2	80.9	416.2
S-VSE	52.1	77.0	86.5	36.3	66.3	76.6	394.7
S-VSE(ft)	**62.9**	**85.3**	**91.5**	**45.8**	**75.4**	**83.6**	**444.5**

Table 2. Results on MS-COCO

Method	Image-to-text			Text-to-image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
DSPE	50.1	79.7	89.2	39.6	75.2	86.9	420.7
VSE++	58.3	86.1	93.3	43.7	77.6	87.8	446.8
VSE++(ft)	64.6	90.0	95.7	52.0	84.3	92.0	478.6
DAN	65.6	89.8	95.5	47.1	79.9	90.0	467.9
S-VSE	60.9	87.7	94.2	44.8	79.0	89.1	455.7
S-VSE(ft)	**66.2**	**90.6**	**96.0**	**54.9**	**86.0**	**93.6**	**487.2**

4.3 Experimental Results and Analysis

We compare our proposed S-VSE network with three classic global visual-semantic embedding methods (DSPE [15], VSE++ [5], DAN [13]). Tables 1 and 2 are our experimental results on Flickr30k and MS-COCO.

Looking at Table 1, we can find that our proposed S-VSE model achieves the best retrieval performance on the Flickr30k dataset. With fine-tuning, compared with the current best model DAN, we achieve a significant improvement of 28.3% in the Rsum metric, and the improvement in other detail metrics is also very obvious. For example, we can also achieve significant improvements of 7.3% and 6.7% on the R@1 metric for text retrieval and image retrieval. Without fine-tuning, the S-VSE model is able to achieve a huge 31.7% improvement on the Rsum metric compared to our base model VSE++. Looking at Table 2, we can find that on the MS-COCO dataset, our S-VSE model can achieve 8.9% and 8.6% improvement in Rsum without fine-tuning and fine-tuning, respectively, compared with VSE++. Clearly, our model does not improve as well on MS-COCO dataset as it does on Flickr30k. The reason may be that the MS-COCO dataset has far more training data than Flcikr30k. With the support of a large number of samples, the network itself has the ability to mine latent semantic information, but the additional supervision information provided by our method cannot cause significant improvement.

In addition, we also mentioned above that retrieval efficiency is also a very important indicator in the retrieval field. Taking testing on the Flickr30k dataset as an example, we compare our S-VSE model with the global embedding model VSE++ and the local alignment model SCAN. Table 3 shows the results of experiments. The "Feature Extraction Time" in the table refers to the total time (second) used to calculate the image global representation and text global representation of the 5000 image-text pairs on the Flickr30k test set. "Retrieval Time" in the table refers to the total time taken to calculate the similarity of 5,000 image-text pairs using the obtained image-text representations and then perform bidirectional retrieval based on the similarity. It is worth noting that our S-VSE model does not need to extract the local features of the image text during testing, and only extracts the global representation of the image text to calculate the similarity.

Table 3. Retrieval efficiency comparison

Method	Feature extraction time	Retrieval time	Rsum
VSE++	24.3	13.2	409.8
S-VSE	25.4	13.2	444.5
SCAN	–	67.5	465.0

Our S-VSE model is equivalent to adding a self-attention module on the basis of VSE++ for the acquisition of text global representation after removing the image-text pair label generation branch. Observing Table 3, it can also be found that the feature extraction time of the VSE++ and S-VSE models is close, and both use the cosine similarity to obtain the similarity between the image representation and the text representation, so their retrieval time is the same. In contrast, SCAN calculates the final similarity through the interaction

Table 4. Results of ablation studies

Method	Image-to-text			Text-to-image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	43.7	71.9	82.1	32.3	60.9	72.1	363.0
VSE++(ft)	52.9	80.5	87.2	39.6	70.1	79.5	409.8
S-VSE-1	45.6	74.2	83.6	33.8	63.4	74.7	375.3
S-VSE-1(ft)	57.0	81.0	88.5	44.0	73.7	82.4	426.6
S-VSE-2	48.0	77.8	85.9	36.0	66.2	76.8	390.4
S-VSE-2(ft)	60.3	84.0	90.6	45.1	75.3	83.2	438.5
S-VSE	52.1	77.0	86.5	36.3	66.3	76.6	394.7
S-VSE(ft)	**62.9**	**85.3**	**91.5**	**45.8**	**75.4**	**83.6**	**444.5**

between the local features of the image and text, so SCAN spends a lot of time for retrieval. Although the retrieval accuracy of our proposed S-VSE is 20.5% lower than that of SCAN, our retrieval efficiency is higher and more suitable for practical production.

4.4 Ablation Study and Analysis

We perform some ablation studies in this section to prove the effectiveness of our proposed method. Specifically, we remove the image-text pair label generation branch in the S-VSE network, and denote the network that only contains the global embedding branch as S-VSE-1. Then remove the self-attention module in the S-VSE network, directly use the mean vector q as the final text representation, and denote the network as S-VSE-2. Table 4 shows the results of our ablation experiments on the Flickr30k dataset. Comparing VSE++ and S-VSE-1, it can be seen that the introduction of self-attention module can bring significant improvement without label information optimization. Comparing S-VSE-1 and S-VSE we can find that additional label information can also significantly improve model performance. However, comparing S-VSE-2 and S-VSE, it can be seen that removing the self-attention module under the action of supervision information does not cause a significant reduction in performance.

5 Conclusion

In this paper, we propose a novel self-supervised visual-semantic embedding network (S-VSE) based on local label optimization. Our S-VSE network uses the label information of the image regions and the interaction between the image-text pairs to obtain the label of the entire image-text pair, and uses this supervision information to optimize the embedding space. Furthermore, we also introduce a self-attention module for enhancing feature representation. The experimental

results on two benchmark datasets also demonstrate that our method can significantly improve the retrieval performance of the global visual-semantic embedding model while ensuring the retrieval efficiency.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
2. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis **28**, 1247–1255 (2013)
3. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014)
4. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database, pp. 248–255 (2009)
5. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives (2018)
6. Hardoon, D.R., Szedmák, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions, pp. 3128–3137 (2015). https://doi.org/10.1109/CVPR.2015.7298932
9. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**, 32–73 (2016)
10. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. arXiv abs/1803.08024 (2018)
11. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval, pp. 4242–4251 (2018)
12. Liu, C., Mao, Z., Liu, A., Zhang, T., Wang, B., Zhang, Y.: Focus your attention: a bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia (2019)
13. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2156–2164 (2017)
14. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2015)
15. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings, pp. 5005–5013 (2016)
16. Wang, Y., et al.: Position focused attention network for image-text matching, pp. 3792–3798 (2019)
17. Wang, Z., et al.: Camp: cross-modal adaptive message passing for text-image retrieval. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5763–5772 (2019)
18. Wen, K., Gu, X., Cheng, Q.: Learning dual semantic relations with graph attention for image-text matching. IEEE Trans. Circuits Syst. Video Technol. **31**, 2866–2879 (2021)

19. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)
20. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval, pp. 1618–1625 (2017)
21. Zhang, Q., Lei, Z., Zhang, Z., Li, S.: Context-aware attention network for image-text retrieval. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3533–3542 (2020)
22. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10386–10395 (2019)