Super Visual Semantic Embedding for Cross-Modal Image-Text Retrieval

Zhixian Zeng
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
zeng_zhixian@yeah.net

Jianjun Cao*
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
caojj@nudt.edu.cn

Guoquan Jiang
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
Jianggq17@nudt.edu.cn

Nianfeng Weng
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
wengnf17@nudt.edu.cn

Yuxin Xu
School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, China
2801343036@qq.com

Zibo Nie
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
2427517759@qq.com

ABSTRACT

Visual semantic embedding network or cross-modal cross-attention network are usually adopted for image-text retrieval. Existing works have confirmed that both visual semantic embedding network and cross-modal cross-attention network can achieve similar performance, but the former has lower computational complexity so that its retrieval speed is faster and its engineering application value is higher than the latter. In this paper, we propose a Super Visual Semantic Embedding Network (SVSEN) for cross-modal image-text retrieval, which contains two independent branch substructures including the image embedding network and the text embedding network. In the design of the image embedding network, firstly, a feature extraction network is employed to extract the fine-grained features of the image. Then, we design a graph attention mechanism module with residual link for image semantic enhancement. Finally, the Softmax pooling strategy is used to map the image fine-grained features to a common embedding space. In the design of the text embedding network, we use the pre-trained BERT-base-uncased to extract context-related word vectors, which will be fine-tuned in training. Finally, the fine-grained word vectors are mapped to a common embedding space by a maximum pooling. In the common embedding space, a soft label-based triplet loss function is adopted for cross-modal semantic alignment learning. Through experimental verification on two widely used datasets, namely MS-COCO and Flickr-30K, our proposed SVSEN achieves the best performance. For instance, on Flickr-30K, our SVSEN outperforms image retrieval by 3.91% relatively and text retrieval by 1.96% relatively (R@1).

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Retrieval models and ranking**;

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487167

KEYWORDS

Cross-modal image-text retrieval, Visual semantic embedding, Feature extraction, Pooling strategy, Common embedding space, Cross-modal triplet loss

ACM Reference Format:
Zhixian Zeng, Jianjun Cao*, Guoquan Jiang, Nianfeng Weng, Yuxin Xu, and Zibo Nie. 2021. Super Visual Semantic Embedding for Cross-Modal Image-Text Retrieval. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3487075.3487167

1 INTRODUCTION

With the development of the mobile Internet, a large amounts of data are generated all the time, including image, text, video, audio, and so on. People's demand for different information is becoming more and more urgent in the real world, including image-text retrieval. Image-text retrieval is one of the important content, which is aim to retrieve images with the same or similar semantic text, or to search for text with the same or similar semantic image, et al [1].

Generally, image-text retrieval methods can be divided into traditional methods and deep learning method. Among the traditional methods, Canonical Correlation Analysis (CCA) [2] is the most representative method, which constructs the correlation between different modal information by learning the mapping matrix. It is not doubted that there are also many improvements on CCA. However, in the field of visual semantic embedding, a lot of works have confirmed that the methods based on deep learning are better than the method based on CCA because the deep learning methods have better nonlinear fitting ability [1]. In deep learning methods, an artificial neural network is usually designed to learn cross-modal image-text information representation and map the image and text to a common embedding space for semantic alignment learning.

Due to the strong nonlinear learning ability of the deep neural network, it has become the mainstream method of image-text retrieval [1]. In cross-modal image-text retrieval, it can be divided into two categories. The one is based on the improvements of the visual semantic embedding framework proposed by Kiros et al. [3], and the other is the improvements of the cross-modal cross-attention

framework proposed by Lee et al. [4]. Recently, Chen et al. [5] have confirmed that the visual semantic embedding framework and the cross-modal cross-attention framework can achieve similar performance, but the former has lower computational complexity and can be deployed in actual scenarios. Specifically, the computational complexity of visual semantic embedding method is O(N), but the other is O(N^2).

In our opinion, under the condition that we keep the feature extraction method unchanged, the existing visual semantic embedding for image-text retrieval mainly improves performance by improving the three aspects including semantic enhanced representation module [6], pooling strategy module [5], and loss function [7]. Since the text feature extraction network uses a sequence-to-sequence model, which can effectively extract the context information about it, the contextual word vectors do not require semantic enhancement. Firstly, the semantic enhanced representation module mainly is used to enhance the fine-grained features of the image, and build the relationship between different fine-grained features. Then, the fine-grained features are used to obtain a global feature representation by the pooling strategy. Moreover, the loss function is adopted to promote the optimization of the model so that the cross-modal image-text similarity with similar semantics is as large as possible, and the cross-modal image-text similarity with a great difference is as small as possible.

In this paper, we propose a Super Visual Semantic Embedding Network (SVSEN), which use the visual semantic embedding framework and improve the performance of image-text retrieval by improving the above three aspects. As shown in Figure 1, we mainly improve the semantic enhanced representation module, pooling strategy module, and loss function to increase the performance of cross-modal image-text retrieval. In terms of the specific implementation, Firstly, we designed a semantic enhanced representation module combining graph attention mechanism [8] with residual structure [9]. Then, a Softmax pooling strategy module [10] is used to generate global feature representation. Finally, a cross-modal triplet loss function based on soft label is adopted to guide the model optimization. Our contributions can be listed below:

- We propose a Super Visual Semantic Embedding Network (SVSEN) by using the general visual semantic embedding framework.
- We improve the performance of visual semantic embedding by improving the semantic enhanced representation module, pooling strategy module, and loss function, which will be proved to be effective in the Section 4.4 ablation study.
- Experiments on MS-COCO and Flickr-30K datasets show that comparing with the best existing visual semantic embedding model, our proposed SVSEN achieves the best performance, and comparing with the method using different modules in ablation study, it is also proved that the performance of image-text retrieval can be improved by improving the above three aspects including semantic enhanced representation module, pooling strategy module, and loss function.

2 RELATED WORKS

The neural network models for cross-modal image-text retrieval mainly include visual semantic embedding networks and cross-modal cross-attention networks, which will be introduced in this section. Moreover, the loss functions commonly used in cross-modal image-text retrieval also will be introduced.

2.1 Visual Semantic Embedding for Cross-Modal Image-Text Retrieval

The visual semantic embedding model general contains two independent embedding branch networks, which are the image embedding network and text embedding network. The two branch networks learn from each other through the loss function. Kiros et al. [3] proposed a unified visual semantic embedding framework with a general cross-modal triplet loss function for the first time, which used sequence-to-sequence network as text embedding network and convolutional neural network as image embedding network. Li et al. [6] proposed a visual semantic reasoning network, which used a graph convolutional neural network [11] with residual link [9] to enhance visual features, and used Bi-GRU [13] network to extract the global image and text features. Recently, Chen et al. [5] found that a simple pooling strategy can achieve a good performance in visual semantic embedding, and proposed a Generalized Pooling Operator (GPO) for visual semantic embedding, which achieved the best performance in visual semantic embedding for image-text retrieval. However, in our opinion, neither the semantic enhanced representation module proposed in literature [6] nor the pooling strategy module proposed in literature [5] is the best method, and there is still room for improvement.

2.2 Cross-Modal Cross-Attention for Cross-Modal Image-Text Retrieval

If we add a cross-modal cross-attention mechanism to the visual semantic embedding model, it makes the model to construct the correlations between different modal information, while the visual semantic embedding can only construct the correlation within the modal. Lee et al. [4] proposed a stacked cross-attention network for image-text retrieval, which used a cross-attention mechanism to construct a two-way semantic association between image and text. Peng et al. [14] used an ensemble learning method and proposed a cross-media bi-attention mechanism for visual semantic alignment. Chen et al. [15] proposed an iterative matching structure, which used the recurrent attention memory module to construct cross-modal deep fine-grained semantic associations. However, due to the introduction of the cross-modal cross-attention mechanism, its computational complexity is increased, resulting in a lower practical application value in engineering practice. It has been confirmed by literature [5] that the computational complexity of cross-modal cross-attention is O(N^2), but the computational complexity of visual semantic embedding is O(N). Fortunately, literature [5] also proved that these two can achieve almost similar performance.

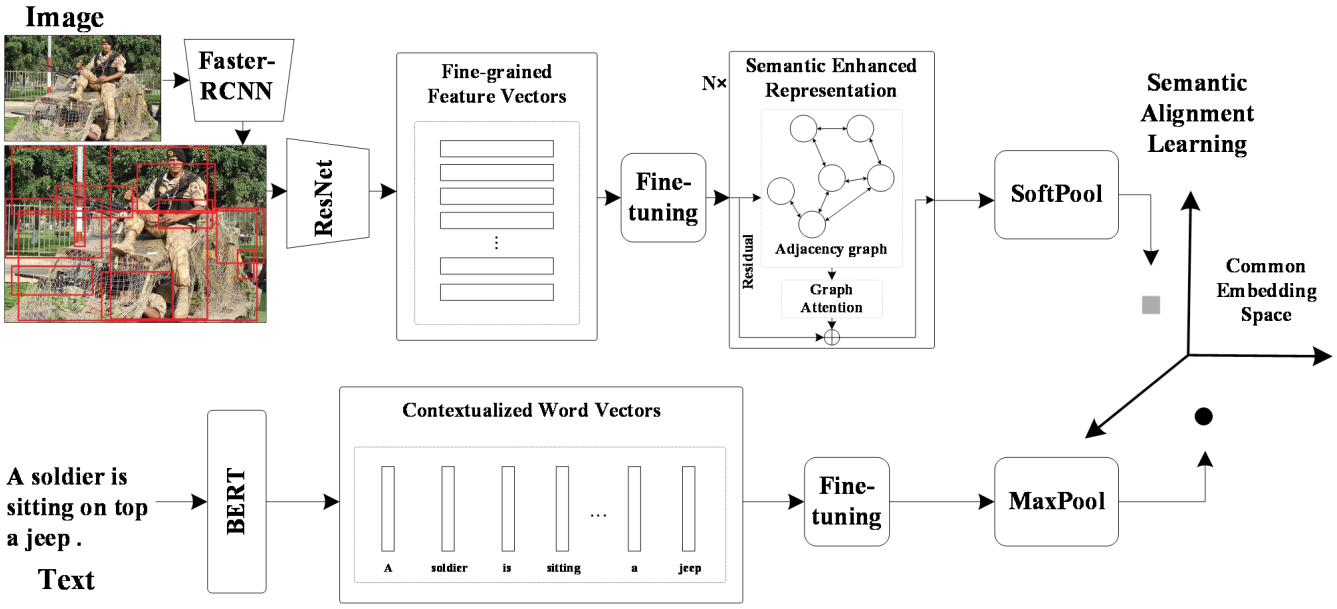

Figure 1: An Overview of Our Proposed Super Visual Semantic Embedding Network (SVSEN) Architecture.

2.3 Loss Functions for Cross-Modal Image-Text Retrieval

The loss function is a degree measure of model optimization. It is no doubt that the loss function has a great impact on the performance of the model, including convergence speed and generalization ability. The design of the loss function has a great influence on the generalization ability and convergence speed of the model. Kiros et al. [3] proposed a general cross-modal triplet loss function, which calculated all triples in each mini-batch indiscriminately. Although convergence rate of the general cross-modal triplet loss function is fast, its generalization ability is poor. Therefore, in order to improve the generalization ability of the model, Lee et al. [7] proposed a cross-modal triplet loss function with hard negative samples (HTL), which maked the model only focus on the negative samples with the greatest similarity to the anchor point in each mini-batch. Due to the excellent generalization ability of the cross-modal loss function with hard negative samples, it has become a commonly used loss function. However, because the cross-modal triplet loss function with hard negative samples reduces the number of triplet samples when calculating the loss, it makes the convergence of the model is slow. Chen et al. [5] used warm-up to solve this problem. Moreover, Chen et al. [16] also proposed an adaptive offline quintuplet loss for cross-modal image-text retrieval. Liu et al. [17] proposed a hub-ness-aware loss function for cross-modal image-text matching. Because the cross-modal triplet loss function is easy to implement, the existing methods generally use the cross-modal triplet loss function with hard negative samples, but its convergence speed is slow. Therefore, it is necessary to improve the convergence speed while maintaining the accuracy.

3 OUR SUPER VISUAL SEMANTIC EMBEDDING NETWORK

As shown in Figure 1, our proposed Super Visual Semantic Embedding Network (SVSEN) consists of four parts: Feature Extraction and Fine-tuning, Semantic Enhanced Representation, pooling strategy, and semantic alignment learning in a common embedding space. Next, we will introduce these four parts in detail.

3.1 Feature Extraction and Fine-tuning

3.1.1 Image Feature Extraction and Fine-tuning. In order to retain more fine-grained information, we use Faster-RCNN [18] for target recognition, which is pre-trained on the Visual Genome dataset [19] by Anderson et al. [21]. The image regions with top-36 confidence in target recognition is selected. ResNet101 [9] pre-trained on ImageNet [22] is used to extract features from the top-36 image regions, and obtain the 2048-dimensional features of the Pool5 layer. Then, the features are fine-tuned by a Fully Connected (FC) network and a Multi-Layer Perceptron (MLP) with residual connections, which is same as the literature [5], and mapped to a 1024-dimensional common embedding space. It should be noted that the pre-trained Faster-RCNN and ResNet101 are not involved in training. The above process can be expressed as Equation 1).

$$\begin{aligned} B &= \text{ResNet101}(\text{FasterRCNN}(Image)), \\ f_i &= \text{MLP}(b_i) + \text{FC}(b_i), \end{aligned} \tag{1}$$

Where $B = \{b_1, b_2, ..., b_{36}\}$, $F = \{f_1, f_2, ..., f_{36}\}$, B is the feature set extracted from each image; f_i is the feature representation by fine-tuning the fine-grained feature b_i.

3.1.2 Text Feature Extraction and Fine-tuning. In the fine-grained feature extraction of text, we use the pre-trained BERT-base-uncased [23] with its default parameter settings to extract the contextualized word vectors representation with 768 dimensions. Then, a fully connected layer is designed to fine-tune the word vectors and map them to a 1024-dimensional common embedding space. It should be noted that BERT-base-uncased will be trained but its learning rate is one-tenth of the overall network. The above process can be expressed as Equation 2).

$$\begin{aligned} C &= \text{BERT\_base\_uncased}(Text), \\ h_k &= \text{FC}(c_k), \end{aligned} \tag{2}$$

Where $C = \{c_1, c_2, ..., c_k, ..., c_n\}$, $H = \{h_1, h_2, ..., h_k, ..., h_n\}$, C is the feature set extracted from each text; n is the length of text; h_k is the feature representation by fine-tuning the fine-grained feature c_k.

3.2 Semantic Enhanced Representation

In the design of the semantic enhanced representation module, we use the graph attention network (GAT) [8] with residual connection [9], which is better than the semantic enhanced representation module with graph convolutional network (GCN) [11] in the literature [6]. In the specific implementation, the adjacency matrix is constructed, which is selected according to the similarity between the fine-grained image regions. the adjacency nodes are set to connected in the top 10% similarity ranking, which is set to 1, and the other nodes are set to disconnected, which is set to 0. The calculation process of similarity is shown in Equation 3), which is also same as the literature [6]

$$\begin{aligned} \phi(f_i) &= W_\phi f_i, \\ \psi(f_j) &= W_\psi f_j, \\ R(f_i, f_j) &= \phi(f_i)^T \phi(f_i), \end{aligned} \tag{3}$$

Where W_ϕ and W_ψ are parameters of fully connected layer that can be learned and adjusted; $R(f_i, f_j)$is the similarity between features f_i and f_j. If $R(f_i, f_j)$ ranks in the top 10% in all similarity of adjacency nodes, f_i is connected with f_j.

After constructing the adjacency matrix, we apply GAT with residual connection to enhance the fine-grained image features. A multi-head attention mechanism is adopted by GAT, where the number of heads K is set to 3, and the average method is used to integrate the output of the multi-head attention mechanism. The specific calculation can be expressed as Equation 4).

$$\begin{aligned} z_i^{k(l)} &= W^{k(l)} f_i^{(l)}, \\ e_{ij}^{k(l)} &= \text{LeakyRELU}(\overrightarrow{a}^{k(l)^T}(z_i^{k(l)} \parallel z_j^{k(l)})), \\ \alpha_{ij}^{(l)} &= \frac{\exp(e_{ij}^{k(l)})}{\sum_{p \in N(i)} \exp(e_{ip}^{k(l)})}, \\ f_i^{(l+1)} &= \sigma(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N(i)} \alpha_{ij}^{(l)} z_{ij}^{k(l)}), \\ f_i^* &= W_r f_i^{(l+1)} + f_i^{(l)}, \end{aligned} \tag{4}$$

Where l represents the l-th layer graph attention network; $W^{k(l)}$and $\overrightarrow{a}^{k(l)}$are the weights of the fully connected layer in the multi-head attention mechanism; $W_r$is the weights of the residual structure; $N(i)$ is the adjacent node set of node i; $\parallel$represents concatenation; LeakyRELU is the activation function; $\sigma$is the Sigmoid activation function; α is the attention weight; $z_i$is the intermediate state representation, and f_i is the input of the l-th layer graph attention network; f_i^* is the output of the graph attention network with the residual structure, namely Semantic Enhanced Representation module.

As shown in Figure 1, the layer of our semantic enhanced representation module is N. In our Super Visual Semantic Embedding Network, we set N to 3. Moreover, we will demonstrate that our Semantic Enhanced Representation module with GAT is superior to the existing methods in Section 4.4 ablation study.

3.3 Pooling Strategy

Pooling strategy can reduce the dimensionality of feature representation, improve feature invariance, prevent overfitting, and improve the generalization ability of the model, while retaining as much feature information as possible. Since Softmax Pooling (SoftPool) [10] can retain more fine-grained features of the image, which is also the best pooling strategy in the field of image classification, we adopt SoftPool to obtain a global feature representation for image. More, Max Pooling (MaxPool) can retain as much texture information as possible. Therefore, MaxPool is adopted to reduce the word vectors dimension of the text and obtain a global feature representation. Moreover, we will prove that our pooling combination is better than the best visual semantic embedding pooling strategy [5] in ablation study.

3.4 Semantic Alignment Learning in a Common Embedding Space

The optimized goal of our network is to make the cross-modal image-text similarity with similar semantics as large as possible, and the cross-modal image-text similarity with different semantics as small as possible. We use dot product to calculate the cross-modal image-text similarity. In the design of the loss function, we adopt a soft label-based cross-modal triplet loss function. Our loss function has certain similarities with the literature [17]. The difference is that the triples between different anchor points have different contributions to our loss, while the triples between different anchor points make the same contributions to the loss proposed by literature [17]. Our triplet loss is shown in Equation 5).

$$\begin{aligned} L = &\frac{1}{\beta} \log \sum_{n=1}^{N} e^{\frac{\beta}{\gamma} \log \sum_{t_n^-} e^{\gamma(\max(0, \lambda - S(i_n^+, t_n^+) + S(i_n^+, t_n^-)))}} \\ &+ \frac{1}{\beta} \log \sum_{n=1}^{N} e^{\frac{\beta}{\gamma} \log \sum_{i_n^-} e^{\gamma(\max(0, \lambda - S(i_n^+, t_n^+) + S(i_n^-, t_n^+)))}} \end{aligned} \tag{5}$$

Where β and γ are scale factors; λ is margin; i_n^+ and t_n^+ are positive samples or anchor points in the triplet; i_n^- and t_n^- are the negative samples in the triplet; N is the number of image or text anchor points in each mini-batch.

Corresponding to Section 2.3, our loss function has faster convergence speed and better accuracy, which will be proved in section 4.4

Table 1: Division of MS-COCO and Flickr-30K Datasets

Dataset	Total	Training	Validation	Testing
MS-COCO [12]	123287	113287	5000	5000
Flickr-30K [20]	31783	29783	1000	1000

4 EXPERIMENTS

4.1 Datasets

In order to verify the effectiveness of our Super Visual Semantic Embedded Network, we adopt two datasets commonly used in cross-modal image-text retrieval, namely MS-COCO [12] and Flickr-30K [20]. The division of datasets is same as [5-7, 14, 15], which is shown in Table 1. It should be noted that each image contains five independent text captions.

4.2 Experimental Environment and Hyper-parameter Settings

Our operating system is CentOS 7 64. Our CPU is Intel(R) Xeon(R) CPU E 5-2630 v4 @ 2.20GHz. We use two parallel Nvidia Tesla P40 GPUs with 24G memory, and our development environments are Python 3.6 .5 and Torch1.6.0.

In the experiment, we use the Adam optimizer. We set the initial learning rate to 5e-4 but set the initial learning rate of pre-trained BERT-base-uncased to 5e-5. The learning rate is decayed by 0.1 for every 15 epochs. Our network will be trained for a total of 25 epochs. The batch size is set to 128. We set the layers of semantic enhanced representation module to 3. We set β=1, γ=256, and λ=0.2in loss function.

4.3 Evaluation Metric

Same as works of literatures [5-7, 14, 15], we adopt R@k and R@sum, where R@k and R@sum are local performance indicator and global overall performance indicator, respectively. R@sum is the sum of R@k. R@k can be expressed as Equation 6).

$$\text{R@}k = \frac{1}{M}\sum_{x=1}^{M} Rel_x, \tag{6}$$

Where M is the number of instances in the testing set, and Rel_x indicates whether there are correct search results in top-k. If there is a correct result among the top-k results, Rel_x is set to 1, otherwise, 0.

4.4 Experimental Results and Analysis

In order to verify effectiveness of our proposed method, we compare our SVSEN with 4 state-of-the-art methods, which can be briefly introduced as follows:

VSE++ [7] proposes a cross-modal triplet loss function with hard negative samples (HTL), which makes the model only focus on the negative samples with the greatest similarity to the anchor point in each mini-batch.

VSRN [6] constructs a visual semantic reasoning network, which uses a graph convolutional neural network with residual link to enhance visual features, and uses Bi-GRU network to extract the global image and text features.

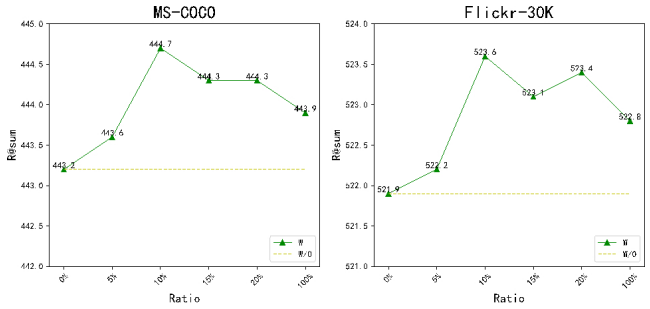

Figure 2: Impact of the Selected Ratio of Adjacent Nodes on Semantic Enhanced Representation in Terms of R@sum.

IMRAM [15] is a method based on a cross-modal cross-attention mechanism. It constructs an iterative matching structure, which uses the recurrent attention memory module to construct cross-modal deep fine-grained semantic associations.

VSE+GPO [5] proposes a Generalized Pooling Operator and achieve a good performance in visual semantic embedding.

As shown in Table 2 and Table 3, our proposed SVSEN achieves the best performance. Comparing with the best method VSE+GPO [5], in terms of local performance R@k, our SVSEN outperforms image retrieval by 5.90% and 3.91% at R@1, respectively, on MS-COCO and Flickr-30K datasets, and in terms of global overall performance R@sum, our SVSEN outperforms by 2.39% and 1.97%, respectively, on MS-COCO and Flickr-30K. Moreover, without a doubt, our method is also better than other cross-modal retrieval methods, such as VSE++ [7], VSRN [6], and IMRAM [15]. All of the above results show that our proposed Super Visual Semantic Embedding Network has significant advantages, thanks to that we improve the semantic enhancement representation, pooling strategy, and cross-modal triplet loss function. In the next section, we will also verify the advanced nature of our three modules including semantic enhancement representation, pooling strategy, and cross-modal triplet loss function.

4.5 Ablation Study

4.5.1 Impact of the Adjacency Matrix of Graph Attention on Performance. In order to verify the impact of the selected ratio of adjacent nodes on semantic enhanced representation, we selected the top 5%, 10%, 15%, 20%, 0%, and 100% of the similarity ranking as adjacent nodes for experiments. Among them, 0% means that we don't adopt semantic enhanced representation, and 100% means that the adjacency matrix is represented as a full connected graph. The experimental results are shown in Figure 2

From the results in Figure 2, it can be seen that the use of semantic enhanced representation (100%) is better than not using it (0%). Our

Table 2: Quantitative Evaluation Results of Cross-Modal Image-Text Retrieval on MS-COCO (5K)

Method	Image retrieval			Text retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Faster-RCNN+BiGRU							
VSE++$_{2018}$ [7]	31.7	61.8	74.2	42.9	74.5	85.1	370.2
VSRN$_{2019}$ [6]	40.5	70.6	81.1	53.0	81.1	89.4	415.7
IMRAM$_{2020}$ [15]	39.7	69.1	79.8	53.7	83.2	91.0	416.5
VSE+GPO$_{2021}$ [5]	39.3	69.9	81.1	56.6	83.6	91.4	421.9
Faster-RCNN+BERT							
VSE++$_{2018}$ [7]	31.0	61.3	73.7	42.1	72.6	83.9	364.7
VSE+GPO$_{2021}$ [5]	42.4	72.7	83.2	58.3	85.3	92.3	434.3
Our: SVSEN	44.9	74.6	84.3	60.7	86.7	93.5	444.7

Table 3: Quantitative Evaluation Results of Cross-Modal Image-Text Retrieval on Flickr-30K (1K)

Method	Image retrieval			Text retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Faster-RCNN+BiGRU							
VSE++$_{2018}$ [7]	45.7	73.6	81.9	62.2	86.6	92.3	442.3
VSRN$_{2019}$ [6]	54.7	81.8	88.2	71.3	90.6	96.0	482.6
IMRAM$_{2020}$ [15]	53.9	79.4	87.2	74.1	93.0	96.6	484.2
VSE+GPO$_{2021}$ [5]	56.4	83.4	89.9	76.5	94.2	97.7	498.1
Faster-RCNN+BERT							
VSE++$_{2018}$ [7]	45.6	76.4	84.4	63.4	87.2	92.7	449.7
VSE+GPO$_{2021}$ [5]	61.4	85.9	91.5	81.7	95.4	97.6	513.5
Our: SVSEN	63.8	87.5	93.1	83.3	97.2	98.7	523.6

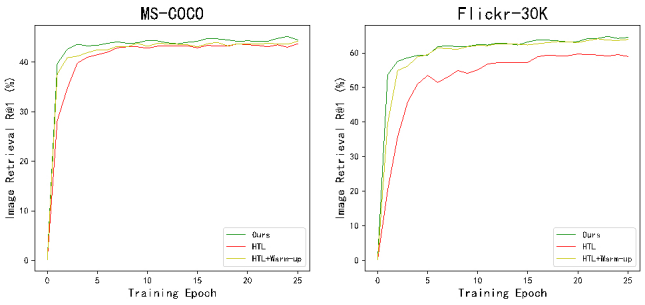

Figure 3: Impact of Different Loss Functions on MS-COCO and Flickr-30K at R@1 (Image Retrival).

SVSEN achieves the best performance when selecting the top 10% of similarity ranking as adjacent nodes.

4.5.2 Impact of Different Loss Functions. In order to verify the advantages of our loss function, experiments with three loss functions are conducted in our SVSEN model, including HTL without a warm-up, HTL with a warm-up, and our loss function. Figure 3 shows the verification accuracy in image retrieval at R@1 after each epoch training under different loss functions. We can see that our loss function has a faster convergence speed and higher accuracy. This is because the more triples in the mini-batch are considered when calculating the loss, and pay different attention to them.

4.5.3 Impact of Different Modules on Performance. In order to verify the effectiveness that we improve semantic enhanced representation (SER), pooling strategy (PS), and cross-modal triplet loss function (CMTLF), three comparative experiments are conducted. In the first one, we carried out experiments using the graph convolutional network with residual link (Reslink) proposed by VSRN [6] to replace the graph attention with residual structure in our SVSEN. Second, we use the GPO+GPO pooling combination in VSE+GPO [5] to replace our SoftPool + MaxPool pooling combination. Third, we use the cross-modal triplet loss function with hard negative samples using warm-up (HTL+Warm-up) proposed by VSE++ [7] to replace our loss function. The experimental results are shown in Table 4

As shown in Table 4, comparing with three comparative experiments, which replace the corresponding modules with other existing modules in our SVSEN, it can be seen that our improvements to the three modules can effectively improve the performance of the cross-modal image-text retrieval. The results show that the performance of the model can be effectively improved by improving the semantic enhanced representation, pooling strategy combination, and loss function.

5 CONCLUSIONS

In this paper, we propose a Super Visual Semantic Embedding Network (SVSEN) for cross-modal image-text retrieval, which includes four parts: Feature Extraction and Fine-tuning, Semantic Enhanced

Table 4: Quantitative Evaluation Results in terms of R@sum of Different Modules, including Semantic Enhanced Representation, Pooling Strategy, and Loss Function

Module	Dataset	
	MS-COCO	Flickr-30K
SER: GAT + Reslink [6]	443.6	522.4
PS: GPO + GPO [5]	442.3	522.0
CMTLF: HTL + Warm-up [7]	441.9	521.6
Our: SVSEN	444.7	523.6

Representation, pooling strategy, and semantic alignment learning in a common embedding space. We propose a semantic enhanced representation module, which combines with graph attention and residual link, a pooling strategy, which combines with Softmax Pooling and Max Pooling, and cross-modal triplet loss function, which is based on soft labels, respectively. Experiments on MSCOCO and Flickr-30K datasets show that our proposed network is better than others. Moreover, Ablation study also show that the performance of our SVSEN can be improved by improving the semantic enhanced representation module, pooling strategy module, and cross-modal triplet loss function.

Our work has a certain reference significance. In the future related research, visual semantic embedding should pay more attention to three aspects, including semantic enhanced representation module, pooling strategy module, and cross-modal triplet loss function.

ACKNOWLEDGMENTS

Our research is supported by the National Natural Science Foundation of China (No.61371196).

REFERENCES

[1] Pengfei Du, Xiaoyong Li and Yali Gao (2021). Survey on multimodal visual language representation learning. Journal of Software, 32(2), 327-348.
[2] Hotelling Harold (1992). Relations between two sets of variates. Breakthroughs in statistics. 23(7): 162-190.
[3] Ryan Kiros, Ruslan Salakhutdinov and Richard S. Zemel (2014). Unifying visual-semantic embeddings with multimodal neural language Models. Computer Science.
[4] KuangHuei Lee, Xi Chen, Gang Hua, *et al.* (2018). Stacked cross attention for image-text matching. Computer Vision ECCV 2018 15th European Conference. Springer, Munich, Germany.
[5] Chen Jiacheng, Hu Hexiang, Wu Hao, *et al.* (2021). Learning the best pooling strategy for visual semantic embedding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Virtual.
[6] Li Kunpeng, Zhang, Yulun, Li, Ka, *et al.* (2019). Visual semantic reasoning for image-text matching. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.
[7] KuangHuei Lee, Xi Chen, Gang Hua, *et al.* (2018). Vse++: improving visual-semantic embeddings with hard negatives. in British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK.
[8] Veličković Petar, Cucurull Guillem, Casanova Arantxa, *et al.* (2018). Graph Attention Networks. International Conference on Learning Representations.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, *et al.* (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, NV, USA.
[10] Stergiou, Alexandros, Poppe, Ronald and Kalliatakis Grigorios (2021). Refining activation downsampling with SoftPool. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Virtual.
[11] Thomas N. Kipf and Max Welling (2017). Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR 2017. OpenReview.net, Toulon, France.
[12] Peng Zhou, Wei Shi, Jun Tian, *et al.* (2016). Attention based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. The Association for Computer Linguistics, Berlin, Germany.
[13] Peng Yuxin, Qi Jinwei and Zhuo Yunkan (2020). MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism. IEEE Transactions on Image Processing, 29(4): 2728-2741.
[14] Hui Chen, Guiguang Ding, Xudong Li, *et al.* (2020). Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. 2020 Conference on Computer Vision and Pattern Recognition. IEEE, Seattle, WA, USA.
[15] Tianlang Chen, Jiajun Deng, and Jiebo Luo (2020). Adaptive offline quintuplet loss for image-text matching. Computer Vision-ECCV 2020-16th European Conference. Springer, Glasgow, UK.
[16] Fangyu Liu, Rongtian Ye, Xun Wang, *et al.* (2020). Hal: Improved text-image matching by mitigating visual semantic hubs. The Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI press, New York, NY, USA.
[17] Shaoqing Ren, Kaiming He, Ross B. Girshick, *et al.* (2018). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6): 1137-1149.
[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, *et al.* (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. 123(1), 32-73.
[19] Peter Anderson, Xiaodong He, Chris Buehler, *et al.* (2018). Bottom-up and top-down attention for image captioning and visual question answering. 2018 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA.
[20] Jia Deng, Wei Dong, Richard Socher, *et al.* (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, Miami, Florida, USA.
[21] Jacob Devlin, Mingwei Chang, Kenton Lee, *et al.* (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, MN, USA.
[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongi, *et al.* (2014). Microsoft COCO: Common objects in context. Computer Vision-ECCV 2014-13th European Conference. Association for Springer, Zurich,Switzerland.
[23] Peter Young, Alice Lai, Micah Hodosh, *et al.* (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics. 2(1): 67-78.