SSM: Semantic Selection and Multi-view Alignment for Image-Text Retrieval

Beiming Yu, Zhenfei Yang, Xiushuang Yi(✉), Yu Wang, and Zhangjun Bao

School of Computer Science and Engineering, Northeastern University, Shenyang, China
xsyi@mail.neu.edu.cn, {2001828,2071737}@stu.neu.edu.cn

Abstract. Image-text retrieval has been a crucial and fundamental task in multi-modal field. Benefiting from the superiority of Transformer encoder in modeling multimodal information, the Transformer-based alignment model has become the mainstream of image-text retrieval. However, current Transformer-based alignment models suffer from two major limitations: (1) The redundancy of modal features and the complexity of correlations between modalities restrict the performance of the model. (2) Current researches are typically limited to a single viewpoint during the modal alignment. To address these issues, in this paper we propose a image-text retrieval model SSM based on Semantic Selection and Multi-view alignment. Specifically, we introduce a gated attention unit to filter unnecessary information, and design an adaptive weighted similarity calculation method to dynamically adjust the importance of different features during the alignment process. On the other hand, we design a multi-view cross-modal alignment method that considers different granularity and different level of information to provide complementary benefits in representation learning. We compare SSM with other advanced image-text retrieval models in MS-COCO and Flickr30K datasets, and the results show that the SSM model has competitive performance without much interaction.

Keywords: Image-text retrieval · Multi-modal · Semantic selection · Multi-view · Contrastive learning

1 Introduction

With the growth of multimedia data on the Internet, cross-modal retrieval has been widely noticed [20]. Cross-modal retrieval aims to understand the natural semantic correlations between different modalities and hence search for semantically similar instances of different modalities. As the core task of cross-modal retrieval, the challenge of image-text retrieval is to accurately learn the semantic relatedness between image and text, and bridge the semantic gap between the two heterogeneous modalities.

Supported by organization x.

X. Yang et al. (Eds.): ADMA 2023, LNAI 14179, pp. 139–153, 2023.
https://doi.org/10.1007/978-3-031-46674-8_10

Early researches on image-text retrieval focus on alignment-based models, which encode image and text independently as feature vector representations and calculate image-text matching score via a similarity function. Faghri et al. [6] encodes the image and text as a global feature vector and aligns the features by contrastive learning. Lee et al. [11] proposes a fine-grained feature alignment method to further improve the performance of the alignment-based model. However, these works remain very inefficient for large scale image-text retrieval, limited by the weakness of CNN and RNN feature encoding capability. Chen et al. [2] proposes an interaction-based model to match image and text features by multiple iterations of neural interaction units, which fully explores the semantic association between the two modalities. But the interaction-based model, while obtaining significant gains in retrieval performance, also leads to a dramatic increase in computational cost and poses challenges for practical deployment in production environments.

In recent years, the successful deployment of Transformer models in the natural language processing [3,27] and multimodal [5,14,26] has demonstrated the superiority of Transformer modeling visual and text information. Transformer employs a multi-head attention mechanism where each part of the input representation interacts with other parts, to obtain better feature representations. Messina et al. [16] improves the alignment model using Transformer and applies it to an image-text retrieval task. Remarkably, their methods maintain the fast inference speed of the alignment-based model while achieving performance close to that of more complex interaction-based models.

Although Transformer-based alignment method has achieved acceptable performance, the current study suffers from two major drawbacks, as shown in Fig. 1: (1) The correlations between image and text are usually complex. In a mutually matching image-text pair, the text may describe only the main content of the image, and an image may require multiple sentences to be described correctly. Therefore, not all regions of image and words of text have matching relationship, especially there will be some region features in image with low contribution to retrieval. Furthermore, current researches commonly employ the Faster-RCNN model to extract image features [1,2,11,16,17], it may lead to excessive border overlap and result in the extraction of image features with redundant information. (2) Multi-layer Transformer in the process of encoding features, the vectors encoded in different layers contain different levels of information [8,22]. For example, the lower layer tend to encode basic features, and the higher layer capture complex semantic information. The previous Transformer-based alignment models [7,17,28], which commonly use the output features of the last layer, ignore the semantic differences between different layers, and these model make limited exploitation of the transformer architecture. Meanwhile, previous models [16] typically focus on local features alignment and ignore the guiding role of global features, which may lead to ambiguous representation due to local features not fully integrated with contextual information.

In this paper, we propose a novel image-text retrieval model SSM. Referring to past work, our model employs Faster-RCNN and Transformer for image fea-

ture encoding and BERT pre-trained model for text encoding. To address the redundancy of modal features and the complexity of correlations between modalities, the SSM model introduces a gated attention mechanism to filter the redundant features in image modalities. In addition, we propose an adaptive weighted similarity calculation method to dynamically attend on representative features and cast aside the interferences of uninformative features in the alignment process. In order to integrate the modal features of different views and learn the ideal modal feature representation, we propose a multi-view cross-modal alignment method to align global features and local features at the semantic level and the feature level to achieve accurate matching of image-text pairs.

Summarizing, the contributions of this paper are the following:

(1) We introduce gated attention units and adaptive weighted similarity calculation method for cross-modal semantic selection.
(2) We propose a multi-view cross-modal alignment method that captures the modal correlations of different views.
(3) We have conducted extensive experiments on two benchmark datasets to validate the effectiveness of SSM. The experimental results show that our methods can significantly improve the metrics of cross-modal retrieval.

2 Related Work

2.1 Image-Text Retrieval

Image-text retrieval is a fundamental task in the field of multimodal where the target is to find a suitable text description for an image or to find a corresponding image for a given text. Existing approaches can be divided into two main types: alignment-based and interaction-based. Notably, due to its low computational cost and fast response speed, the alignment-based method has been widely used in industry and has attracted a lot of attention in academia. The alignment-based method leverages a neural network model to encode images and text as feature representations separately and performs inter-modal alignment by contrastive learning. However, the results achieved by alignment-based method in earlier researches are not satisfactory due to encoder performance limitations [6,10,11].

Benefiting from the excellent performance of the Transformer encoder, recent studies have applied it to cross-modal alignment. Messina et al. [17] first applies Transformer as a modal encoder for image-text alignment. Qu et al. [23] enhances the feature representation capability of the model by leveraging the BERT pre-trained model and the feature summarization module. Messina et al. [16] proposes a fine-grained alignment model based on Transformer encoder to align regions of image and words of text, and achieves approximate results with the interaction model of that time. The Transformer-based alignment model has achieved promising results. However, its retrieval accuracy still has much space for improvement. In this paper, we introduce a gated attention unit and an adaptive weighted similarity calculation method to better align the image-text semantics.

2.2 Contrastive Learning

Contrastive learning is a representation learning method. It essentially aims to learn a better representation of the input by maximizing agreement between two similar data samples. The concept of contrastive learning is widely applied in cross-modal alignment. Radford et al. [24] implements the idea of contrastive learning based on large scale image-text datasets, achieving excellent performance on several multi-modal downstream tasks. Shukor et al. [25] introduces a novel triplet losses with dynamic margins that adapt to the difficulty of the task. In this work we follow the line of previous work on image-text retrieval and use triplet loss as contrastive learning loss [2,6,11,16]. Different from previous work, we design a multi-view cross-modal alignment by considering the features of different Transformer layers and the information of different granularities to obtain a high-quality modal representation.

3 Methodology

The overall framework of SSM is shown in Fig. 2, it contains three parts: image encoder, text encoder, and alignment module. In this section, we elaborate our proposed methods. Firstly, we introduce the image encoder and text encoder in Sect. 3.2 and 3.3. We then describe the adaptive weighted similarity calculation method (AWS) for local feature alignment in Sect. 3.4. Finally, we introduce the multi-view cross-modal alignment method (MVA) and objective function for image-text retrieval in Sect. 3.5.

3.1 Problem Definition

Formally, given an image-text pair, the image is represented as a visual feature of regions $I = \{r_i | i \in [1, m]\}$ and the text is represented as a text feature of words $T = \{w_i | i \in [1, n]\}$, where m and n denote the number of image regions and text words, respectively. The object of the task is to evaluate the matching score between them, thus enabling cross-modal retrieval from the database.

3.2 Text Encoder

SSM uses the pre-trained BERT as a text encoder. Considering that image features are generated by pre-trained deep neural networks, this paper uses a deeper text encoder to model the semantic relationships between words. Concretely, for the input text, each word is mapped to the embedding representation $T^e = \{T_i^e | i \in [1, n]\}$ as the input to the text encoder. We add an embedding T_0^e at the first position to aggregate the global representation of the text. The text embedding T^e consists of three parts: word embedding, position embedding, and segment embedding.

$$T^e = W + P + S \tag{1}$$

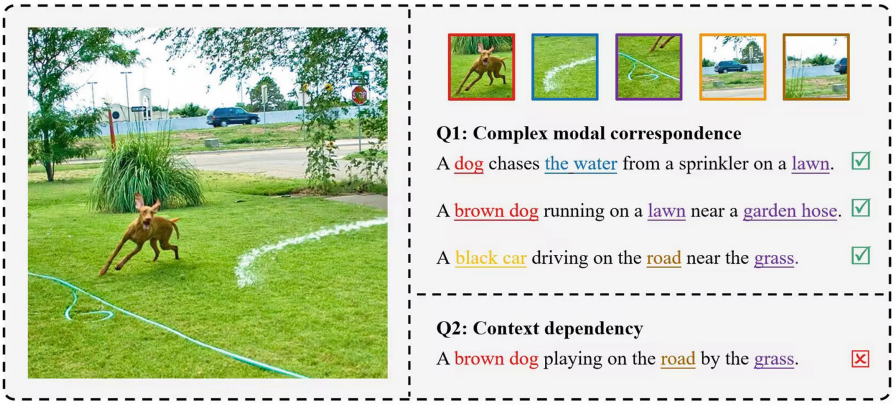

Fig. 1. Illustrations of major drawbacks for Image-Text Retrieval.

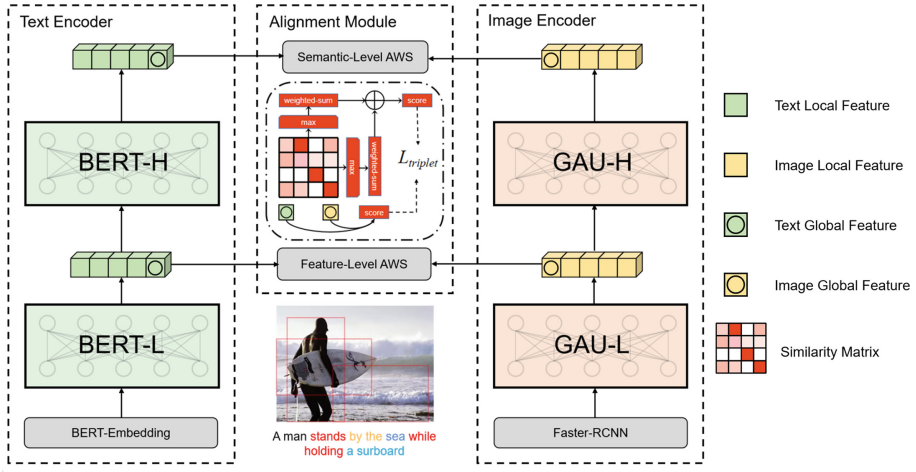

Fig. 2. Model framework of SSM.

In the Transformer architecture, different layers capture information with various semantic clues. SSM uses the first 8 layers of BERT as the low layer text encoder ($BERT_L$) to obtain a feature-level representation of the text $T^f = \{T_i^f | i \in [0, n]\}$.

$$T^f = BERT_L(T^e) \tag{2}$$

The last 4 layers of BERT are used as the high layer text encoder ($BERT_H$) to obtain the semantic-level features of the text $T^s = \{T_i^s | i \in [0, n]\}$.

$$T^s = BERT_H(T^f) \tag{3}$$

3.3 Image Encoder

Following recent work, we leverage Bottom-up Attention model to pre-extract features from image regions with high confidence [2,4,11]. Specifically, we pre-extract features from the image I by Faster-RCNN to obtain a set of visual sequence features $I^e = \{r_i | i \in [1, m]\}$. Notably, we add an embedding I_0^e at the first position of the visual sequence feature to aggregate the global representation.

In order to filter noise and redundant information in image features, we introduce the gated attention unit (GAU) as the image encoder. The gated attention unit, which is based on the Transformer architecture, controls the internal information flow through a gate mechanism to adaptively capture contextual information and refine high-quality image representations.

The GAU firstly projects I^e through three linear layers to obtain Q, K, and V, respectively.

$$\begin{cases} Q = W_q I^e + b_q \\ K = W_k I^e + b_k \\ V = W_v I^e + b_v \end{cases} \tag{4}$$

where Q, K and V respectively denote the query, key, and value, and W_q, W_k, W_v, b_q, b_k, b_v are learnable parameters.

The attention weight matrix $attn$ is then calculated. The formula is as follows:

$$attn = relu^2(\frac{QK^T}{\sqrt{d}}) \tag{5}$$

The GAU adds a gated linear unit for filtering unnecessary information. Specifically, I^e is linearly projected to obtain the gating weights U, after which the intermediate features I^h are obtained by the gated self-attention calculation.

$$U = W_u I^e + b_u \tag{6}$$

$$I^h = W_h(U \otimes attnV) \tag{7}$$

where W_u, W_h, b_u are learnable parameters.

Subsequently, the image features are mapped into the same vector space as the text feature dimension by linear projection layer. The feature-level representation $I^f = \{I_i^f | i \in [0, m]\}$ of the image is calculated as follows:

$$I^f = W_f I^h + b_f \tag{8}$$

where W_f, b_f are learnable parameters.

The SSM model uses another GAU module as a high layer encoder of image features to obtain the semantic-level image features I^s.

$$I^s = GAU_H(I^f) \tag{9}$$

3.4 Adaptive Weighted Similarity Calculation

For the image-text retrieval task, different regions of the image and different words of the text make different contributions to the image-text alignment, as shown in Fig. 3. In this paper, we devise an adaptive weight similarity calculation method (AWS) to balance the importance of different features in the similarity calculation.

First, given an image I and a text T, compute the similarity matrix $M \in R^{m \times n}$ between all regions and words.

$$M_{ij} = \frac{I_i^T T_j}{||I_i|| ||T_j||} \qquad i \in [1, m], j \in [1, n] \tag{10}$$

where M_{ij} denotes the similarity between the i-th region feature and the j-th word feature.

We use the combination of linear layer and $softmax$ function to measure the weight α of different features in similarity matching, this process is represented as:

$$\alpha = softmax(W_o I_i + b_o) \qquad i \in [1, m] \tag{11}$$

where W_o, b_o are learnable parameters.

The final image-to-text similarity score can be calculated as:

$$Sim_{i2t} = \sum\nolimits_{i=1}^{m} \alpha_i max_{j=1}^{n}(M_{ij}) \qquad i \in [1, m], j \in [1, n] \tag{12}$$

For text-to-image similarity calculation, we use the same calculation as above to obtain the word-region similarity score Sim_{t2i}. The final similarity score is calculated as follows:

$$Sim = Sim_{i2t} + Sim_{t2i} \tag{13}$$

3.5 Multi-view Alignment

In this paper, we propose a multi-view cross-modal alignment method (MVA) for image-text retrieval that combines the hierarchical and granular information. We use the low-layer information for feature-level multi-grained alignment, and the high-layer information is used for multi-grained alignment at the semantic-level.

Specifically, we perform feature-level alignment using the image features I^f and text features T^f obtained from the low-layer encoder. For the local features, we use the adaptive weighted similarity calculation method proposed above to obtain the image-text similarity matrix $S^{fl} \in R^{B \times B}$, where B denotes the batch size and S_{ij}^{fl} denotes the local matching score of the i-th image and the j-th text within the same batch at the feature-level. For the global features, we calculates the cosine similarity between the feature-level global representations I_0^f and T_0^f as follows:

$$S^{fg} = \frac{I_0^{fT} T_0^f}{||I_0^f|| ||T_0^f||} \tag{14}$$

where $S^{fg} \in R^{B \times B}$ and S_{ij}^{fg} denotes the feature-level global matching score of the i-th image and the j-th text within the same batch.

For the semantic-level alignment, we use the last layer features as the semantic features of the image and text and calculate the global similarity matrix S^{sg} and the local similarity matrix S^{sl}. We only consider the local similarity at the semantic-level during the model validation process, the similarity of the other views is only used for the calculation of the alignment loss during model training process.

In this paper, we use a triplet contrastive loss as the optimization objective. Following Faghri et al. [6], we focus the attention on hard negatives. Our triplet contrastive loss is defined as:

$$L^* = [\lambda + S_{ij'}^* - S_{i+}^*]_+ + [\lambda + S_{i'j}^* - S_{+j}^*]_+ \tag{15}$$

where $S^* \in \{S^{fl}, S^{fg}, S^{sl}, S^{sg}\}$, $[x]_+ = max(x, 0)$, S_{i+}^* denotes the similarity between the i-th image and the matched text, $S_{ij'}^*$ denotes the similarity between the i-th image and the hardest negative sample of text within the same batch. λ defines the minimum distance that should be maintained between a truly matched text-image positive sample pair and a negative pair. The hardest negative samples i' and j' are denoted as:

$$\begin{cases} i' = argmax(S_{*,j}) & i' \neq j \\ j' = argmax(S_{i,*}) & j' \neq i \end{cases} \tag{16}$$

The overall training objective of our model is:

$$L = L^{fl} + L^{fg} + L^{sl} + L^{sg} \tag{17}$$

4 Experiments

We evaluate our methods on two widely used benchmark datasets including MS-COCO [13] and Flickr30K [21], and compare the SSM model to current advanced models. We also conduct ablation studies to incrementally verify our methods.

4.1 Datasets

MS-COCO is a more general dataset for image-text retrieval, with a total of 123,287 images. Each image is given a set of 5 manual descriptions. Following the split by Karpathy and FeiFei [9] we utilize 5,000 images for validation and 5,000 images for testing and the rest for training. Flickr30K contains 31,783 images collected from social network, and each image is associated 5 captions. We use 1,000 images for validation, 1,000 images for testing and the rest for training.

4.2 Evaluation Metric

We measure the model performance with $R@k(k = 1, 5, 10)$ and $R@sum$. where $R@k$ denotes the fraction of queries for which the correct item is retrieved in the closest k points to the query, and $R@sum$ denotes the sum of recall rates of retrieval tasks.

4.3 Implementation Details

The SSM model is trained on an A5000 graphics card for 50 epochs. The batch-size is set to 30 for all experiments. The initial learning rate is set to 0.00001 and then decay to 0.1 times every 20 epochs. For the text, we utilize BERT for feature encoding, where the feature dimension is 768. For the image, we take the Faster-RCNN detector for feature pre-extraction. Each image has 36 region proposals, where the feature dimension is 2048. After feeding the region features into a GAU_L module, we add a linear layer to transform the GAU_L output to a 768-dimension vector. The layer of GAU_L and GAU_H is set to 4. The margin λ for the triplet contrastive loss is set to 0.2.

4.4 Main Results

We compare the model SSM proposed in this paper with other baseline models on two benchmark datasets, include the traditional alignment-based VSE++ [6] and SCAN [11], the interaction-based IMRAM [2] and SGRAF [4], and the Transformer-based alignment model TERN [17]. The results show that SSM significantly outperforms all other baseline models.

Table 1 shows the performance of SSM with the baseline model on the Flickr30k dataset, achieving 80.3%, 94.9%, and 98.2% for R@1, R@5 and R@10 in text retrieval, the metric on image retrieval is 62.3%, 85.9%, and 91.4% for R@1,

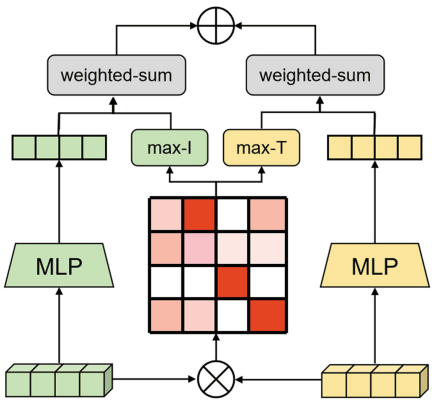

Fig. 3. Adaptive weighted similarity calculation method (AWS).

Table 1. The experimental results on Flickr30K dataset.

Models	Text Retrieval			Image Retrieval			*R*@*sum*
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	
VSE++ [6]	52.9	80.5	87.2	39.6	70.1	79.5	409.8
SCAN [11]	67.4	90.3	95.8	48.6	77.7	85.2	465
VSRN [12]	70.4	89.2	93.7	53.0	77.9	85.7	469.9
IMRAM [2]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
TERN [17]	53.2	79.4	86.9	41.1	71.9	81.2	413.7
MMCA [28]	74.2	92.8	96.4	54.8	81.4	87.8	487.4
CAMERA [23]	76.5	95.1	97.2	58.9	84.7	90.2	502.6
TERAN [16]	75.8	93.2	96.7	59.5	84.9	90.6	500.7
GASA [18]	74.9	92.7	96.8	55.3	82.5	89.3	491.5
SGRAF [4]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
SSM(Ours)	**80.3**	**94.9**	**98.2**	**62.3**	**85.9**	**91.4**	**513.9**

Table 2. The experimental results on MS-COCO 1K dataset.

Models	Text Retrieval			Image Retrieval			*R*@*sum*
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	
VSE++ [6]	64.6	90.0	95.7	52.0	84.3	92.0	478.6
SCAN [11]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
VSRN [12]	76.2	94.8	98.2	62.8	89.7	95.2	516.9
IMRAM [2]	76.7	95.6	98.5	61.7	89.1	95.0	516.6
TERN [17]	63.7	90.5	96.2	51.9	85.6	93.7	481.6
MMCA [28]	74.8	95.6	97.7	61.6	89.8	95.2	514.7
CAMERA [23]	75.9	95.5	98.5	62.3	90.1	95.2	517.5
TERAN [16]	77.7	95.9	98.5	65.0	91.2	96.4	524.7
GASA [18]	77.9	96.5	98.8	63.4	90.7	96.0	523.3
SGRAF [4]	79.6	96.2	98.5	63.2	90.7	96.1	524.3
SSM(Ours)	**82.2**	**97.7**	**99.4**	**68.2**	**92.6**	**97.2**	**537.3**

R@5 and R@10. Compared to the traditional interaction method IMRAM, SSM improves retrieval speed while maintaining higher accuracy without the complex interactions. Compared with CAMERA, which also uses the BERT pre-trained model, SSM achieves a 3.4% improvement in *R*@1 for text retrieval and an even greater improvement (3.8%) for image retrieval. The SSM model also has better evaluation metrics than the Transformer-based fine-grained model TERAN [16].

Table 3. The experimental results on MS-COCO 5K dataset.

Models	Text Retrieval			Image Retrieval			*R*@*sum*
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	
VSE++ [6]	41.3	71.1	81.2	30.3	59.4	72.4	355.7
SCAN [11]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN [12]	50.3	79.6	87.9	37.9	68.5	79.4	403.6
IMRAM [2]	53.6	83.2	91.0	39.7	69.1	79.8	416.4
TERN [17]	38.4	69.5	81.3	28.7	59.7	72.7	350.3
MMCA [28]	54.0	82.5	90.7	38.7	69.7	80.8	416.4
CAMERA [23]	53.1	81.3	89.8	39.0	70.5	81.5	415.2
TERAN [16]	55.6	83.9	91.6	42.6	72.5	82.9	429.1
GASA [18]	56.7	84.8	91.8	42.3	71.2	83.1	429.9
SGRAF [4]	57.8	-	91.6	41.9	-	81.3	-
SSM(Ours)	**60.1**	**86.3**	**92.7**	**45.5**	**75.7**	**85.0**	**445.3**

Table 2, Table 3 show the bidirectional retrieval results on MS-COCO dataset with 1K and 5K test images. The results show that *R*@1 is 68.2% for image retrieval and *R*@1 is 82.2% for text retrieval on MS-COCO 1K. For MS-COCO 5K, our proposed SSM model still has a performance advantage over other models. It demonstrates that the SSM model has great generalization and robustness. Meanwhile, the performance achieved by SSM on *R*@1 verifies that the proposed methods in this paper can effectively enhance the ability of encoder.

4.5 Ablation Studies

To demonstrate the effectiveness and stability of each component in the SSM model, we carry a series of ablation experiments on the Flickr30K dataset in this section. The baseline model for comparison utilizes BERT as the text encoder and Transformer as the image encoder, and uses only the normal local alignment method during the similarity calculation. Table 4 investigates the impact of each component, where GAU denotes gated attention units, AWS denotes the adaptive weighted similarity calculation method, MVA denotes the multi-view alignment method, and w/o denotes that the current component is not used. For example, w/o AWS denotes that the adaptive weighted similarity calculation is replaced with the mainstream adopted Max-Sum fusion method, while the multi-view alignment and gated attention units are retained.

Table 4. Ablation study on Flickr30K to investigate contributions of each component.

Models	Text Retrieval			Image Retrieval			*R*@*sum*
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	
Baseline	75.0	92.3	95.7	59.1	83.6	90.1	495.8
w/o GAU	78.7	94.5	97.5	61.4	85.3	91.1	508.5
w/o AWS	77.2	93.8	96.5	60.7	84.6	90.6	503.4
w/o MVA	78.4	94.2	96.9	61.1	84.8	91.0	506.4
SSM(Ours)	**80.3**	**94.9**	**98.2**	**62.3**	**85.9**	**91.4**	**513.9**

In Table 4 we can observe that each strategy brings an improvement on the baseline model. GAU improves 1.6% for text retrieval and 0.9% for image retrieval on R@1, demonstrating that gated attention units can filter redundant information and bring positive profits. AWS achieves a more comprehensive improvement in all metrics. It indicates that the adaptive weighted similarity calculation method, compared to the common local alignment method, is able to highlight the role played by important information in the alignment process. The results of whether or not to use MVA demonstrate that the different view information can be complementary. Finally, the final SSM model using all strategy achieves optimal result.

Table 5. Model performance with different fusion methods on Flickr30K.

Methods	Text Retrieval			Image Retrieval			*R*@*sum*
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	
Mean-Mean	66.6	90.0	94.5	54.2	80.8	88.1	474.2
Max-Max	73.9	93.1	96.6	56.0	82.4	89.3	491.3
Max-Mean	71.7	92.5	96.5	56.6	82.3	89.3	488.9
Max-Sum	75.0	92.3	95.7	59.1	83.6	90.1	495.8
AWS(Ours)	**77.3**	**94.1**	**97.0**	**60.2**	**84.1**	**90.6**	**503.3**

Table 5 explores the effectiveness of the proposed AWS compared to the conventional Max-Mean and other variants on the baseline model without any strategy. We can observe the Mean-Mean fusion strategy is less effective, and the Max-Mean, Max-Sum and Max-Max achieve better retrieval accuracy. This may be because the Mean-Mean strategy considers the value of each feature completely equally, leading to the interference of some unnecessary features. The best performance is achieved by the AWS strategy, which indicates that our methods is able to dynamically consider the importance of different features compared to the Mean strategy. Meanwhile, compared to the Max-Max strategy which only considers the features with the highest similarity scores, our strategy is able to better utilize the information of each feature.

Table 6. Ablation study on Flickr30K to investigate contributions of multi-view alignment.

Methods				Text Retrieval			Image Retrieval			$R@sum$
sl	fl	sg	fg	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
✓				78.4	94.2	97.5	61.1	84.8	91.0	506.4
✓	✓			80.1	94.6	97.9	61.6	85.4	91.2	510.8
✓	✓	✓		79.9	94.7	98.0	61.9	85.8	**91.5**	511.8
✓	✓	✓	✓	**80.3**	**94.9**	**98.2**	**62.3**	**85.9**	91.4	**513.9**

Table 6 explores the impact of different view alignment on the Flickr30K dataset, where sl denotes semantic-level local alignment, fl denotes feature-level local alignment, sg denotes semantic-level global alignment, and fg denotes feature-level global alignment. The experimental results verify that MVA can improve the retrieval accuracy of the model.

5 Conclusion

In this paper, we present a Transformer-based image-text retrieval model SSM based on semantic selection and multi-view alignment. SSM utilizes gated attention units and the adaptive weighted similarity calculation method for semantic selection and performs cross-modal alignment in multiple views. The experimental results on MS-COCO dataset and Flickr30K dataset show that SSM has excellent cross-modal retrieval performance, and the ablation experiments also demonstrate the effectiveness of each component. Our next work will explore the effectiveness in our methods on multimodal pre-trained models and investigate how to distill the knowledge from the interaction-based model to alignment-based models to achieve an overall improvement in accuracy and speed.

Acknowledgments. This work is supported by the National Key R&D Program of China under Grant No. 2021YFC3300300; the National Natural Science Foundation of China under Grant No. 62032013, 62132004.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
2. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12655–12663 (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

4. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1218–1226 (2021)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
7. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 214–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_13
8. Hao, Y., Dong, L., Wei, F., Xu, K.: Visualizing and understanding the effectiveness of BERT. arXiv preprint arXiv:1908.05620 (2019)
9. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
10. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
11. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 212–228. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_13
12. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4654–4662 (2019)
13. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
15. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 638–647 (2022)
16. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **17**(4), 1–23 (2021)
17. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5222–5229. IEEE (2021)
18. Miao, L., Lei, Y., Zeng, P., Li, X., Song, J.: Granularity-aware and semantic aggregation based image-text retrieval network. Comput. Sci. **49**(11), 134–140 (2022)
19. Min, S., et al.: Hunyuan_tvr for text-video retrivial. arXiv preprint arXiv:2204.03382 (2022)
20. Peng, Y., Huang, X., Zhao, Y.: An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges. IEEE Trans. Circuits Syst. Video Technol. **28**(9), 2372–2385 (2017)
21. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer

image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
22. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. arXiv preprint arXiv:1904.07531 (2019)
23. Qu, L., Liu, M., Cao, D., Nie, L., Tian, Q.: Context-aware multi-view summarization network for image-text matching. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1047–1055 (2020)
24. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
25. Shukor, M., Couairon, G., Grechka, A., Cord, M.: Transformer decoders with multimodal regularization for cross-modal food retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4567–4578 (2022)
26. Su, W., et al.: VL-BERT: pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
28. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10941–10950 (2020)