Multi-task Visual Semantic Embedding Network for image-text retrieval

Xueyang Qin
Dalian University of Technology
qinxueyang@snnu.edu.cn

Lishuang Li*
Dalian University of Technology
lils@dlut.edu.cn

Jingyao Tang
Dalian University of Technology
tangjingyao@mail.dlut.edu.cn

Fei Hao
Shaanxi Normal University
fhao@snnu.edu.cn

Meiling Ge
Weifang University
gemeiling@wfu.edu.cn

Guangyao Pang
Wuzhou University
pangguangyao@gmail.com

Abstract

Image-text retrieval aims to capture the semantic correspondence between images and texts, which serves as a foundation and crucial component in multi-modal recommendations, search systems, and online shopping. Existing mainstream methods primarily focus on modeling the association of image-text pairs while neglecting the advantageous impact of multi-task learning on image-text retrieval. To this end, a Multi-task Visual Semantic Embedding Network (MVSEN) is proposed for image-text retrieval. Specifically, we design two auxiliary tasks, including text-text matching and multi-label classification, for semantic constraints to improve the generalization and robustness of visual semantic embedding from a training perspective. Besides, we also present an intra- and inter-modality interaction scheme to learn discriminative visual and textual feature representations by facilitating information flow within and between modalities. Subsequently, we utilize multi-layer graph convolutional networks in a cascading manner to infer the correlation of image-text pairs. Experimental results show that MVSEN outperforms state-of-the-art methods on two publicly available datasets, Flickr30K and MSCOCO, with *rSum* improvements of 8.2% and 3.0%, respectively.

Keywords: Image-text Retrieval, Cross-modal Retrieval, Multi-task Learning, Graph Convolutional Networks

1. Introduction

Vision and language are fundamental information patterns for people to understand the world. The interplay between what we perceive visually and how we communicate through language forms the cornerstone of our experience. Image and text, as the most direct reflections of vision and language, serve as potent conduits through which we exchange ideas, express emotions, and weave narratives. Exploring the relationship between them has become a hotspot in the field of multimodality and spawned some specific applications, such as image-text retrieval [39, 24], multimodal recommendation [19, 32], and visual commonsense reasoning [15, 16]. In this paper, we focus on image-text retrieval that aims to bridge the semantic gap between images and textual descriptions. Despite significant efforts in recent years, there remains a challenge in measuring the relevance of images and texts due to the heterogeneity and distributional differences in the data of these two modalities.

To cope with the challenge, early approaches for image-text retrieval project the whole visual and textual information together into a shared subspace, where the correlations between images and texts are easily measured. For instance, Wang et al. [27] presented a two-branch embedding network to model images and texts, respectively, while employing a similarity network with element-wise product operation followed by a fully-connected layer to compute the correlations of image-text pairs. Similar methods, such as [20, 12, 25, 23, 4, 33], also design different networks to acquire global semantic features of images and texts to mine visual-linguistic associations. Although such coarse-grained methods are impressive, they overlook the details of image-text alignment, resulting in poor performance. Intuitively, when observing an image, people tend to pay more attention to salient regions and less attention to non-salient regions. Considering this, some works begin to explore fine-grained alignments between image regions and text words to discover the connection between these two modalities. A common practice is to use pre-trained object detection tools to identify objects in an image and extract region-specific visual features using convolutional neural networks, while adopting recurrent neural networks to obtain word-level features. After that, the similarity scores between images and texts are inferred using paired region-word similarity matrices. SCAN [10] is a typical representative of such methods, which introduces a cross-modal attention network to explore the fine-grained relationship between images and texts. Subsequently, many attention-based methods [17, 30, 8, 37, 38, 31, 29] are proposed to realize fine-grained alignment. Compared with coarse-grained methods, fine-grained methods show great potential for cross-modal image-text retrieval.

However, the fine-grained methods are more about improving performance from a model design perspective and ignore the view of training optimization. Therefore, we introduce two auxiliary tasks to further enhance cross-modal retrieval performance from the model optimization perspective. Theoretically, the optimization of single-task learning moves toward loss reduction during training. If a model is trapped in a local minimum, it is difficult to be optimized further, which is not conducive to finding the global optimum. Distinctively, multi-task learning jointly trains multiple differentiated tasks. Due to the differences between tasks, the optimization directions between various tasks may be different. When the target task falls into a local optimum, it may jump out of the local optimum under the action of other tasks, which provides a possibility for finding the global optimum. Furthermore, some sub-networks or parameters are shared in multi-task learning, which helps to learn general feature representations and improve the model's robustness and generalization ability.

Based on the above analysis and discussion, a Multi-task Visual Semantic Embedding Network (MVSEN) is proposed to explore cross-modal image-text retrieval from both the model design and optimization perspectives. Firstly, we present an intra- and inter-modality interaction mechanism from the model design perspective to obtain discriminative transformed visual and textual features. Subsequently, a similarity vector function is utilized to acquire similarity matrix vectors that will be used to infer the correlations between images and texts through graph convolutional networks with residual connection followed by two fully-connected layers. Secondly, From the perspective of training optimization, we design two semantically constrained auxiliary tasks, including text-text matching and multi-label classification, to train together with the target task cross-model image-text retrieval to improve model's performance. Finally, we employ a weighted approach to combine the optimization objectives of these three tasks as the optimization loss of the proposed MVSEN. In brief, our contributions can be summarized as threefold:

- We present a multi-task visual semantic embedding network that exploits intra- and inter-modality interaction strategy to enhance the discrimination of visual and textual features, as well as utilizes multi-task collaborative training to improve the performance of cross-modal image-text retrieval.

- We introduce two auxiliary tasks into image-text retrieval, which is beneficial for improving the robustness and generalization of visual-semantic embedding. To the best of our knowledge, we are the first time to introduce the task of text-text matching as a semantic constraint into cross-modal retrieval.

- We conduct extensive experiments on two benchmarks, Flickr30K and MSCOCO, and the experimental results show that the proposed MVSEM achieves advanced performance compared to the state-of-the-art methods, with improvements of 8.2% and 3.0% on evaluation metric $rSum$, respectively.

2. Related Work

2.1. Image-text Retrieval

According to the way of image-text modeling, existing methods can be roughly divided into two categories: global-based methods [20, 25, 23, 4, 2] and local-based methods [17, 18, 3, 37, 38, 31]. The global-based methods aim to project heterogeneous multimodal data into a joint embedding space, generating corresponding global feature representations for each modality and then measuring their correlation using a distance metric function. Liu et al. [20] utilized ResNet152 and RNN to encode visual and textual information, and learned consistent visual-semantic embedding through deep mapping and reconstructed mapping. Sarafianos et al. [25] developed an adversarial network to learn discriminative feature representations in these two modalities jointly. Chen et al. [2] proposed an effective generalized pooling strategy to learn optimized visual-semantic features. Since these methods only consider the global semantic information of image-text pairs, it is not advantageous in the face of complex scenes. Compared with global-based methods, local-based methods pay more attention to the details of image-text alignment. Lee et al. [10] presented a stacked attention framework to identify the potential alignment between visual regions and textual words. Considering that different fragment-level features contribute differently to inferring the correspondence between vision and language, Liu et al. [17] proposed a bidirectional focal attention approach to focus on relevant regions and words and eliminate irrelevant ones. Wu et al. [31] adopted the method of reassigning region-word attention weights to alleviate the impact of unimportant fragment-level features on model performance. Additionally, some methods consider global- and local-based strategies to explore the correlation of image-text pairs from different perspectives. For example, Diao et al. [5] exploited graph convolutional networks to perform similarity inference on fused global-based alignment and local-based alignment. Wang et al. [28] designed a global-local alignment strategy that considers both global semantic and local segment information. Analogously, similar approaches such as [9, 14] and [11] also optimized the whole network in a joint manner.

2.2. Multi-task Learning

Different from single-task learning, multi-task learning learns a shared feature representation by jointly modeling multiple associated single tasks, where each task can act as a semantic constraint for other tasks, which is beneficial to improve the performance of models. In recent years, multi-task learning has been applied to various computer vision and natural language processing tasks, such as dense prediction [35], emotion recognition [6], and biomedical relation extraction [22] and so on. Vandenhende et al. [26] proposed a multi-scale task interaction framework to determine the information interaction between different tasks in multi-task learning through distillation units. Xu et al. [35] presented a shared encoder-decoder strategy to jointly model multi-task learning and capture task-specific features via cross-task attention mechanisms. Similarly, Moscato et al. [22] also designed shared encoder layers, including lexicon encoder and transformer encoder, and task-specific layers to realize biomedical relation extraction. Foggia et al. [6] developed a convolution-based shared encoding layer and a task-specific layer, while employing an independent classification layer to make predictions for different tasks. Similarly, there are also some works [12, 21, 36, 34, 13] that adopt multi-task learning in the field of cross-modal image-text retrieval, but the difference is that cross-modal retrieval is the target task, and other tasks are auxiliary tasks. For example, Luo et al. [21] presented correlation recognition and context reconstruction tasks combined with two regularization terms to jointly improve the performance of cross-modal image-text retrieval. Xu et al. [34] integrated image and text multi-classification tasks into cross-modal retrieval to constrain the global semantic features. Li et al [12, 13] exploited the image captioning task to model the ground-truth caption and the generated caption association through the log-likelihood function. Analogously, Yuan et al. [36] also adopted the joint training of image captioning and image-text retrieval to improve the performance of cross-modal retrieval.

2.3. Differences with existing methods

The proposed MVSEN is a multi-task learning-based method. However, it differs from previous methods as follows: (1) Compared with existing multi-task learning approaches, such as [12, 21, 36, 34, 13], we introduce a new auxiliary task text-text matching into cross-modal retrieval. Theoretically, if an image has a high correlation with a text, then other texts that are highly related to the text will be related to the image. Considering this, we employ a text-text matching task to constrain the global semantic information to improve the performance of cross-modal image-text retrieval. To the best of our knowledge, we are the first time to introduce the text-text matching task into image-text retrieval. (2) Although CASC [34] also employs a multi-label classification task, it treats images and texts as separate multi-label classifications. In contrast, our approach performs multi-label classification on the fused visual and textual information. Furthermore, we consider the interaction within each modality to preserve shared information between images and texts before the multi-label classification, while CASC ignores this. Based on the above analysis, it can be observed that our method MVSEN is different from existing cross-modal retrieval approaches.

3. Methodology

In this section, we will present the proposed Multi-task Visual Semantic Embedding Network (MVSEN) in detail. As illustrated in Figure 1, the overall framework of MVSEN mainly consists of two components: feature representation for encoding shared visual and textual features, and multi-task learning for jointly learning and optimizing various tasks. Specifically, we first introduce the feature representation module in Section 3.1. Then, we explain the multi-task learning in Section 3.2. Finally, we introduce the optimization objective of MVSEN in Section 3.3.

3.1. Feature Representation

Visual Feature Representation. Given an image I, we employ $v = \{v_1, v_2, ..., v_i, ..., v_k\}$ to represent the features of image I, where $v_i \in R^{1\times d}$ denotes the i-th regional feature and k indicates the number of salient region. Concretely, a well-known object detection tool Faster-RCNN pre-trained by Anderson et al. [1] on Visual Genome, is used to detect visual regions. Then, we select top-k ($k = 36$) salient regions with the highest confidence scores and utilize ResNet [7] to extract the feature of each region. This process can be formalized as

$$v = f_{rn}(f_{fr}(I, \theta_{fr}), \theta_{rn}), \tag{1}$$

where $f_{rn}(\cdot)$ and $f_{fr}(\cdot)$ denote the Faster-RCNN and ResNet functions, respectively, and θ_{rn} and θ_{fr} are the corresponding learnable parameters.

Textual Feature Representation. Given a text T containing n words, since each word in the text is not isolated, we utilize a Bidirectional Gated Recurrent Unit (Bi-GRU) to model the sequential information of the text. The Bi-GRU consists of a forward GRU to capture semantic features in the forward direction and a backward GRU to capture semantic features in the reverse direction. Therefore, for the given text $T = \{w_1, w_2, ..., w_j, ..., w_n\}$, the

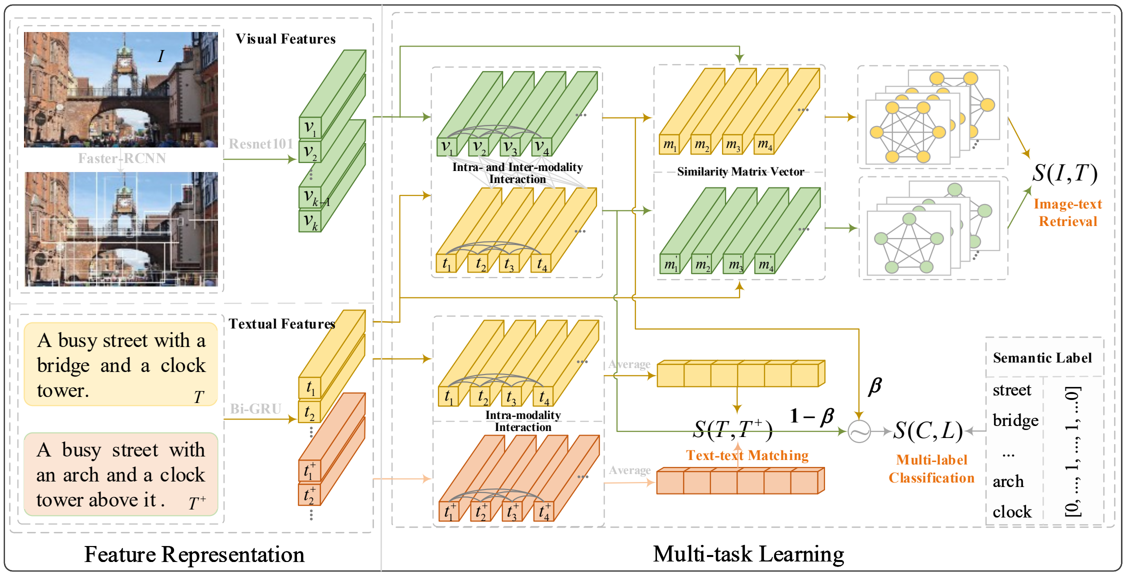


Figure 1: The overall framework of the proposed MVSEN.

encoding using Bi-GRU can be expressed as

$$\begin{aligned} \overrightarrow{h}_j &= f_{\overrightarrow{g}}(\overrightarrow{h}_{j+1}, w_j, \theta_{\overrightarrow{g}ru}), \\ \overleftarrow{h}_j &= f_{\overleftarrow{g}}(\overleftarrow{h}_{j+1}, w_j, \theta_{\overleftarrow{g}}), \end{aligned} \tag{2}$$

where $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ indicate the j-th hidden state representations from the forward $f_{\overrightarrow{g}}(\cdot)$ and backward $f_{\overleftarrow{g}}(\cdot)$, respectively. $\theta_{\overrightarrow{g}ru}$ and $\theta_{\overleftarrow{g}}$ are the learnable parameters. To better exploit the encoding information of Bi-GRU, we adopt the average of $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ as the feature representation t_j of the word w_j, *i.e.*, $t_j = (\overrightarrow{h}_j + \overleftarrow{h}_j)/2 \in R^{1\times d}$. Subsequently, the features of text T can be written as $t = \{t_1, t_2, ..., t_j, ...t_n\}$. Similarly, the features of the positive sample T^+ of text T can be denoted as $t^+ = \{t_1^+, t_2^+, ...t_j^+, ..., t_m^+\}$.

3.2. Multi-task Learning

In this subsection, we will illustrate the proposed multi-task learning module, including image-text retrieval, text-text matching, and multi-label classification tasks. We will first discuss the intra- and inter-modality interaction module since it is adopted in image-text retrieval and multi-label classification.

Intra-modality Interaction. Intuitively, when describing an image, people will focus on some salient regions according to their preferences and enrich the image with words that have no real meaning. As shown in Figure 1, some contents such as "*buildings*" and "*people*" in the image cannot find the corresponding semantics in the sentence "*A busy street with a bridge and a clock tower*". Also, the words "*a*" and "*with*" in the sentence have no real meaning. Obviously, these will affect the performance of image text retrieval. Therefore, we adopt the self-attention mechanism to strengthen meaningful words and weaken meaningless words. At the same time, we integrate self-learned "*external information*" into the self-attention mechanism to alleviate the inconsistency between visual and textual information.

Specifically, taking textual feature t as an example, we first map the textual feature t into the query matrix $Q = tW_q \in R^{n\times d_v}$, key matrix $K = tW_k \in R^{n\times d_v}$, and value matrix $V = tW_v \in R^{n\times d_v}$, where W_q, W_k and W_v are learnable parameters. Subsequently, we employ Gaussian distribution to initialize two matrices $K_m \in R^{l\times d_v}$ and $V_m \in R^{l\times d_v}$ as self-learned "*external information*" to extend K and V. A weighted output of text feature t can be formalized as

$$h_i = f_{sm}(\frac{Q f_{cc}(K, K_m)^T}{\sqrt{d}}) f_{cc}(V, V_m), \quad (3)$$

where $f_{sm}(\cdot)$ denotes the *softmax* function. $f_{cc}(\cdot)$ means a function that performs a concatenation operation on the first dimension, such as $f_{cc}(K, K_m) \in R^{(n+l)\times d_v}$. Finally, we adopt H ($d = H \times d_v$) substructures in Eq. 3 to ensure the diversity and robustness of the encoded information, *i.e.*, $t^{'} = concat(h_1, h_2, ..., h_i, ..., h_H) \in R^{n\times d}$. For simplicity, we utilize the function $f^T_{intra}(\cdot)$ to represent the intra-modality interaction within the text T, and the entire above process can be rewritten as

$$t^{'} = f^T_{intra}(t, \theta^T_{intra}), \quad (4)$$

where θ^T_{intra} is learnable parameter. Analogously, we can obtain visual feature $v^{'} = f^I_{intra}(v, \theta^I_{intra}) \in R^{k\times d}$ after performing the intra-modality interaction.

Inter-modality Interaction. Unlike the intra-modality interaction, the inter-modality interaction primarily focuses on the information correlation between different modalities, which is crucial for enhancing the performance of cross-modal retrieval. As we can see from Figure 1, some visual regions in the image can be matched with corresponding textual fragments and vice versa. To quantify this, we adopt the dot product followed by the "*softmax*" function to calculate the correlation scores between different visual regions and textual words. The process can be formalized as

$$a_{t2i} = f_{sm}(\alpha \cdot f_{lr}(v^{'}(t^{'})^T)), \quad (5)$$

where $a_{t2i} \in R^{k\times n}$ denotes the correlation score matrix between visual regions and textual words. α is a control factor used to regulate the correlation scores. $f_{lr}(\cdot)$ indicates the "*LeakyRelu*" activation function.

After that, we can consider this question: can textual information be converted into visual information? Actually, the j-th column of the score matrix a_{t2i} can be interpreted as the importance of word w_j to each visual region. Therefore, an approximate approach is to obtain the transformed features through matrix multiplication. Formally, the visual feature v_{t2i} transformed from textual feature $t^{'}$ can be denoted as $v_{t2i} = a_{t2i}t^{'} \in R^{k\times d}$. For convenience, we employ the function f^{t2i}_{inter} to represent the inter-modality interaction from text to image direction and the above process can be formalized as

$$v_{t2i} = f^{t2i}_{inter}(v^{'}, t^{'}, \alpha). \quad (6)$$

Similarly, we can acquire the textual feature t_{i2t} transformed from visual feature $v^{'}$, *i.e.*, $t_{i2t} = f^{i2t}_{inter}(t^{'}, v^{'}, \alpha) \in R^{n\times d}$. Theoretically, the transformed features retain some common information in both modalities, preventing the model from focusing on non-important information and helping improve the model's performance.

Image-text Retrieval. As discussed earlier, the transformed features are beneficial to maintaining common information between these two modalities. Considering this, we utilize the transformed and original features to explore the correlations between images and texts. Specifically, taking the features v_{t2i} and v as an example, we first calculate the similarity matrix vector $M_{t2i} = \{m_1, m_2, ..., m_k\} \in R^{k\times d_c}$ between v_{t2i} and v by

$$M_{t2i} = \frac{|v_{t2i} - v|^2 w}{||v_{t2i} - v|^2 w||_2}, \quad (7)$$

where $|\cdot|^2$ denotes element-wise square. $w \in R^{d\times d_c}$ is learnable parameter matrix, and $||\cdot||_2$ is L_2-norm.

Subsequently, we treat M_{t2i} as a fully connected graph G_{t2i} containing k nodes, where each row in M_{t2i} contributes to a node. To enhance the representation of these nodes, we utilize stacked graph convolutional networks (GCN) followed by residual connections to establish the association between them. We implement the update of node features of the l-th ($l \geq 0$) layer by

$$M^{l+1}_{t2i} = f_{gcn}(M^l_{t2i}, w^l_{t2i}, \theta^l_{t2i}) + M^l_{t2i}, \quad (8)$$

$$w^l_{t2i} = (M^l_{t2i} w^l_1 + b^l_1)(M^l_{t2i} w^l_2 + b^l_2)^T, \quad (9)$$

$$M^0_{t2i} = M_{t2i}, \quad (10)$$

where $M^l_{t2i} \in R^{k\times d_c}$ denotes the node features of the l-th layer, and $w^l_{t2i} \in R^{k\times k}$ is corresponding edge weight matrix. $w_1^l \in R^{d_c \times d_c}$, $w_2^l \in R^{d_c \times d_c}$, $b_1^l \in R^{d_c \times 1}$ and $b_2^l \in R^{d_c \times 1}$ are learnable parameters. Then, two cascaded fully connected layers are employed to infer semantic relevance score S_{t2i} of image-text pairs.

$$S_{t2i} = f_m((f_{th}(M_{t2i}^l w_1 + b_1))w_2 + b_2), \quad (11)$$

where $f_m(x)$ means averaging all elements in the matrix x. $f_{th}(\cdot)$ indicates the "*tanh*" activation function. $w_1 \in R^{d_c \times d_v}$, $w_2 \in R^{d_v \times 1}$, $b_1 \in R^{d_c \times 1}$ and $b_2 \in R^{d_v \times 1}$ are learnable parameters. Likewise, we can get S_{i2t}.

To optimize image-text retrieval, following previous approaches [3, 5, 10], we employ the bidirectional triplet loss as the optimization objective, that is

$$\begin{aligned} \mathcal{L}_r = \sum_{(I,T)} [\lambda_1 - S(I,T) + S(I,T^-)]_+ \\ +[\lambda_1 - S(I,T) + S(I^-,T)]_+, \end{aligned} \quad (12)$$

where λ_1 is a margin. I and T are matched image-text pair, and I^- and T^- are corresponding negatives, respectively. For each batch size B, the top $C/B + 1$ most relevant examples are selected as negatives, where C indicates the number of $S(I,T) + \lambda_1 - f_{diag}(S(I,T)) > 0$, $S(I,T) \in \{S_{t2i}, S_{i2t}\}$.

Text-text Matching. Different from image-text retrieval, we view the text-text matching task as an auxiliary task to perform semantic constraint from a global perspective. As shown in Figure 1, the texts T and T^+ are the ground-truth captions of the image I. Theoretically, if the text T and image I exhibit a high degree of semantic similarity, and text T^+ and image I also show a solid semantic correspondence, then it can be inferred that there exists a high semantic correlation between text T and text T^+. Based on the above analysis, we can improve the performance of image-text retrieval by constraining the semantic consistency of these two texts to alleviate the semantic bias that may occur during the model's training process.

Specifically, we first utilize the intra-modality interaction module to capture textual semantics information. It is worth noting that the intra-modality interaction module here differs from Eq. 4. In Eq. 4, the matrices K and V are extended, but this operation is not performed here, constituting the sole distinction between them. To avoid duplication and distinguish them, we employ the function $g_{intra}^T(\cdot)$ to represent the intra-modal interaction here. Taking text T as an example, the feature $t_{T,intra}$ of text T after intra-modality interaction can be written as

$$t_{T,intra} = g_{intra}^T(t, \theta_{T,intra}), \quad (13)$$

where $t_{T,intra} = \{t_{1,intra}, t_{2,intra}, ..., t_{n,intra}\}$ and $\theta_{T,intra}$ is learnable parameter. Then, we obtain the global semantic feature t_g of text T through the mean operation, *i.e.*,

$$t_g = \frac{1}{n} \sum_{j=1}^{n} t_{j,intra} \quad (14)$$

Analogously, the global semantic feature of text T^+ can be denoted as t_g^+. Subsequently, we exploit cosine similarity to measure the correlation score between t_g and t_g^+, that is, $S(T,T^+) = cos(t_g, t_g^+)$. Similar to image-text retrieval, we also employ bidirectional triplet loss as the optimization objective of text-text matching, *i.e.*,

$$\begin{aligned} \mathcal{L}_m = \sum_{(T,T^+)} [\lambda_2 - S(T,T^+) + S(T,\widehat{T}^+)]_+ \\ +[\lambda_2 - S(T,T^+) + S(\widehat{T},T^+)]_+, \end{aligned} \quad (15)$$

where λ_2 is a margin. $\widehat{T}$ and $\widehat{T}^+$ are text negatives of T and T^+, respectively, obtained in the same manner as described in Eq. 12 for negatives.

Multi-label Classification. Similar to the text-text matching task, the multi-label classification is also considered an auxiliary task for implementing global semantic constraint. Intuitively, each image or text will display some semantic information. For example, the image and texts in Figure 1 contain significant semantic features such as "*street*" and "*clock tower*". By constraining the shared semantic information between image and text, it can be ensured that the matched images and texts always maintain semantic consistency during the model training process. Since existing image-text retrieval datasets such as Flickr30K and MSCOCO do not contain semantic label information, a straightforward approach is to build semantic labels directly from these datasets. Given that nouns can embody the semantic information of images better than other types of words, we adopt nouns to construct semantic labels for image-text pairs.

Concretely, we first employ NLTK[1] to identify nouns in the text and then form the semantic label vector $L \in R^{1\times N}$ by selecting the top-N nouns with the highest frequencies. Then, a "0-1" encoded semantic label $y \in R^{1\times N}$ is assigned to each matched image-text pair (I, T), where "1" indicates that the semantic information at the same position in the label dictionary L is present in the image-text pair. Subsequently, we adopt the transformed features $v_{t2i} = \{t^{'}_{1,v}, t^{'}_{2,v}, ..., t^{'}_{k,v}\}$ and $t_{i2t} = \{v^{'}_{1,t}, v^{'}_{2,t}, ..., v^{'}_{n,t}\}$ to perform the multi-classification task. Formally, a mean function and two fully connected layers are utilized to predict the semantic labels $y^{'} \in R^{1\times N}$. The process can be formalized as

$$y^{'} = (((1-\beta)\cdot\tfrac{1}{k}\sum_{r=1}^{k} t^{'}_{r,v} + \beta\cdot\tfrac{1}{n}\sum_{s=1}^{n} v^{'}_{s,t})w_{11} \\ +b_{11})w_{22} + b_{22}, \tag{16}$$

where β is a balancing factor. $w_{11} \in R^{d\times N}$, $w_{22} \in R^{N\times N}$, $b_{11} \in R^{d\times 1}$ and $b_{22} \in R^{N\times 1}$ are learnable parameters. Then, for the matched image-text pair (I, T), the predicted label $y^{'} = \{y^{'}_1, y^{'}_2, ..., y^{'}_u, ..., y^{'}_N\}$ and ground-truth label $y = \{y_1, y_2, ..., y_u, ..., y_N\}$ can be viewed as N binary classification problems, and the optimization objective can be written as

$$\mathcal{L}_c = -\sum_{u=1}^{N} y_u \log\sigma(y^{'}_u) + (1-y_u)\log\sigma(1-y^{'}_u), \tag{17}$$

where σ indicates the *Sigmoid* activation function.

3.3. Optimization Objective

As discussed in the section "Introduction", multi-task learning is beneficial to enhance the generalization and robustness of the model. Therefore, we adopt a joint approach to optimize these three tasks, consisting of image-text retrieval, text-text matching and multi-label classification, and the optimization objective is defined as

Algorithm 1: Model training process of the proposed MVSEN.

```
Input: Matched image-text pairs $(I, T)$;
       Positive sample $T^+$ of $T$; Parameters: $\alpha$,
       $\lambda_1$, $B$, $\lambda_2$, $\beta$, $\delta$, $\gamma$; Optimized network
       parameters: $\theta$.
Output: $\theta$.
for epoch $= 1, 2, ..., E$ do
    for each batch size $B$ do
        Initial feature representations: $v$, $t$, $t^+$
          via Eqs. 1 and 2;
        Perform intra- and inter-modality
          feature interaction via Eqs. 3 and 6;
        Compute loss $\mathcal{L}_r$ for image-text
          retrieval via Eq. 12;
        Obtain loss $\mathcal{L}_m$ for text-text matching
          via Eq. 15;
        Calculate loss $\mathcal{L}_c$ for multi-label
          classification via Eq. 17;
    end
    Get losses $\mathcal{L}_r$, $\mathcal{L}_m$ and $\mathcal{L}_c$ ;
    Compute $\mathcal{L}$ via Eq. 18;
    $\theta \leftarrow$ Backward ($\mathcal{L}$).
end
```

$$\mathcal{L} = \mathcal{L}_r + \delta\cdot(\gamma\mathcal{L}_m + (1-\gamma)\mathcal{L}_c), \tag{18}$$

where δ is a balancing factor used to balance the weights between the target task and the auxiliary task. γ is employed to balance the weight between these two auxiliary tasks.

Additionally, it should be noted that all three tasks participate in the model training process, but the auxiliary tasks are not involved during the testing stage. Algorithm 1 is the entire training process, where θ is the parameter that the proposed MVSEN needs to be trained and optimized.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the proposed approach MVSEN and all baselines on two publicly avail-

[1] https://www.nltk.org/

able datasets Flickr30K [17] and MSCOCO [30]. Flickr30K consists of 31,000 images in total and each image is associated with five matched textual descriptions. Following the split protocol in [10], we adopt 29,000 images for training, 1,000 images for validation, and 1,000 images for testing. MSCOCO includes 123,287 images, and each image is manually annotated with five sentences, where 11,3287 images are employed for training, 5,000 images for validation, and 5,000 images for testing. It should be noted that the performance of MSCOCO is evaluated by averaging 5-folds of 1K and all 5K test images.

Evaluation Metrics. Following previous methods [3, 18], we utilize *Recall at* K ($R@K, K = 1, 5, 10$) as an evaluation metric to assess the performance of cross-modal retrieval. Also, a comprehensive evaluation metric $rSum$, indicating the sum of $R@K$ in cross-modal retrieval, is also adopted to evaluate the model's performance.

Implementation Details. The proposed model MVSEN is implemented using Python 3.7.0 and PyTorch 1.7.0 frameworks and trained on an NVIDIA GeForce RTX 3090 GPU with the Adam optimizer. We set the training epoch to 30 with a learning rate of 0.0002 on Flickr30K and MSCOCO, where the learning rate is reduced by 10% after 10 epochs. The batch size is set to 64. The control factor α in Eq. 5 is 10, *i.e.*, $\alpha = 10$. The margins λ_1 in Eq. 12 and λ_2 in Eq. 15 are set to 0.2 and 0.1, respectively. The balancing factor β in Eq. 16 is set to 0.99 and δ in Eq. 18 is set to 0.01. Additionally, some detailed settings can be found in our code: `https://github.com/FlyCuteBird/MVSEN`.

Results on Flickr30K.[2] Table 1 reports the quantitative results on Flickr30K dataset. It can be observed that the proposed MVSEN exceeds all baseline models in the evaluation metric $rSum$, with a gain of 8.2% $\sim$ 26.2%. Compared with the state-of-the-art method NAAF [38], the proposed MVSEN obtains performance improvements of 0.4% and 2.3% on $R@1$ for text retrieval and image retrieval, respectively. Also, compared with other baseline models, such as SGRAF [5], MEMBER [11], CGMN [3], etc. MVSEN continues to demonstrate notable performance enhancements in terms of $R@1$, exhibiting improvements ranging from 2.5% to 6.6% for text retrieval and from 2.5% to 8.6% for image retrieval, which confirms the efficacy of MVSEN in the realm of image-text retrieval.

Table 1: Quantitative results on Flickr30K dataset.

Methods	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
SGRAF$_{2021}$ [5]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
MEMBER$_{2021}$ [11]	77.5	94.7	97.3	59.5	84.8	91.0	504.8
CGMN$^{*}_{2022}$ [3]	77.9	93.8	96.8	59.9	85.1	90.6	504.1
UARDA$_{2022}$ [37]	77.8	95.0	97.6	57.8	82.9	89.2	500.3
NAAF$^{*}_{2022}$ [38]	81.3	**95.6**	98.1	60.8	84.8	90.7	511.3
GLFN$_{2023}$ [39]	75.1	93.8	97.2	54.5	82.8	89.9	493.3
RAAN$_{2023}$ [29]	77.1	93.6	97.3	56.0	82.4	89.1	495.5
VSRN++$_{2023}$ [13]	79.2	94.6	97.5	60.6	85.6	91.4	508.9
MVSEN (ours)	**81.7**	**95.6**	**98.2**	**63.1**	**88.0**	**92.9**	**519.5**

Table 2: Quantitative results on MSCOCO dataset (1K).

Methods	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
SGRAF$_{2021}$ [5]	79.6	96.2	98.5	63.2	90.7	96.1	524.3
MEMBER$_{2021}$ [11]	78.5	**96.8**	98.5	63.7	90.7	95.6	523.8
CGMN$^{*}_{2022}$ [3]	76.8	95.4	98.3	63.8	90.7	95.7	520.7
UARDA$_{2022}$ [37]	78.6	96.5	**98.9**	63.9	90.7	96.2	524.8
NAAF$^{*}_{2022}$ [38]	79.7	96.4	98.6	63.0	89.5	95.2	522.4
GLFN$_{2023}$ [39]	78.4	96.0	98.5	62.6	89.6	95.4	520.5
RAAN$_{2023}$ [29]	76.8	96.4	98.3	61.8	89.5	95.8	518.6
VSRN++$_{2023}$ [13]	77.9	96.0	98.5	64.1	91.0	96.1	523.6
MVSEN (ours)	**80.5**	96.5	98.7	**64.6**	**91.1**	**96.4**	**527.8**

Table 3: Quantitative results on MSCOCO dataset (5K).

Methods	Text Retrieval			Image Retrieval			$rSum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
MEMBER$_{2021}$ [11]	54.5	82.3	90.1	40.9	71.0	81.8	420.6
CGMN$^{*}_{2022}$ [3]	53.4	81.3	89.6	41.2	71.9	82.4	419.8
UARDA$_{2022}$ [37]	56.2	83.8	91.3	40.6	69.5	80.9	422.3
NAAF$^{*}_{2022}$ [38]	58.1	**85.5**	**92.0**	42.1	70.7	80.8	429.2
VSRN++$_{2023}$ [13]	54.7	82.9	90.9	42.0	72.2	82.7	425.4
MVSEN (ours)	**58.7**	84.0	91.7	**42.5**	**72.0**	**82.7**	**431.6**

Results on MSCOCO. Tables 2 and 3 show the experimental performance on MSCOCO (1K) and

[2] "*" indicates that we reproduce the results by the publicly available code.

Table 4: Ablation studies on Flcikr30K and MSCOCO (1K) datasets.

Models	Settings					Flickr30K dataset							MSCOCO dataset (1K)						
						Text Retrieval			Image Retrieval			$rSum$	Text Retrieval			Image Retrieval			$rSum$
	R	M	C	$t2i$	$i2t$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
#1	✓			✓		76.9	93.5	96.9	60.6	85.7	91.6	505.2	78.0	95.8	98.2	63.0	90.5	96.0	521.5
#2	✓				✓	76.7	93.0	96.9	58.6	84.5	90.6	500.3	76.1	95.7	98.6	61.3	90.1	95.8	517.6
#3	✓			✓	✓	79.1	94.5	97.9	61.7	87.4	92.4	513.0	79.1	96.2	98.5	64.4	91.2	96.4	525.8
#4	✓	✓		✓		77.1	94.2	98.0	60.4	85.2	91.5	506.4	77.8	95.5	98.3	63.0	90.5	96.0	521.1
#5	✓	✓			✓	75.5	94.2	97.6	58.6	85.1	90.7	501.7	77.5	96.0	98.5	62.3	90.5	95.8	520.6
#6	✓	✓		✓	✓	80.2	95.5	97.9	62.2	86.9	92.6	515.3	79.7	96.3	98.6	64.3	**91.5**	**96.5**	526.9
#7	✓		✓	✓		77.2	93.2	97.1	61.4	85.8	91.8	506.5	77.8	95.3	98.4	62.1	90.7	96.1	520.4
#8	✓		✓		✓	75.4	93.5	96.9	58.9	85.2	90.8	500.7	77.1	95.7	98.5	61.4	90.2	95.9	518.8
#9	✓		✓	✓	✓	79.7	93.9	97.8	**63.6**	87.4	92.7	515.1	80.0	96.3	98.5	64.4	91.3	**96.5**	527.0
#10	✓	✓	✓	✓		79.5	94.6	97.7	61.2	86.6	91.7	511.3	77.5	96.0	98.3	62.7	90.3	95.8	520.6
#11	✓	✓	✓		✓	77.8	93.3	97.2	59.2	85.0	90.5	503.0	77.7	95.8	98.5	61.6	90.3	95.9	519.8
#12	✓	✓	✓	✓	✓	**81.7**	**95.6**	**98.2**	63.1	**88.0**	**92.9**	**519.5**	**80.5**	**96.5**	**98.7**	**64.6**	91.1	96.4	**527.8**

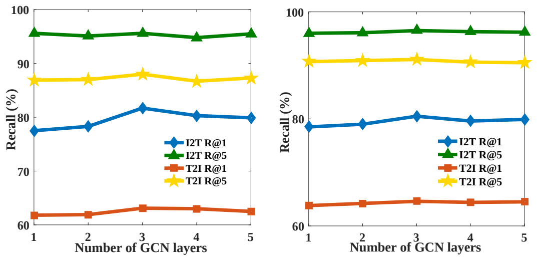
MSCOCO (5K), respectively. From Table 2, we can observe that the proposed MVSEN obtains the best results $R@1 = 80.5\%$ for text retrieval and $R@1 = 64.6\%$ for image retrieval compared with all baseline models. At the same time, MVSEN also achieves the best performance in terms of evaluation metric $rSum$, with $rSum = 527.8\%$. In addition, when the test set data increases, it can be seen from Table 3 that MVSEN still has advantages on the evaluation metric $R@1$ compared with baseline models, which indirectly shows that the proposed MVSEN can handle cross-modal image-text retrieval under different data scales well.

4.2. Effect of Different Network Modules

To assess the influence of different modules on the model's performance, we conduct ablation studies on Flickr30K and MSCOCO datasets, as shown in Table 4, where "✓" indicates that the corresponding module is adopted. R, M and C represent image-text retrieval, text-text matching and multi-classification tasks, respectively. $t2i$ and $i2t$ denote that we employ S_{t2i} and S_{i2t} to measure the semantic correlation of image-text pairs, respectively. From Table 4, we can see that the model (#12) achieves the best performance on the evaluation metric $rSum$ when all auxiliary tasks are employed, with improvements of $rSum = 6.5\%$ on Flickr30K and $rSum = 2.0\%$ on MSCOCO compared to the baseline model (#3). Additionally, when only one auxiliary task M or C is used, the model's performance is still enhanced, with a boost of 2.3% (#6 $vs.$ #3) and 2.1% (#9 $vs.$ #3) in terms of $rSum$ on Flickr30K dataset. Similarly, we can observe similar properties on MSCOCO dataset, which indicates that both text-text matching and multi-label classification tasks contribute to improving the performance of cross-modal retrieval. Moreover, their combined use yields even better results. Furthermore, it can be seen that the multi-label classification task contributes more to Flickr30K than to MSCOCO. One possible reason is that Flickr30K describes people's daily lives, and the images share many similarities. The multi-label classification task effectively constrains sentences with high similarity. In contrast, MSCOCO encompasses diverse categories, and the images exhibit more considerable variations. In this case, the role of the multi-label classification task is relatively diminished, resulting in slight performance improvement on MSCOCO dataset.

4.3. Effect of Different Parameters

In this subsection, we select two representative hyper-parameters, l and γ, to explore the impact of different parameter settings on the model's performance, where l indicates the number of GCN layers, ranging from 1 to 5, and is critical for the overall performance of image-text retrieval. The hyper-parameter γ is employed to balance the loss



(a) Results on Flickr30K (b) Results on MSCOCO

Figure 2: Effect of the number of GCN layers (l) on Flickr30k and MSCOCO (1K) datasets, where I2T and T2I indicate text retrieval and image retrieval, respectively.

weights between the target and auxiliary tasks, facilitating the training of multi-task learning. The value of γ is set from 0.1 to 0.9 with a step of 0.2.

The influence of parameter l on Flickr30K and MSCOCO is depicted in Figure 2. From these results, it can be observed that increasing the parameter l from 1 to 3 can enhance the performance of image-text retrieval on evaluation metrics $R@1$ and $R@5$. The possible reason is that increasing the parameter l explores deep feature representation, which is beneficial to improving the model's performance. However, the performance decreases when the parameter l is larger than 3, such as $l = 4$ or $l = 5$. One possible explanation is that the "over-smoothing" problem occurs when the number of GCN layers increases to a certain level, resulting in performance degradation.

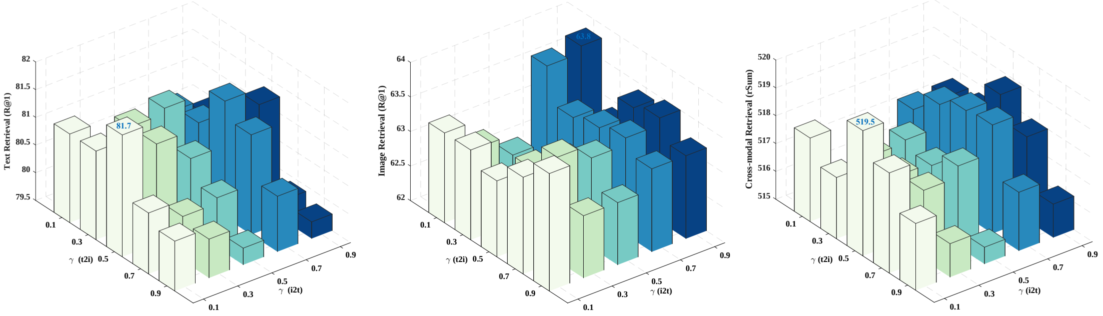
Figures 3 and 4 show the impact of parameter γ on Flickr30K and MSCOCO, respectively, where γ (o), $o \in \{t2i, i2t\}$, indicates that we employ S_o to measure the similarity of image-text pairs. Clearly, when the value of γ increases from 0.1 to 0.9 (step size is 0.2), the experimental results also change, which indirectly shows that when the weight of the auxiliary task changes, the optimization direction of the model will also be different, resulting in the performance difference. When γ(t2i)=0.5 and γ(i2t)=0.1, MVSEN obtains the best $R@1 = 81.7\%$ for text retrieval and $rSum = 519.5\%$ for cross-modal retrieval on Flickr30K. Analogously, it can be seen from Figure 4 that MVSEN achieves the best text retrieval (γ(t2i)=0.3 and γ(i2t)=0.5) and image retrieval (γ(t2i)=0.1 and γ(i2t)=0.1) by taking different values of γ.

By analyzing the impact of parameters l and γ on experimental results, we can discern that different parameter configurations substantially influence the model's performance. An appropriate parameter is helpful to improve the model's efficiency. Furthermore, it can be seen that when the model achieves the best performance, the same parameter, such as the parameter γ, may have different values on different datasets due to differences in data distribution.

4.4. Analysis of Retrieval Time

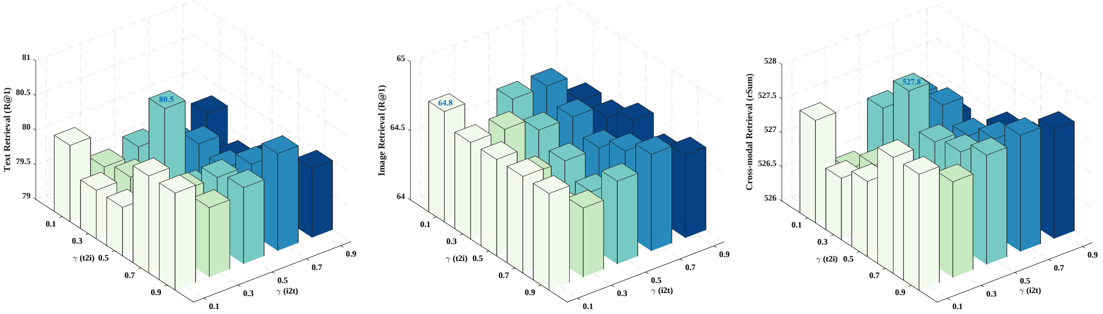
To validate that the proposed method achieves a good balance between performance and efficiency, we compare the retrieval time with three advanced approaches, including SGRAF [5], CGMN [3], and NAAF [38] based on the publicly available codes they provide. For a fair comparison, we obtain a bidirectional retrieval time by averaging the retrieval time of 5,000 image-text pairs, and all experiments are performed on an Intel(R) Core(TM) i9-10920X CPU@3.50GHz, 64G memory and an NVIDIA GeForce RTX 3090 GPU with 24G memory.

As shown in Figure 5, we can observe that when performing a bidirectional retrieval, the proposed approach MVSEN is significantly lower than SGRAF and NAAF in retrieval time but higher than CGMN. Furthermore, we can observe that the retrieval time of NAAF is the longest among these four methods. The reason is that NAAF employs the correlations between all visual regions and textual words to assess the semantic similarity of the entire image and text, resulting in a significant time overhead. In contrast, CGMN utilizes cosine to measure the similarity of image-text pairs, so the cost is minimal. Although the retrieval time of the proposed MVSEN is higher than that of CGMN, it is still within the same order of magnitude. Moreover, the performance of MVSEN on the evaluation metric $rSum$ exceeds CGMN by a large margin, with 15.3% on Flickr30K and 7.1% on MSCOCO (1K). Therefore, the proposed approach performs better in balancing performance and retrieval effi-



(a) (b) (c)

Figure 3: Effect of parameter γ on Flickr30k, and the best result is marked in the figure. (a) Text retrieval (R@1) on Flickr30k. (b) Image retrieval (R@1) on Flickr30k. (c) Cross-modal retrieval (rSum) on Flickr30k.



(a) (b) (c)

Figure 4: Effect of parameter γ on MSCOCO, and the best result is marked in the figure. (a) Text retrieval (R@1) on MSCOCO. (b) Image retrieval (R@1) on MSCOCO. (c) Cross-modal retrieval (rSum) on MSCOCO.

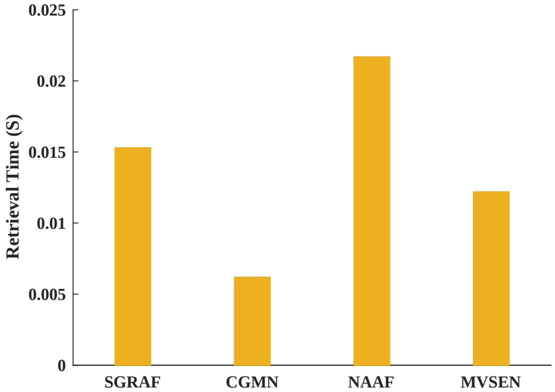


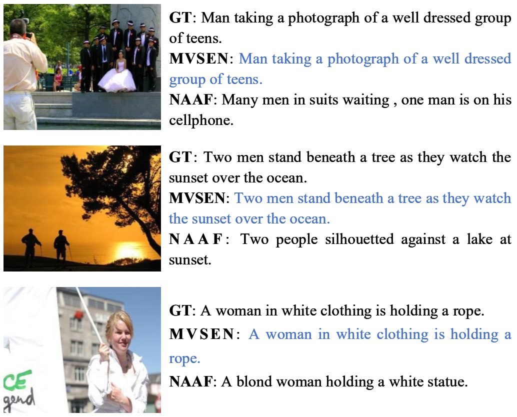
Figure 5: The comparison of bidirectional retrieval time.

ciency.

4.5. Visualization and Analysis

To further assess the performance of the proposed MVSEN, we visualize the retrieval results and compare MVSEN with the advanced method NAAF. For text retrieval, we offer the ground-truth caption (GT), and the top-ranked text retrieved by MVSEN and NAAF. Also, we report the top-2 results for image retrieval, where the correct result is marked with a green box.

Figures 6 and 7 are the results of text retrieval on Flickr30K and MSCOCO datasets, respectively. As can be seen from these retrieval results, MVSEN performs better than NAAF. Taking the third example on Flcikr30K as an illustration, although NAAF identifies some vital elements such

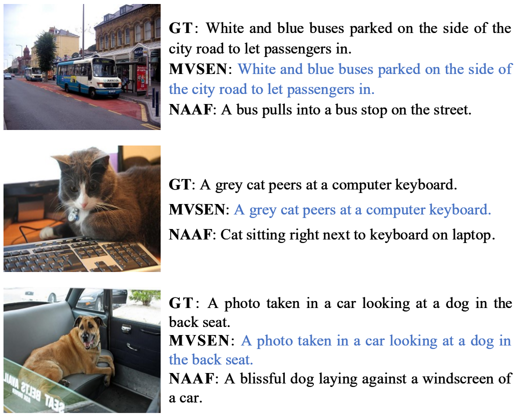


GT: Man taking a photograph of a well dressed group of teens.
MVSEN: Man taking a photograph of a well dressed group of teens.
NAAF: Many men in suits waiting , one man is on his cellphone.

GT: Two men stand beneath a tree as they watch the sunset over the ocean.
MVSEN: Two men stand beneath a tree as they watch the sunset over the ocean.
NAAF: Two people silhouetted against a lake at sunset.

GT: A woman in white clothing is holding a rope.
MVSEN: A woman in white clothing is holding a rope.
NAAF: A blond woman holding a white statue.

Figure 6: Qualitative results of text retrieval on Flickr30K.

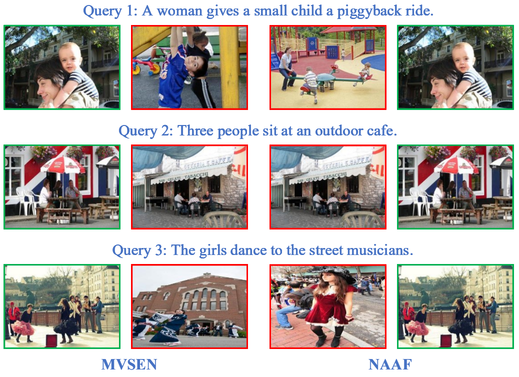


GT: White and blue buses parked on the side of the city road to let passengers in.
MVSEN: White and blue buses parked on the side of the city road to let passengers in.
NAAF: A bus pulls into a bus stop on the street.

GT: A grey cat peers at a computer keyboard.
MVSEN: A grey cat peers at a computer keyboard.
NAAF: Cat sitting right next to keyboard on laptop.

GT: A photo taken in a car looking at a dog in the back seat.
MVSEN: A photo taken in a car looking at a dog in the back seat.
NAAF: A blissful dog laying against a windscreen of a car.

Figure 7: Qualitative results of text retrieval on MSCOCO.

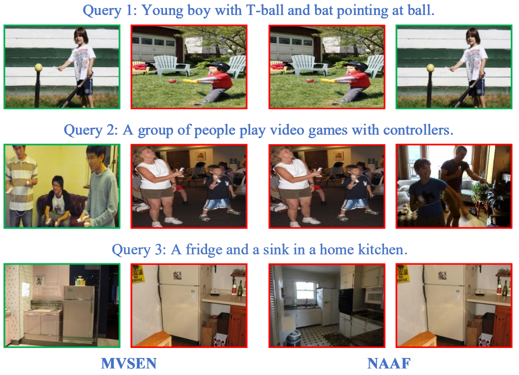


Query 1: A woman gives a small child a piggyback ride.

Query 2: Three people sit at an outdoor cafe.

Query 3: The girls dance to the street musicians.

Figure 8: Qualitative results of image retrieval on Flickr30K.



Query 1: Young boy with T-ball and bat pointing at ball.

Query 2: A group of people play video games with controllers.

Query 3: A fridge and a sink in a home kitchen.

Figure 9: Qualitative results of image retrieval on MSCOCO.

as “*blond woman*”, “*holding*” and “*white*”, it mistakes “*rope*” for “*statue*”, leading to a wrong judgment. In contrast, MVSEN can detect these small gaps, which also verifies the effectiveness of multi-task learning in text retrieval.

Figures 8 and 9 show qualitative results of image retrieval on Flickr30K and MSCOCO datasets, respectively. From these retrieval results, we can observe that for any given query text, all the images retrieved by NAAF and the proposed MVSEN exhibit high similarity. The difference is that MVSEN can distinguish them well and retrieve the ground-truth in the top-ranked result. However, NAAF is confused by these similar results, leading to “incorrect” choices. Therefore, it can be inferred from these results that the proposed MVSEN can make sound judgments when facing high-similarity images.

5. Conclusions

In this paper, we propose a Multi-task Visual Semantic Embedding Network (MVSEN) that leverages the collaboration of different tasks to explore the semantic relevance of image-text pairs. Experimental results on two publicly available benchmarks, Flickr30K and MSCOCO, show that MVSEN performs better than state-of-the-art approaches. Furthermore, through ablation experiments, it can be observed that both text-text matching and multi-label classification tasks contribute to improving the performance of cross-modal retrieval. In fact, the two auxiliary tasks designed for semantic constraints are more suitable for handling highly similar images. The effects are not as pronounced for images with significant differences, which is also a limitation of the proposed multi-task scheme. Additionally, qualitative results also confirm the effectiveness of the proposed method. In future work, we will explore the feasibility and efficacy of multi-task learning in other cross-model tasks.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 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*, pages 6077–6086, 2018. 4

[2] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021. 3

[3] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–23, 2022. 3, 7, 9, 11

[4] J. Chi and Y. Peng. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1173–1187, 2019. 2, 3

[5] H. Diao, Y. Zhang, L. Ma, and H. Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1218–1226, 2021. 3, 7, 9, 11

[6] P. Foggia, A. Greco, A. Saggese, and M. Vento. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Engineering Applications of Artificial Intelligence*, 118:105651, 2023. 3, 4

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[8] Y. He, X. Liu, Y.-M. Cheung, S.-J. Peng, J. Yi, and W. Fan. Cross-graph attention enhanced multimodal correlation learning for fine-grained image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1865–1869, 2021. 2

[9] Z. Ji, K. Chen, and H. Wang. Step-wise hierarchical alignment network for image-text matching. In *Proceedings of the 31th Intrnational Joint Conference on Artificial Intelligence*, 2021. 3

[10] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 2, 3, 7, 9

[11] J. Li, L. Liu, L. Niu, and L. Zhang. Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. *IEEE Transactions on Image Processing*, 30:9193–9207, 2021. 3, 9

[12] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662, 2019. 2, 4

[13] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):641–656, 2023. 4, 9

[14] W. Li, S. Yang, Y. Wang, D. Song, and X. Li. Multi-level similarity learning for image-text retrieval. *Information Processing & Management*, 58(1):102432, 2021. 3

[15] Z. Li, Y. Guo, K. Wang, F. Liu, L. Nie, and M. Kankanhalli. Learning to agree on vision attention for visual commonsense reasoning. *IEEE Transactions on Multimedia*, 2023. 1

[16] Z. Li, Y. Guo, K. Wang, Y. Wei, L. Nie, and M. Kankanhalli. Joint answering and explanation for visual commonsense reasoning. *IEEE Transactions on Image Processing*, 2023. 1

[17] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 3–11, 2019. 2, 3, 9

[18] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020. 3, 9

[19] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, and R. Hong. Multimodal graph contrastive learning for

multimedia-based recommendation. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 1

[20] Y. Liu, Y. Guo, L. Liu, E. M. Bakker, and M. S. Lew. Cyclematch: A cycle-consistent embedding network for image-text matching. *Pattern Recognition*, 93:365–379, 2019. 2, 3

[21] J. Luo, Y. Shen, X. Ao, Z. Zhao, and M. Yang. Cross-modal image-text retrieval with multitask learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2309–2312, 2019. 4

[22] V. Moscato, G. Napolano, M. Postiglione, and G. Sperlì. Multi-task learning for few-shot biomedical relation extraction. *Artificial Intelligence Review*, pages 1–21, 2023. 3

[23] Y. Peng and J. Qi. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(1):1–24, 2019. 2, 3

[24] X. Qin, L. Li, F. Hao, G. Pang, and Z. Wang. Cross-modal information balance-aware reasoning network for image-text retrieval. *Engineering Applications of Artificial Intelligence*, 120:105923, 2023. 1

[25] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824, 2019. 2, 3

[26] S. Vandenhende, S. Georgoulis, and L. Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*, pages 527–543. Springer, 2020. 3

[27] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 2

[28] X. Wang, L. Zhu, and Y. Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 3

[29] Y. Wang, Y. Su, W. Li, Z. Sun, Z. Wei, J. Nie, X. Li, and A. Liu. Rare-aware attention network for image–text matching. *Information Processing & Management*, 60(3):103280, 2023. 2, 9

[30] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10941–10950, 2020. 2, 9

[31] J. Wu, C. Wu, J. Lu, L. Wang, and X. Cui. Region reinforcement network with topic constraint for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):388–397, 2022. 2, 3

[32] Y. Wu, L. Liao, G. Zhang, W. Lei, G. Zhao, X. Qian, and T.-S. Chua. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*, 25:3113–3124, 2023. 1

[33] Y. Xie, X. Zeng, T. Wang, L. Xu, and D. Wang. Multiple deep neural networks with multiple labels for cross-modal hashing retrieval. *Engineering Applications of Artificial Intelligence*, 114:105090, 2022. 2

[34] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020. 4

[35] Y. Xu, X. Li, H. Yuan, Y. Yang, and L. Zhang. Multi-task learning with multi-query transformer for dense prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[36] H. Yuan, Y. Huang, D. Zhang, Z. Chen, W. Cheng, and L. Wang. Vsr++: Improving visual semantic reasoning for fine-grained image-text matching. In *Proceedings of the 25th International Conference on Pattern Recognition*, pages 3728–3735, 2021. 4

[37] K. Zhang, Z. Mao, A. Liu, and Y. Zhang. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Transactions on Multimedia*, 25:1320–1332, 2022. 2, 3, 9

[38] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2022. 2, 3, 9, 11

[39] G. Zhao, C. Zhang, H. Shang, Y. Wang, L. Zhu, and X. Qian. Generative label fused network for image-text matching. *Knowledge-Based Systems*, page 110280, 2023. 1, 9