Deep Top-k Ranking for Image-Sentence Matching

Lingling Zhang, Minnan Luo*, Jun Liu, Xiaojun Chang, Yi Yang, and Alexander G. Hauptmann

***Abstract*—Image-sentence matching is a challenging task for the heterogeneity-gap between different modalities. Ranking-based methods have achieved excellent performance on this task in the past decades. Given an image query, these methods typically assume that the correct matched image-sentence pair must rank before all the other mismatched ones. However, this assumption may be too strict and prone to the overfitting problem, especially when some sentences in a massive database are similar and confusable with one another. In this paper, we relax the traditional ranking loss and propose a novel deep multi-modal network with a top-k ranking loss to mitigate the data ambiguity problem. With this strategy, query results will not be penalized unless the index of ground truth is outside the range of top-k query results. Considering the non-smoothness and non-convexity of the initial top-k ranking loss, we exploit a tight convex upper bound to approximate the loss and then utilize the traditional back-propagation algorithm to optimize the deep multi-modal network. Finally, we apply the method on three benchmark datasets, namely, Flickr8k, Flickr30k, and MSCOCO. Empirical results on metrics R@K(K=1,5,10) show that our method achieves comparable performance to state-of-the-art methods.**

***Index Terms*—Image-sentence matching, Cross-modal retrieval, Deep learning, Top-k ranking.**

I. Introduction

Data related to a certain concept typically appear in diverse modalities with the rapid development of the Internet. Learning a certain concept from different modality data is beneficial for human cognition [1]. For example, when we aim to learn to play tennis, we prefer to find relevant results across various modalities, such as some articles describing tennis skills, images illustrating body movements, and videos recording tennis teaching. Consequently, cross-modal retrieval, where the queries and results are typically from different modalities, plays a significant role in numerous real-world applications. In this paper, we focus on the bi-directional retrieval between images and sentences, *i.e.*, image-sentence matching task. Previous studies on this task can be divided into two types, namely weakly and strongly supervised methods. The former only requires the image-sentence pairs for training [2], [3]; the latter demands the category labels for images and sentences to better mine their semantics [4]–[6]. The strongly supervised methods are impractical in the real world because of two reasons: (1) The labeled images and sentences are not available due to the annotation cost [7]; (2) the query in the testing stage may fall outside the categories in the training phase. Therefore, this study focuses on image-sentence matching with weakly supervised pair information.

Two main strategies for implementing the image-sentence matching task are available for past weakly supervised methods. The first group focuses on learning a common representations for images and sentences based on canonical correlation analysis (CCA) [8]. This group seeks the linear or nonlinear projections by maximizing the correlation between the projected vectors of images and sentences [9]. These CCA-based methods have two limitations: 1) They do not maximize the evaluation criterion related to the ultimate cross-modal ranking performance [10]; 2) they commonly require high memory cost to compute the covariance matrix between the entire images and sentences [11]. For these issues, the second group aims to minimize a ranking-based loss for image-sentence matching. Specifically, the ranking-based methods first map the entire sentences and the given image query to a shared embedding space, and then compute the relevance scores of the entire image-sentence pairs. Notably, the traditional ranking loss forces the correctly matched sentence for the given image query to be ranked higher than other mismatched ones [12]. Compared with CCA-based methods, on the one hand, these ranking-based methods reflect the performance of cross-modal ranking directly. On the other hand, they are more appropriate for large-scale image-sentence matching without constructing the data covariance matrix. Thus the ranking-based methods have attracted increasing attention for their improved interpretation and minimal computational consumption.

The traditional ranking loss commonly assumes that the query result is correct if and only if the matched instance for the given query ranks first. This assumption is of course appropriate for retrieving from a database only with distinguishable instances. However, for a large-scale cross-modal retrieval task, having certain highly similar and confusable instances in the query database is inevitable. These instances may deteriorate the training process of the retrieval model if we continue to use the previous ranking loss. As shown in Figure 1, there are two image-sentence pairs about playing tennis, *i.e.* (image$_1$, sentence$_1$) and (image$_2$, sentence$_2$). Although the first sentence is correctly matched to the first image in the query database, it can also describe the content of the second image to a certain extent. Taking the first sentence as a query, forcing the first image to rank before the second image is over strict. Thus, for a real-world information retrieval task with certain indistinguishable data, the traditional ranking function is inapplicable and may be prone to the overfitting

Corresponding author: Minnan Luo.

Lingling Zhang is with Ministry of Education Key Lab For Intelligent Networks and Network Security, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China, e-mail: zhanglingling@stu.xjtu.edu.cn

Minnan Luo and Jun Liu are with National Engineering Lab for Big Data Analytics, Xian Jiaotong University, Xi'an 710049, China (e-mail: {minnluo, liukeen}@mail.xjtu.edu.cn).

Xiaojun Chang is with Faculty of Information Technology, Monash University, Australia (e-mail: cxj273 @gmail.com).

Yi Yang is with Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia (e-mail: yee.i.yang@gmail.com).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, PA 15213, USA (e-mail: alex@cs.cmu.edu).

Sentence1: Many people are playing tennis together in tennis hall.

Sentence2: There are two persons playing tennis on a professional court.

Fig. 1: The example of confusable image-sentence pairs. Although the first sentence is matched to the first image in the query database, it could also describe the detailed content of the second image.

problem. Considering the above mentioned issues, we propose a novel deep top-k ranking loss in a multi-modal network to mitigate the instance ambiguity problem. Given a query, the top-k ranking loss allows the correctly matched instance to be among the largest k query results, rather than be ranked first. In other words, the top-k ranking loss is not penalized for the first $(k-1)$ mismatched instances. This strategy diminishes the extravagant requirement of previous ranking loss. Three contributions of this paper are summarized as follows:

- To mitigate the instance ambiguity problem, we propose a deep multi-modal network with a top-k ranking loss for large scale image-sentence matching with indistinguishable data. This method allows k query to be guessed and penalizes $(n-k)$ mismatched image-sentence pairs.
- We propose an effective convex upper bound function to approximate the initial non-smooth and non-convex top-k ranking loss. Notably, the convex top-k ranking loss does not increase the time complexity in comparison with the traditional ranking loss in theory.
- We conduct extensive experiments on benchmark datasets Flickr8k, Flickr30k, and MSCOCO to illustrate the effectiveness and superiority of the proposed method. The experimental results demonstrate that our method achieves comparable performance to state-of-the-art methods.

II. Related Work

A. Image-Sentence Matching

At present, CCA-based methods and ranking-based methods are two main streams for accomplishing the image-sentence matching task. Certain methods based on deep learning (DL) have been proposed recently to learn the semantic representations for images and sentences effectively.

1) ***CCA-Based Methods****:* Several recent studies on image-sentence matching task are based on CCA. CCA learns the common representations for images and sentences by mutually maximizing the correlation between their projections [9]. Kernel CCA (KCCA) is an extension of CCA, which finds maximally correlated nonlinear projections restricted to reproducing kernel Hilbert spaces with corresponding kernels [13]. Sparse KCCA (SKCCA) [14] incorporates sparsity into KCCA by penalizing the ℓ_1-norm of dual vectors. To improve the scalability of CCA on large data training, deep CCA (DCCA) [15] is proposed to optimize the CCA objective function in a deep learning framework. Hodosh *et al.* [16] pioneered the application of CCA to sentence-based image description, where the image and sentence features were obtained with minimal supervision. Gong *et al.* [17] started with CCA and incorporated a third view to capture the high-level image semantics, which accomplished the image-to-image, tag-to-image, and image-to-tag three-directional information retrieval. Wang *et al.* [18] scaled the columns of the CCA projection matrices by a power of the corresponding eigenvalues, which achieved a significant improvement on retrieval results. Klein *et al.* [19] utilized the CCA algorithm to accomplish the sentence-image matching task, where the sentences were represented as the new variants of fisher vectors (FV). Specifically, Yan *et al.* [20] proposed a GPU implementation and a novel strategy to handle overfitting issues for the original DCCA. Additionally, the CCA algorithm has been utilized in other fields, such as chemometrics [21], speech processing [22], [23], and multimodal signal processing [24]. Although CCA has been confirmed to be a surprisingly effective method for the image-sentence matching task, it requires high memory cost to compute the covariance matrix for the entire images and sentences. Although the proposed DCCA utilizes the stochastic gradient descent (SGD) algorithm to optimize the original CCA objective function, it cannot obtain a favorable solution because covariance estimation in each minibatch is unequal to real covariance over all data.

2) ***Ranking-Based Methods****:* Another notable body of image-sentence matching tasks is to learn a shared embedding space based on a ranking loss function. A single-directional ranking loss ensures that the correct text description ranks higher than other irrelevant descriptions for a given image. This ranking loss has been applied in [25], [26] to achieve image annotation. After that, some researchers have proposed the bi-directional ranking loss to strengthen the intensity of single-directional loss [27], [28]. With a given sentence query, the bi-directional ranking loss adds a margin-based penalty to a mismatched image when it ranks higher than the matched image [29]. Luo *et al.* [10] integrated diversity self-paced learning theory into the bi-directional ranking loss to enhance the model's robustness to outliers. To hold the neighborhood structure within each single-model view, Wang *et al.* [30] combined the cross-view bi-directional ranking constraints with within-view structure-preserving ranking constraints. They [31] also introduced an appropriate image-sentence pair sampling for ranking loss to improve the matching performance. Nam *et al.* [32] integrated the attention mechanisms into the bi-directional retrieval through a novel dual attention network. Huang *et al.* [33] constructed a selective multi-modal long short-term memory network (LSTM) with a bi-directional ranking loss for instance-aware image and sentence matching. Compared with CCA, the ranking loss function does not require the construction of the data covariance matrix, which can be readily utilized for large-scale image-sentence matching tasks. Nevertheless, the present ranking constraints strictly demand that the score of matched image-sentence pair is higher than that of other mismatched image-sentence pairs. This requirement is severe because the data ambiguity typically exists

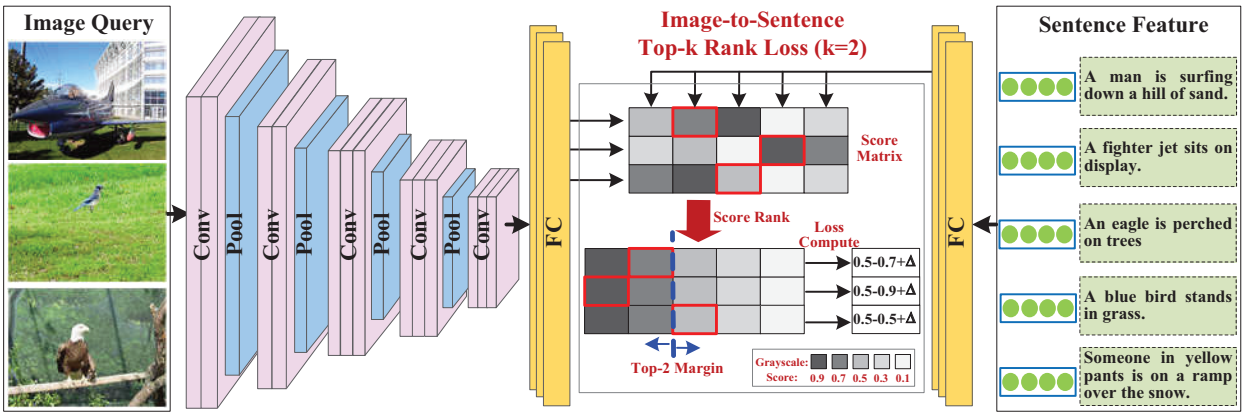


Fig. 2: Diagram of the image-query-sentence process based on top-k ranking loss. Left: image network (VGG16). Right: text network (MLP). Center: the top-k ranking loss, where k=2. There are three given image queries and five sentences in query database. For each given image, we utilize five gray boxes to denote the five relevance scores of image-sentence pairs (*i.e.* the score matrix part in the Figure). The larger grayscale represents the larger score. The red box is the score of correct matched image-sentence pair. We rank the relevant scores for each image query and stipulate that the score of matched image-sentence pair is higher than the score of 3-th image-sentence pairs by at least a margin of Δ.

when large amounts of image-sentence pairs are used for training. For this issue, we propose a novel top-k ranking loss to better accomplish image-sentence matching in this paper.

*3) **DL-Based Methods**:* Deep learning has achieved great considerable success in image recognition and natural language processing. To better achieve image-sentence matching, most researchers utilize the deep nonlinear mappings to capture powerful semantic features for images and sentences. Donahue *et al.* [34] developed a novel end-to-end recurrent convolutional architecture for large-scale visual learning, and demonstrated its effectiveness on retrieval problems. Kiros *et al.* [35] described an image-text multi-modal neural language model and showed jointly learning of the word representations and image features for image-text retrieval. Ma *et al.* [36] provided an end-to-end framework with convolutional architectures to exploit image representations, word compositions, and matching relations between the two modalities. Eisenschtat *et al.* [37] utilized two tied neural network channels to project image and sentence views into a common, maximally correlated space. In this paper, our embedding functions are designed with deep networks, where the 16-layers VGG convolutional network (VGG16) [38] and multi-layer perceptrons (MLP) [39] are utilized as image and text networks respectively.

B. Top-k Theory

The top-k strategy has received renewed attention with the advent of large-scale problems [40]. At present, the top-k strategy is applied to mitigate the category ambiguity problem in large-scale multi-class classification. To be specific, the traditional multi-class methods demand that the top-1 predicted category with the largest score is exactly identical with its ground truth label. This requirement is suitable for the simple classification only with clearly distinguishable categories. However, for large-scale multi-class classification, some classes are highly similar or confusable with one another [41]. In this case, traditional multi-class methods are too strict and may suffer from overfitting. To this issue, Lapin *et al.* [42] relaxed the penalty for making mistakes by considering the top-k predictions rather than only the top one. Furthermore, they [43] also introduced the multi-class smooth top-k hinge loss and provided an efficient optimization scheme for it. After that, Chang *et al.* [44] presented a generic, robust top-k multi-class method for visual category recognition. Yan *et al.* [45] integrated multiple feature fusion into the top-k multi-class framework to improve the classification performance. In this paper, we incorporate the top-k strategy into large-scale cross-modal retrieval to relax the extravagant requirement of the traditional ranking-based methods.

III. Deep Top-k Image-Sentence Matching

In the framework of the image-sentence matching, we are provided with two collections of images and sentences, *i.e.* $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and $\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_n\}$ correspondingly. $\mathbf{x}_i \in \mathbb{R}^{224\times224\times3}$ represents the RGB pixel values of the i-th image. $\mathbf{t}_i \in \mathbb{R}^m$ refers to a m-dimensional feature vector extracted from the i-th sentence. Notably, the order of images in $\mathcal{X}$ and sentences in $\mathcal{T}$ corresponds to each other. In other words, each image $\mathbf{x}_i \in \mathcal{X}$ is associated with the sentence description $\mathbf{t}_i \in \mathcal{T}$, such that the training dataset consists of n image-sentence pairs, *i.e.* $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i) : i = 1, 2, \cdots, n\}$. We take the image-query-sentence for an example. Given image query $\mathbf{x}_i$, the image-sentence matching task aims to return the matched sentence $\mathbf{t}_i$ from the large database $\mathcal{T}$. To perform this task, we first construct an end-to-end deep multi-modal network to project images $\mathcal{X}$ and sentences $\mathcal{T}$ into a shared embedding space (Section III-A). Then we propose a novel top-k ranking loss to optimize the parameters of the deep multi-modal network (Section III-B).

A. Network Architecture

In Figure 2, the deep multi-modal network consists of two single-modal networks, *i.e.* the image network and the text

network. This multi-modal network aims to map the original images and sentences to a d-dimensional shared embedding space. Considering the success of deep convolutional network on image recognition, we adopt VGG16 [38] with a modified objective function for the image network. In the left part of Figure 2, VGG16 is composed of thirteen 3×3 convolutional layers ($conv1 - conv13$, the purple layers), four max-pooling layers (the blue layers), and three fully connected layers ($fc14 - fc16$, the orange layers). The raw images first pass the thirteen convolutional layers and four max-pooling layers to extract the deep features, then through three fully connected layers to be represented as d-dimensional semantic features. The text network is illustrated in the right part of Figure 2. The network is designed as a MLP [46] network with three fully connected layers ($fc1 - fc3$, the orange layers). The original sentence features, such as FV [19], pass three fully connected layers to be encoded as d-dimensional vectors in the shared embedding space. Considering the good performance of ReLU activation, we adopt the $a(x) = \max(0, x)$ as the activation function for all convolutional and fully connected layers in our multi-modal network.

B. Cross-Modal Top-k Ranking

In this section, a novel top-k ranking loss is introduced to optimize the deep multi-modal network. With the input of images $\mathcal{X}$ and sentences $\mathcal{T}$, the objective function (1) finetunes the network by jointly minimizing a bi-directional top-k ranking loss and a regularization term as follows:

$$\min_{\mathbf{W}_x, \mathbf{W}_t} \left(\lambda r(\mathbf{W}_x, \mathbf{W}_t) + \sum_{i=1}^{n} \underbrace{L\left(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}; \mathbf{W}_t), k\right)}_{image-query-sentence} + \sum_{i=1}^{n} \underbrace{L\left(\Psi(\mathbf{t}_i; \mathbf{W}_t), \Phi(\mathcal{X}; \mathbf{W}_x), k\right)}_{sentence-query-image} \right). \tag{1}$$

The variables $\mathbf{W}_x$ and $\mathbf{W}_t$ denote the trainable parameter sets of the image and text networks respectively. $r(\mathbf{W}_x, \mathbf{W}_t)$ is the regularization term with a shrinkage coefficient λ. It denotes the squared sum of parameters in $\mathbf{W}_x$ and $\mathbf{W}_t$, *i.e.* $r(\mathbf{W}_x, \mathbf{W}_t) = \frac{1}{2} \sum_{w \in \{W_x, W_t\}} w^2$. $\Phi(\cdot; \mathbf{W}_x)$ and $\Psi(\cdot; \mathbf{W}_t)$ indicate the deep non-linear mapping functions for image and text networks with the parameters $\mathbf{W}_x$ and $\mathbf{W}_t$, correspondingly. They encode the single instance $\mathbf{x}_i, \mathbf{t}_i$ or batch instances $\mathcal{X}, \mathcal{T}$ as new d-dimensional representations in the shared embedding space. The novel top-k ranking loss terms $L(\cdot)$ during image-query-sentence and sentence-query-image processes can be obtained using Equations (2) and (3).

$$L\left(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}; \mathbf{W}_t), k\right) = \max\left(0, S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta\right), \tag{2}$$

$$L\left(\Psi(\mathbf{t}_i; \mathbf{W}_t), \Phi(\mathcal{X}; \mathbf{W}_x), k\right) = \max\left(0, S(\mathcal{X}^{\setminus \mathbf{x}_i}, \mathbf{t}_i)_{[k]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta\right), \tag{3}$$

where k and Δ are two hyper-parameters that can be obtained by cross-validate. The collections $\mathcal{X}^{\setminus \mathbf{x}_i}$ and $\mathcal{T}^{\setminus \mathbf{t}_i}$ separately remove the i-th instances in $\mathcal{X}$ and $\mathcal{T}$, *i.e.* $\mathcal{X}^{\setminus \mathbf{x}_i} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_n\}$ and $\mathcal{T}^{\setminus \mathbf{t}_i} = \{\mathbf{t}_1, \cdots, \mathbf{t}_{i-1}, \mathbf{t}_{i+1}, \cdots, \mathbf{t}_n\}$. The value of $S(\mathbf{x}_i, \mathbf{t}_j)$ denotes the similarity (relevance score) between image $\mathbf{x}_i$ and sentence $\mathbf{t}_j$ in the shared embedding space, which is obtained by computing their cosine distance as follows:

$$S(\mathbf{x}_i, \mathbf{t}_j) = \Phi(\mathbf{x}_i; \mathbf{W}_x)^\top \Psi(\mathbf{t}_j; \mathbf{W}_t). \tag{4}$$

In this case, we define the relevance scores between image $\mathbf{x}_i$ and the entire sentences as $S(\mathbf{x}_i, \mathcal{T})$, which is represented as:

$$S(\mathbf{x}_i, \mathcal{T}) = \{S(\mathbf{x}_i, \mathbf{t}_1), \cdots, S(\mathbf{x}_i, \mathbf{t}_i), \cdots, S(\mathbf{x}_i, \mathbf{t}_n)\}. \tag{5}$$

In the image-query-sentence stage, given image $\mathbf{x}_i$, we sort the matching scores of all sentences $\mathcal{T}$ in descending order:

$$S(\mathbf{x}_i, \mathcal{T})_{[1]} \geq S(\mathbf{x}_i, \mathcal{T})_{[2]} \geq \cdots \geq S(\mathbf{x}_i, \mathcal{T})_{[n]}, \tag{6}$$

where the bracket $[\cdot]$ denotes a permutation of querying results. In other words, $S(\mathbf{x}_i, \mathcal{T})_{[r]}$ is the r-th largest score in $S(\mathbf{x}_i, \mathcal{T})$. Traditional ranking loss forces the matched image-sentence pair to rank in front of all other mismatched pairs, *i.e.*, $S(\mathbf{x}_i, \mathcal{T})_{[1]} = S(\mathbf{x}_i, \mathbf{t}_i)$. Apparently, this assumption is improper because some sentences in a large database may be highly similar and confusable to the correctly matched sentence $\mathbf{t}_i$. To mitigate this issue, we relax the traditional ranking loss and propose a top-k ranking loss for the cross-modal retrieval. With this strategy, the querying results are not penalized unless the index of the ground truth sentence is outside the range of top-k query results. Formally, the top-k ranking constraint assumes that the query result is correct for a given image $\mathbf{x}_i$ if its corresponding sentence $\mathbf{t}_i$ satisfies the condition $S(\mathbf{x}_i, \mathbf{t}_i) \geq S(\mathbf{x}_i, \mathcal{T})_{[k+1]}$, *i.e.*,

$$S(\mathbf{x}_i, \mathbf{t}_i) \geq S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}, \tag{7}$$

where $S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})$ is obtained by removing the score value $S(\mathbf{x}_i, \mathbf{t}_i)$ in collection $S(\mathbf{x}_i, \mathcal{T})$. In that case, the top-$k$ ranking loss with image query $\mathbf{x}_i$ can be defined as

$$\begin{aligned} &L(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}; \mathbf{W}_t), k) \\ &= \max(0, S(\mathbf{x}_i, \mathcal{T})_{[k+1]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta) \\ &= \max(0, S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta). \end{aligned} \tag{8}$$

Equation (8) guarantees that the score with the matched sentence for image query $\mathbf{x}_i$, *i.e.*, $S(\mathbf{x}_i, \mathbf{t}_i)$, must be higher than the score with $(k+1)$-th sentence, *i.e.*, $S(\mathbf{x}_i, \mathcal{T})_{[k+1]}$, by at least a margin of Δ. When k is set to one, Equation (8) forces that the first rank result of image query $\mathbf{x}_i$ must be sentence $\mathbf{t}_i$, which is apparently overstrict due to the data ambiguity problem. Specifically, the term $S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}$ in Equation (8), *i.e.*, returning the k-th largest element in $S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})$, is non-convex and non-smooth when $k \geq 2$. With the strategy in [47], we propose a tight upper bound as expressed in the following equation to approximate it:

$$\widetilde{S}(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]} = \frac{1}{k} \sum_{r=1}^{k} (S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[r]}) \geq S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}. \tag{9}$$

Notably, the sum of k largest components in $\widetilde{S}(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}$ is convex according to the research in [48]. Specifically, $\widetilde{S}(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]} = S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}$ if and only if all k largest image-sentence pair scores in $S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})$ are the same. The value of $\left(\widetilde{S}(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]} - S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}\right)$ can be regarded as a extra slack. It can expand the score margin between the first k image-sentence pairs and the remaining $(n-k)$ least similar pairs. This strategy may lead to an improvement over the initial top-k ranking loss. Therefore, $\widetilde{S}(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}$ is a much better approximation for $S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[k]}$. In this case, for image-query-sentence, we propose the following convex top-k ranking loss to approximate the original non-convex loss (2):

$$\widetilde{L}\left(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}; \mathbf{W}_t), k\right) = \max\left(0, \frac{1}{k}\sum_{r=1}^{k}(S(\mathbf{x}_i, \mathcal{T}^{\setminus \mathbf{t}_i})_{[r]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta\right). \quad (10)$$

We omit the detailed explanation of top-k ranking loss during the sentence-query-image procedure, *i.e*, the meaning of $L(\Psi(\mathbf{t}_i; \mathbf{W}_t), \Phi(\mathcal{X}; \mathbf{W}_x), k)$ in Equation (3). Apparently, this explanation is analogous to illustrating the top-k ranking loss mentioned in the image-query-sentence stage.

C. Time Complexity Analysis

In this section, we analyze the time complexity of traditional ranking loss and top-k ranking loss in theory. For large-scale cross-modal retrieval, the objective function is generally applied in each mini-batch to optimize the deep network. We assume b (b = 128 in our experiment) image-sentence pairs for each mini-batch. These images $\mathcal{X}_b$ and sentences $\mathcal{T}_b$ first pass through the multi-modal network to be represented in a shared embedding space. The time cost during this procedure is only related to the network architecture and is independent of the loss function. After that, the ranking loss is measured over these deep representations to update the network parameters. Taking image-query-sentence as an example, the traditional ranking loss is defined in accordance with the work [49]:

$$\hat{L}\left(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}_b; \mathbf{W}_t)\right) = \sum_{j \neq i}\max(0, S(\mathbf{x}_i, \mathbf{t}_j) - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta). \quad (11)$$

The top-k ranking loss over the mini-batch is computed as:

$$\widetilde{L}\left(\Phi(\mathbf{x}_i; \mathbf{W}_x), \Psi(\mathcal{T}_b; \mathbf{W}_t), k\right) = \max\left(0, \frac{1}{k}\sum_{r=1}^{k}(S(\mathbf{x}_i, \mathcal{T}_b^{\setminus \mathbf{t}_i})_{[r]} - S(\mathbf{x}_i, \mathbf{t}_i) + \Delta\right). \quad (12)$$

By comparing Equation (12) and (11), given image query $\mathbf{x}_i$, the top-k ranking loss and traditional ranking loss all consume the same time to obtain b matching scores for all sentences, *i.e.*, $S(\mathbf{x}_i, \mathcal{T}_b)$. After that, the traditional ranking loss stipulates that the score for matched image-sentence pair $S(\mathbf{x}_i, \mathbf{t}_i)$ must be higher than all other mismatched scores $S(\mathbf{x}_i, \mathcal{T}_b^{\setminus \mathbf{t}_i})$ by a margin of Δ. This procedure requires $O(b)$ time cost to compute the loss value for the given image query $\mathbf{x}_i$. For top-k ranking loss, we compute the average of the largest k matching scores, and then obtain the loss value for one image query. According to the work [50], the time cost for top-k ranking is $O(b + klogk)$ if we use the quick top-k sorting algorithm. Note that $O(b + klogk)$ is approximately equal to $O(b)$ because the k in our top-k ranking loss is generally small for image-sentence retrieval. Therefore, the top-k and traditional ranking losses consume the nearly equivalent time cost for the image-sentence matching task.

IV. Experiment

In this section, we conduct extensive experiments on three benchmark datasets to validate the effectiveness of our deep top-k ranking loss for image-sentence matching. Three benchmark datasets, eleven competitors, and implementation details are introduced in Section IV-A. Some image-sentence matching examples and performance comparisons are discussed in Section IV-B. We analyze the impacts of the top-k strategy and the embedding dimension of multi-modal space on retrieval performance in Sections IV-C and IV-D.

A. Experimental Setup

1) Datasets: We perform the experiments on three benchmark datasets, namely, Flickr8k, Flickr30k, and MSCOCO.

- **Flickr8k** [16]: The dataset consists of 8,000 images from Flickr.com website. The dataset focuses on some activities of people and animals. Each image is annotated with five sentences using Amazon Mechanical Turk. These sentences generally describe the objects, scenes, and activities for the corresponding images.
- **Flickr30k** [51]: The dataset is an extension of Flickr8k with 31,783 images. Each image also has five corresponding sentences to describe its content, where the sentences are obtained in a style similar to Flickr8k.
- **MSCOCO** [52]: The dataset contains 123,287 images, and each image corresponds to five descriptive sentences. The dataset includes 91 object categories and 82 of them have more than 5,000 images.

Some examples of images and the corresponding sentences in Flickr8k, Flickr30k, and MSCOCO are demonstrated in Figure 3. For each dataset, we utilize 1,000 images for validation, 1,000 images for testing, and the remainder for training.

2) Competitors: We compare our method with eleven state-of-the-art baselines. The CCA (Mean-Vector), CCA (FV HGLMM), CCA (FV GMM+HGLMM), and DCCA are four representative CCA-based methods. The mCNN, DSPE, Two-branch Nets (Embedding), Dual-Attention Nets, and sm-LSTM are five ranking-based methods. The m-RNN and Two-way Nets are two other DL-based methods. The detailed introduction of these competitors is presented as follows:

- **CCA (Mean-vector)** [19]: A primary CCA-based method for the image-sentence matching task. For the method, each sentence is mapped to a set of word2vec vectors and then the average vector of this set is utilized for the sentence representation.
- **CCA (FV)** [19]: A CCA-based method that extracts the sentence FV features as the input of the model. Specifically, the method "CCA (FV HGLMM)" adopts HGLMM

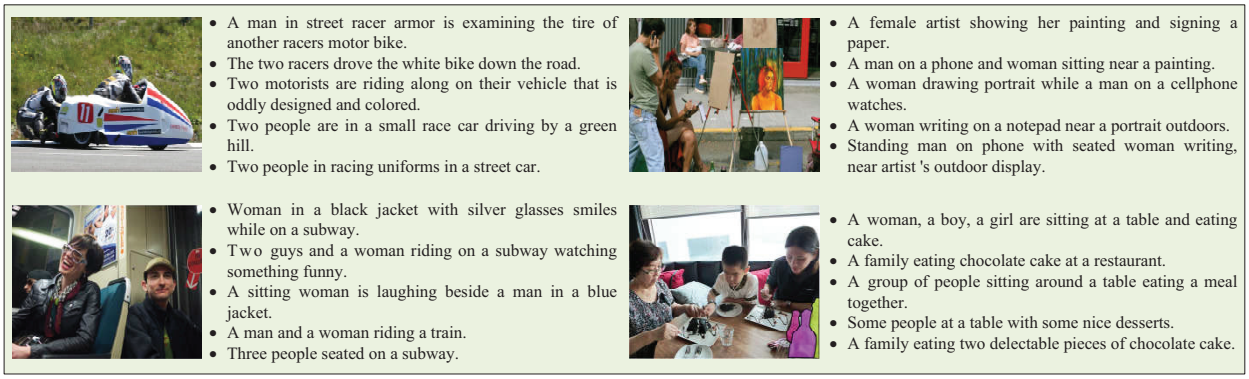


Fig. 3: Some examples of image-sentence pairs from three datasets.

pooling strategy to generate better fisher vectors. And the "CCA (FV GMM+HGLMM)" method utilizes the fusion of GMM with HGLMM pooling strategies.

- **DCCA** [20]: An alternative end-to-end learning scheme based on the deep CCA to achieve cross-modal retrieval, where the image part of the deep network is initialized using a pre-trained AlexNet model [54]. This method employs simple term frequency inverse document frequency (TF-IDF) features to represent text descriptions.
- **mCNN** [36]: A novel bi-directional ranking-based method for cross-modal retrieval, which includes the matching CNN and MLP two parts. A matching CNN takes images and words as the input and produces the joint multi-modal representations. Then MLP summarizes the joint representation and outputs the matching score.
- **DSPE** [30]: A state-of-the-art ranking-based method for the image-sentence matching task. This method considers not only the cross-view ranking constraints, but also the within-view neighborhood structure preservation constraints inspired by metric learning literature.
- **Two-branch Nets** [31]: Embedding or similarity network for the image-text matching task. The Two-branch Nets with optimal results is an embedding network that is the extension of competitor DSPE [30] by utilizing a neighborhood sampling strategy.
- **Dual-Attention Nets** [32]: A ranking-based method that jointly leverages visual and textual attention mechanisms, where we only show results with VGG image features because we also use this type of features in our model.
- **sm-LSTM** [33]: An bi-directional ranking-based method for cross-modal retrieval, which recurrently selects salient pairs of image-sentence instances, and then measures their local similarities within several time steps.
- **m-RNN** [53]: A multi-modal deep network for generating novel image captions. It consists of two sub-networks: a RNN network for sentences and a CNN network for images. For the image part, we respectively utilize AlexNet for Flickr8k, and VggNet for Flickr30k and MSCOCO datasets. We denote them as m-RNN-AlexNet and m-RNN-VggNet in the experiment.
- **Two-way Nets** [37]: A novel neural network for matching vectors from two data sources. This method uses two tied neural network channels to project the two views into a common, maximally correlated space.

3) Implementation: We implement our model based on the open-source *PyTorch* framework. To augment the number of images, we crop 224×224 regions from the original 256×256 images in ten different ways, including the four corners, the center, and their x-axis mirror images. We extract the sentence FV features as the input of the MLP text network. Specifically, we first represent each sentence word as a 300-dimensional word2vec semantic feature, and then construct a codebook with 30 centers through independent component analysis (ICA) on the word2vec features. The FV features for sentences with $300 \times 30 \times 2 = 18000$ dimensions are obtained based on HGLMM pooling strategy [19]. Finally, we conduct principal components analysis (PCA) [55], [56] to reduce the sentence dimension from 18,000 to 6,000. In this case, each sentence is denoted as a 6,000-dimensional vector.

For the image network, the parameters of convolutional layers ($conv1 - conv13$) in VGG16 are pre-trained on ImageNet to avoid overfitting. The numbers of neurons for fully connected layers ($fc14 - fc16$) are set to 4,096, 1,024, and 512, respectively. The output dimensions of layers ($fc1 - fc3$) in the text network are set to 1,024, 512, and 512, correspondingly. Note that we add the dropout layers with probability 0.3 and apply batch normalization for the fully connected layers to improve the network stability. In the training stage, each mini-batch contains 128 image-sentence pairs. We utilize the SGD algorithm with momentum 0.9 and weight decay 0.001 to optimize our multi-modal network. The learning rate is initialized as 0.1 and then decreased by 5% after each epoch.

B. Performance Comparison

In Figures 4 and 5, we show some query results of image retrieval and sentence retrieval on Flickr30k dataset, respectively. We only list the top-5 predicted results, where the correct predictions appear in red font or red box. The R@K(K=1,5,10) results over three datasets are resulted in Tables I, II, and III, where the retrieval results for all competitors are directly obtained from the corresponding published papers. To illustrate the effectiveness of top-k ranking loss adequately, we assign the value of hyper-parameter k in the interval $\{4, 5, 6\}$. According to these results, we have the following observations:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2931352, IEEE Transactions on Multimedia

- Three children pose for a photograph by a rock.
- Three children pose outdoors on a hike.
- Children sitting on rock in the desert.
- A young boy sitting on a rock pile.
- Children are playing on a playground set with one another.

- The white and black dog runs on the field with his tongue hanging out.
- A dog is running through a field with its tongue hanging out.
- Dog walking with his lease in his mouth.
- A dog with its mouth open is running.
- A brown and white dog runs away carrying a big stick.

- A dirt biker rides up a rocky hill on a motorized dirt bike.
- A young man pushes his motocross bike up a dirt hill.
- A person on a motorcycle going uphill on rocky terrain.
- An off road motorbike climbs a rocky hill.
- A man wearing a helmet and safety suit goes uphill over rocks on a dirt bike.

- A man and woman stand next to a table covered in beer glasses and pitchers.
- A woman leaning her head on a man 's shoulder.
- A smiling woman sitting on a smiling man 's lap.
- A woman and a man are talking at a bar.
- A woman leaning her head on a man 's shoulder.

Fig. 4: Some testing results of image-query-sentence on Flickr30k dataset. We show top-5 sentences for each image query, and the retrieved correct sentences are shown in red font.

Fig. 5: Some testing results of sentence-query-image on Flickr30k dataset. We show top-5 images for each sentence query, and the retrieved correct images are marked with red borders.

- Compared with the four CCA-based methods, mCNN, and m-RNN, our method achieves the optimal retrieval results over three benchmark datasets. Specifically, the R@1 for image-query-sentence improves by 2.3% on Flickr8k, 7.0% on Flickr30k, and 8.3% on MSCOCO; the R@1 for sentence-query-image improves by 1.8% on Flickr8k, 6.3% on Flickr30k, and 10.0% on MSCOCO.
- Our method achieves comparable performance to five state-of-the-art methods, including DSPE, Two-branch Nets (Embedding), Two-way Nets, Dual-Attention Nets (VGG), and sm-LSTM. Specifically, our method obtains better results on R@5 but poorer on R@1 than Two-way Nets. This result is consistent with the conclusion in [31]. Apparently, the top-k ranking is very promising if it considers the structure information within modalities like methods DSPE and Two-branch Nets, or if it is applied on better image-sentence features and network architectures like methods Dual-Attention Nets and sm-LSTM.
- The majority of approaches perform optimally on dataset MSCOCO, then Flickr30k, and lastly Flickr8k. For our method with top-4 ranking loss, the R@1 values over datasets Flickr8k, Flickr30k, and MSCOCO are 33.3%, 42.4%, and 51.1% respectively in the image-query-sentence stage. This is reasonable because more training data in MSCOCO is conductive to improving the model's generalization ability.

C. Impact of the top-k strategy

The top-k ranking loss is an improvement on traditional ranking loss, which can be extensively applied to any ranking-based retrieval methods. To confirm its effectiveness further, we conduct an experiment to compare the retrieval performance of traditional and top-k ranking losses in the same experimental setting. Considering the simplified version of ranking-based methods DSPE, Two-branch Nets, and Dual-Attention Nets have been accomplished and opened by some researchers, we thus directly revise these codes to replace the original cross-modal ranking loss as the top-k loss to verify whether the top-k strategy can improve the retrieval performance. Notably, numerous versions of Two-branch Nets

TABLE I: The ranking performance comparison in terms of R@K over Flickr8k.

Methods	Image-query-Sentence			Sentence-query-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CCA(FV Mean-vector) [19]	22.6	48.8	61.2	19.1	45.3	60.4
CCA (FV HGLMM) [19]	28.5	58.4	71.7	20.6	49.4	64.0
CCA (FV GMM+HGLMM) [19]	31.0	59.3	73.7	21.3	50.0	64.8
DCCA [20]	28.2	56.1	69.8	26.3	54.0	67.5
mCNN [36]	24.8	53.7	67.1	20.3	47.6	61.7
m-RNN-AlexNet [53]	14.5	37.2	48.5	11.5	31.0	42.4
Two-way Nets [37]	**43.4**	**63.2**	-	**29.3**	49.7	-
TOP-k Ranking (k=4)	33.3	61.8	74.6	28.1	55.9	68.1
TOP-k Ranking (k=5)	32.7	62.7	75.3	27.3	**57.8**	**69.3**
TOP-k Ranking (k=6)	32.1	**63.2**	**75.5**	24.1	54.7	68.7

TABLE II: The ranking performance comparison in terms of R@K over Flickr30k.

Methods	Image-query-Sentence			Sentence-query-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CCA(FV Mean-vector) [19]	24.8	52.5	64.3	20.5	46.3	59.3
CCA (FV HGLMM) [19]	34.4	61.0	72.3	24.4	52.1	65.6
CCA (FV GMM+HGLMM) [19]	35.0	62.0	73.8	25.0	52.7	66.0
DCCA [20]	27.9	56.9	68.2	26.8	52.9	66.9
mCNN [36]	33.6	64.1	74.9	26.2	56.3	69.6
m-RNN-VggNet [53]	35.4	63.8	73.7	22.8	50.7	63.1
DSPE [30]	40.3	68.9	79.9	29.7	60.1	72.1
Two-branch Nets (Embedding) [31]	43.2	**71.6**	79.8	31.7	61.3	72.4
Two-way Nets [37]	**49.5**	67.5	-	**36.0**	55.6	-
Dual-Attention Nets (VGG) [32]	41.4	**73.5**	82.5	31.8	61.7	72.5
sm-LSTM [33]	42.4	67.5	79.9	28.2	57.0	68.4
TOP-k Ranking (k=4)	42.4	69.0	77.4	30.2	58.3	70.1
TOP-k Ranking (k=5)	41.3	70.3	79.8	33.1	61.5	72.9
TOP-k Ranking (k=6)	37.8	68.3	**82.6**	31.2	**63.1**	**75.2**

TABLE III: The ranking performance comparison in terms of R@K over MSCOCO.

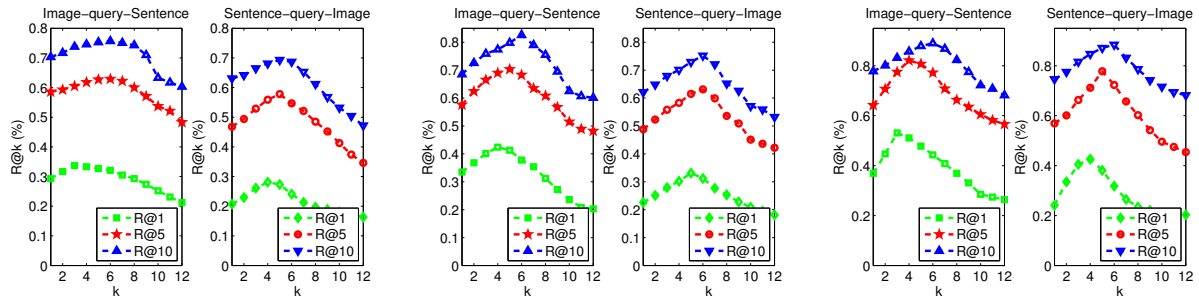
Methods	Image-query-Sentence			Sentence-query-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CCA(FV Mean-vector) [19]	33.2	61.8	75.1	24.2	56.4	72.4
CCA (FV HGLMM) [19]	37.7	66.6	79.1	24.9	58.8	76.5
CCA (FV GMM+HGLMM) [19]	39.4	67.9	80.9	25.1	59.8	76.6
mCNN [36]	42.8	73.1	84.1	32.6	68.6	82.8
m-RNN-VggNet [53]	41.0	73.0	83.5	29.0	42.2	77.0
DSPE [30]	50.1	79.7	89.2	39.6	75.2	86.9
Two-branch Nets (Embedding) [31]	54.0	**84.0**	**91.2**	**43.3**	76.8	87.6
Two-way Nets [37]	**55.8**	75.2	-	39.7	63.3	-
sm-LSTM [33]	52.4	81.7	90.8	38.6	73.4	84.6
TOP-k Ranking (k=4)	51.1	82.1	85.7	42.6	71.2	84.7
TOP-k Ranking (k=5)	47.8	80.7	87.9	38.1	**77.8**	87.1
TOP-k Ranking (k=6)	44.3	77.2	89.0	31.8	72.3	**88.4**

exist in [31], where previous competitor DSPE is the one by considering the neighborhood structure within modalities. To be different from the setting of DSPE, we select the version of Two-branch Nets that do not contain the within-view structure-preserving constraints. Table IV presents the retrieval performance with traditional and top-k ($k = 5$) ranking losses on dataset Flickr30k. The R@K (K=1,5,10) performances of methods DSPE, Two-branch Nets (Embedding), and Dual-Attention Nets with traditional ranking loss are slightly lower than the published results in original papers. It is because the procedure of data processing is not public, and the opened codes are partly consistent with the details in these papers. This situation does not affect the fair comparison of traditional ranking loss and our top-k ranking loss because these methods continue to be in the same experimental setting. The results indicate that our top-5 ranking loss consistently performs better than traditional ranking loss in any experimental settings, thus the top-k strategy effectively improves retrieval performance by mitigating the data ambiguity problem.

In addition, we conduct an experiment to evaluate how the selection of k in the top-k ranking loss impact on retrieval performance on three datasets. In Figure 6, we plot the R@K (K=1,5,10) performance curves as the increase of k in the image-query-sentence and sentence-query-image stages. The results indicate that the retrieval performance on R@K (K=1,5,10) improves with the increase of k; after reaching its maximum, the performance decreases gradually. For R@1 evaluation criteria, the optimal retrieval results are obtained constantly when k equals 3 or 4 over three datasets. In most cases, the R@5 metric gets the maximum while $k = 5$. When $k = 5$ or 6, the performance value of R@10 obtains the peak. The above mentioned results are reasonable to some extent. The larger value of hyper-parameter k brings the looser restriction, which is more beneficial for the R@K metric with larger K. In summary, a proper k, such as $k = 4, 5$, or 6, can achieve the desired performance for cross-modal retrieval with

TABLE IV: The retrieval performance with traditional ranking loss and top-k ranking loss in the same experimental setting.

Ranking-based Methods		Image-query-Sentence			Sentence-query-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
DSPE [30]	Traditional ranking loss	39.6	68.2	78.1	28.3	58.2	71.4
	Top-5 ranking loss	**41.8**	**70.9**	**81.6**	**32.1**	**62.3**	**74.2**
Two-branch Nets (Embedding) [31]	Traditional ranking loss	38.2	67.6	78.8	28.1	57.9	70.8
	Top-5 ranking loss	**41.2**	**70.2**	**80.9**	**32.5**	**61.7**	**73.4**
Dual-Attention Nets (VGG) [32]	Traditional ranking loss	40.3	72.1	80.1	30.0	59.8	70.9
	Top-5 ranking loss	**42.1**	**73.4**	**81.3**	**31.7**	**62.4**	**73.8**
Our Setting	Traditional ranking loss	39.2	68.1	76.8	30.9	58.4	69.4
	Top-5 ranking loss	**41.3**	**70.3**	**79.8**	**33.1**	**61.5**	**72.9**



(a) Flickr8k (b) Flickr30k (c) MSCOCO

Fig. 6: R@K curves with increasing of hyper-parameter k over three datasets.

indistinguishable data.

D. Impact of Embedding Dimension

The embedding dimension of the multi-modal space impacts the image-sentence matching results. In this section, with the condition of $k = 5$ in the top-k ranking loss function, we study the influence of the embedding dimension by assigning it to be tuned from 200 to 800 with a step-size of 50. In Figure 7, we plot the R@5 performance curves as the increase of embedding dimension on three datasets. It indicates that retrieval results improve with the increase of embedding dimension at first. After reaching the maximum, the performance decreases gradually. Note that the R@5 value fluctuates extensively when the embedding dimension is too small. The results demonstrate that a proper value of the embedding dimension can achieve the optimal retrieval results by capturing the common semantics of images and sentences. The optimal embedding dimension distinguishes over different datasets because of the diverse properties related to the datasets. When the embedding dimension is in the interval $[450, 550]$, the retrieval results can be satisfactory and relatively stable.

V. Conclusion

In this paper, we propose an end-to-end deep multi-modal network with a novel top-k ranking loss for large-scale image-sentence matching. This network learns the shared embedding space by combining the CNN and MLP networks. We propose a relatively slack top-k ranking loss to relieve the data ambiguity problem in massive query databases. An effective convex upper bound function is exploited to approximate the initial top-k loss in the framework of a multi-modal deep network. The experimental results over three benchmark datasets, *i.e.*, Flickr8k, Flickr30k, and MSCOCO, indicate that our method generally obtains the comparable performance with state-of-the-art methods in terms of the R@K metric. In future work, it is potential to replace the MLP text network as a LSTM network for extracting the text representations automatically.

Acknowledgment

This work is supported by National Key Research and Development Program of China (2016YFB1000903), National Natural Science Foundation of China (61532004, 61532015, 61672418, 61672419 and 61877050), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology, Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340, Australian Research Council Discovery Early Career Researcher Award (DE190100626), Defense Advanced Research Projects Agency (DARPA-BAA-HR0011-18-S-0044).

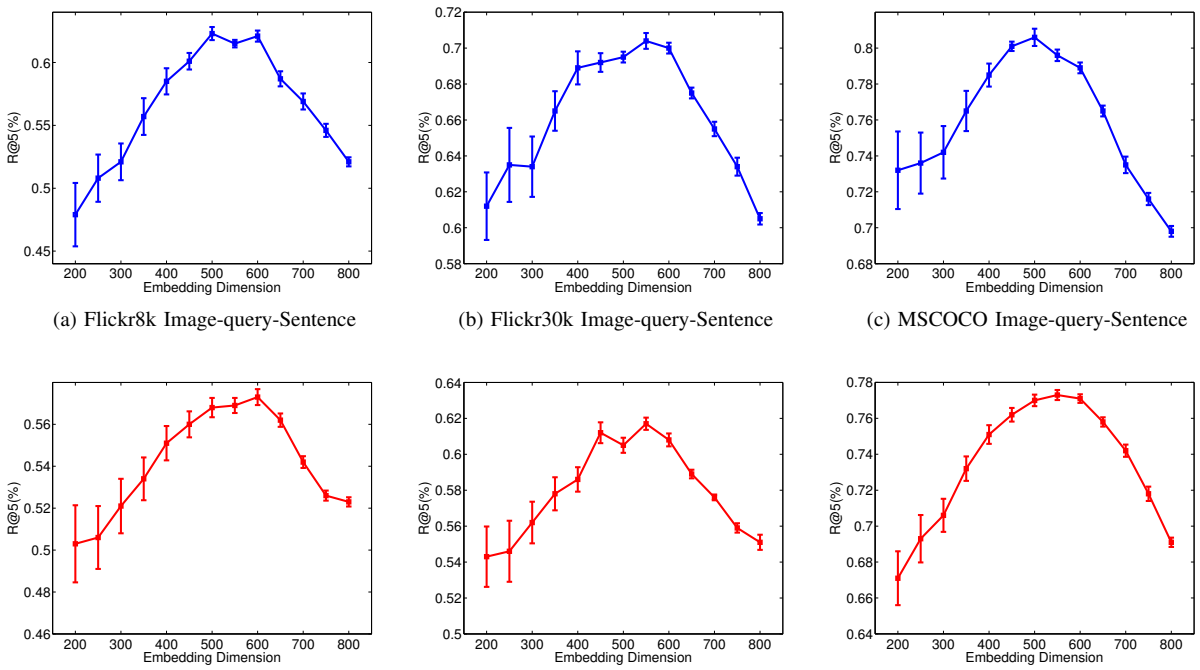
References

[1] M. Lux, M. Granitzer, W. Kienreich, V. Sabol, W. Klieber, and W. Sarka, "Cross media retrieval in knowledge discovery," *Practical Aspects of Knowledge Management*, pp. 343–352, 2004.

[2] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang, "A low rank structural large margin method for cross-modal ranking," in *SIGIR*, 2013, pp. 433–442.

[3] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[4] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multi-grained fusion by hierarchical network," *IEEE Transactions on Multimedia*, 2017.



(a) Flickr8k Image-query-Sentence (b) Flickr30k Image-query-Sentence (c) MSCOCO Image-query-Sentence

(d) Flickr8k Sentence-query-Image (e) Flickr30k Sentence-query-Image (f) MSCOCO Sentence-query-Image

Fig. 7: The influence of embedding dimension on retrieval results for three datasets.

[5] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1201–1216, 2016.

[6] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2017.

[7] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: A real-world dataset for weakly supervised cross-media retrieval," *IEEE Transactions on Multimedia*, 2017.

[8] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1220–1233, 2017.

[9] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[10] M. Luo, X. Chang, Y. Yang, L. Nie, A. G. Hauptmann, and Q. Zheng, "Simple to complex cross-modal learning to rank," *arXiv preprint arXiv:1702.01229*, 2017.

[11] Z. Ma, Y. Lu, and D. Foster, "Finding linear structure in large datasets with scalable canonical correlation analysis," in *ICML*, 2015, pp. 169–178.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[13] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," in *IJCNN*, 2000, p. 4614.

[14] X. Zhang, D. Chu, L.-Z. Liao, and M. K. Ng, "Sparse kernel canonical correlation analysis via ℓ_1-regularization," *arXiv preprint arXiv:1701.04207*, 2017.

[15] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[16] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[17] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[18] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *ECCV*, 2014, pp. 529–545.

[19] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation," *arXiv preprint arXiv:1411.7399*, 2014.

[20] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015, pp. 3441–3450.

[21] L. Montanarella, M. R. Bassani, and O. Bréas, "Chemometric classification of some european wines using pyrolysis mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 9, no. 15, pp. 1589–1593, 1995.

[22] F. Rudzicz, "Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics," in *ICASSP*, 2010, pp. 4198–4201.

[23] R. Arora and K. Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, 2013, pp. 7135–7139.

[24] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[25] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[26] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, vol. 11, 2011, pp. 2764–2770.

[27] Y. Verma and C. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," in *British Machine Vision Conference (BMVC)*, vol. 1, 2014, p. 2.

[28] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *NIPS*, 2014, pp. 1889–1897.

[29] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[30] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

[31] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *Pattern Analysis and Machine Intelligence*, 2018.

[32] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," *arXiv preprint arXiv:1611.00471*, 2016.

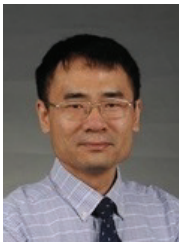
[33] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal lstm," in *CVPR*, vol. 2, no. 6, 2017, p. 7.

[34] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional

networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.
[35] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *ICML*, 2014, pp. 595–603.
[36] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *ICCV*, 2015, pp. 2623–2631.
[37] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *CVPR*, 2017.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[39] L. Deng, D. Yu *et al.*, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.
[40] L.-P. Liu, T. G. Dietterich, N. Li, and Z.-H. Zhou, "Transductive optimization of top-k precision," *arXiv preprint arXiv:1510.05976*, 2015.
[41] M. R. Gupta, S. Bengio, and J. Weston, "Training highly multiclass classifiers." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1461–1492, 2014.
[42] M. Lapin, M. Hein, and B. Schiele, "Top-k multiclass svm," in *NIPS*, 2015, pp. 325–333.
[43] ——, "Loss functions for top-k error: Analysis and insights," in *CVPR*, 2016, pp. 1468–1477.
[44] X. Chang, Y.-L. Yu, and Y. Yang, "Robust top-k multiclass svm for visual category recognition," in *KDD*. ACM, 2017, pp. 75–83.
[45] C. Yan, M. Luo, H. Liu, Z. Li, and Q. Zheng, "Top-k multi-class svm using multiple features," *Information Sciences*, 2017.
[46] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons-from back-propagation to adaptive learning algorithms," *Computer Standards & Interfaces*, vol. 16, no. 3, pp. 265–278, 1994.
[47] M. Lapin, M. Hein, and B. Schiele, "Top-k multiclass svm," in *NIPS*, 2015, pp. 325–333.
[48] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
[49] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in *ACM MM*. ACM, 2015, pp. 69–78.
[50] Z. Wang and S.-S. Tseng, "Knee point search using cascading top-k sorting with minimized time complexity," *The Scientific World Journal*, vol. 2013, 2013.
[51] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[53] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[55] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, and Q. Zheng, "Avoiding optimal mean robust pca/2dpca with non-greedy l1-norm maximization," in *IJCAI*, 2016, pp. 1802–1808.
[56] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Avoiding optimal mean 2, 1-norm maximization-based robust pca for reconstruction," *Neural computation*, vol. 29, no. 4, pp. 1124–1150, 2017.

Lingling Zhang is currently working toward the PhD degree in computer science at Xian Jiaotong University. She received the BS degree in Computing Science from Xian Jiaotong University in 2015. Her research interests include cross-media information mining and few-shot learning.

Minnan Luo received the Ph. D. degree from the Department of Computer Science and Technology, Tsinghua University in 2014. Currently, she is an Assistant Professor in the School of Electronic and Information Engineering at Xi'an Jiaotong University. She is also a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, USA. Her research interests include machine learning and optimization, video analysis, cross-media retrieval.



Jun Liu received the B.S, M.S and Ph.D degrees in computer science in 1995, 1998, and 2004 from Xi'an Jiaotong University, China. He is currently a professor in the School of Electronic and Information Engineering at Xi'an Jiaotong University. His research interests include text mining, data mining, Intelligent Network learning Environment and multimedia E-learning.

Xiaojun Chang was a Post-Doctoral Research Associate with the School of Computer Science, Carnegie Mellon University. He is currently a Faculty Member with the Faculty of Information Technology, Monash University, Clayton Campus, Australia. He is also affiliated with the Centre for Data Science, Monash University. He has spent most of his time working on exploring multiple signals (visual, acoustic, and textual) for automatic content analysis in unconstrained or surveillence videos. He has achieved top performance in various international competitions, such as TRECVID MED, TRECVID SIN, and TRECVID AVS. He was an ARC Discovery Early Career Researcher Award Fellow from 2019 to 2021.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University in 2010. He was a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, USA. He is currently an Associate Professor with the University of Technology Sydney, Australia. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval and video content understanding.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, USA, the degree in computer science from the Technische Universitat Berlin, Germany, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU),USA. He is currently with the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include natural language processing, speech understanding and synthesis and video analysis.