Multi-Networks Joint Learning for Large-Scale Cross-Modal Retrieval

Liang Zhang[1,3], Bingpeng Ma[1,2,3*], Guorong Li[1,3], Qingming Huang[1,2,3*], Qi Tian[4]
[1] University of Chinese Academy of Sciences, Beijing, 100049, China
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[3] Key Lab of Big Data Mining and Knowledge Management, CAS, Beijing, 100190, China
[4] Department of Computer Science, University of Texas at San Antonio, TX, 78249, USA
zhangliang14@mails.ucas.ac.cn,{bpma,liguorong,qmhuang}@ucas.ac.cn,qitian@cs.utsa.edu

ABSTRACT

This paper proposes a novel deep framework of multi-networks joint learning for large-scale cross-modal retrieval. For most existing cross-modal methods, the processes of training and testing don't care about the problem of memory requirement. Hence, they are generally implemented on small-scale data. Moreover, they take feature learning and latent space embedding as two separate steps which cannot generate specific features to accord with the cross-modal task. To alleviate the problems, we first disintegrate the multiplication and inverse of some big matrices, usually involved in existing methods, into that of many sub-matrices. Each sub-matrix is targeted to dispose one pair of image-sentence, for which we further design a novel sampling strategy to select the most representative samples to construct the cross-modal ranking loss and within-modal discriminant loss functions. By this way, the proposed model consumes less memory each time such that it can scale to large-scale data. Furthermore, we apply the proposed discriminative ranking loss to effectively unify two heterogenous networks, deep residual network for images and long short-term memory for sentences, into an end-to-end deep learning architecture. Finally, we can simultaneously achieve specific features adapting to cross-modal task and learn a shared latent space for images and sentences. Extensive evaluations on two large-scale cross-modal datasets show that the proposed method brings substantial improvements over other state-of-the-art ranking methods.

KEYWORDS

Multi-modal analysis; Cross-modal retrieval; Learning to rank; Deep feature representation

1 INTRODUCTION

Cross-modal retrieval has become a popular research topic in recent years due to the increasing prevalence of multi-modal data in search engines and social media. Subsequently, exploiting the correlation across different modalities is imperative to many practical applications. For example, when a web user types into a textual description, he usually expects to obtain a set of images that visually best illustrate it. Considering the fact that heterogeneous samples have different feature dimensions and distinct data distribution, a large number of methods focus on learning coupled transformations by which similarity search across different modalities is feasible. Generally, most of the traditional cross-modal methods mainly adopt four different techniques, *i.e.*, maximizing correlations [15, 17, 28–30, 32, 34], manifold learning [23, 25], learning to rank [5, 16, 40, 43], and labeling approximation [20, 38, 42].

The above methods have achieved decent performance on various small-scale datasets. However, the cross-modal datasets are becoming larger and larger in the big data era. Generally speaking, the large number of samples can help to learn discriminative transformations for more semantic association can be used to model the correlations between different modalities. However, the volume and dimension of multimedia data will be growing significantly on large-scale datasets. For most of existing methods, their optimizations usually involve multiplication, inverse and eigenvalue decomposition of some big matrices. Computational efficiency and memory space are big challenges for these operations. Hence, it is hard for these methods to be tested on the large-scale datasets.

Besides, traditional cross-modal methods usually take feature learning and latent space embedding as two separate steps. As a common practice, they first encode images and texts into vectors of hand-crafted features. And then they generate a low-dimensional joint embedding space where the heterogeneous similarity can be calculated. Although the optimal results are ensured in the respective stages, it may produce sub-optimal results for the cross-modal task because such visual and textual feature vectors may not be optimally compatible with the embedding process. Furthermore, feature extractions of different modalities are independent to each other such that the semantic correlation between heterogeneous features is neglected.

To alleviate these problems, we propose a novel deep framework named **M**ulti-**N**etworks jo**i**nt **L**earning (MNiL) for large-scale cross-modal retrieval. In this work, we take a step further to realize efficient retrieval of sentences in response to image query or vice versa, as shown in Figure 1. Concretely, we design a discriminative ranking loss function to easily meet the memory requirement when we deal with large-scale dataset. For the problem that the multiplication and inverse of some big matrices will consume huge memory space, we first disintegrate the training process of large-scale data into a set of sub-problems. Each time we dispose one pair of image-sentence. Furthermore, based on a novel sampling strategy, we can

*Corresponding author.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

DOI: https://doi.org/10.1145/3123266.3123317

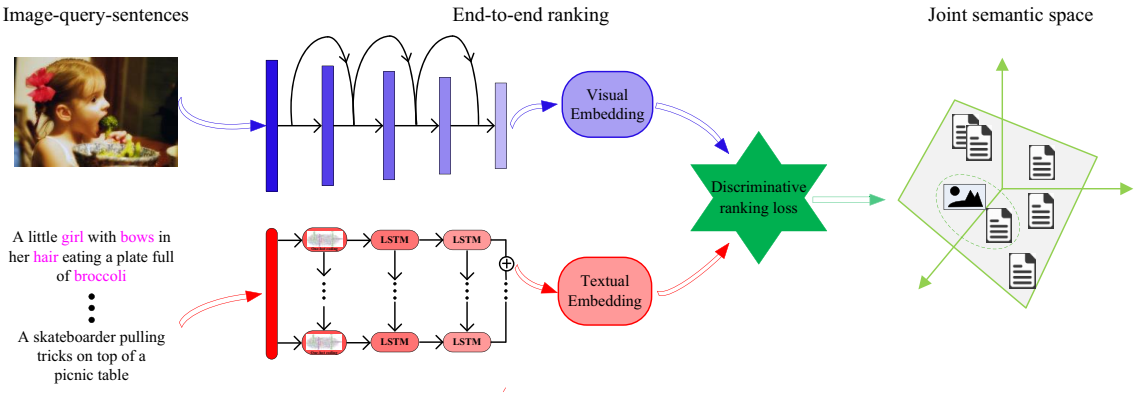


Figure 1: Flowchart of the proposed deep cross-modal architecture. Images and sentences are first encoded into heterogenous feature vectors. Then these vectors are embedded as low-dimensional representations via two different *feature-map layers*. After that, the discriminative ranking loss is intended to drive two heterogenous networks. Lastly, similar points are pushed together while dissimilar points are separated as far as possible in the learned shared latent space.

select the most representative samples to define the cross-modal ranking loss function and within-modal discriminant loss function for each image-sentence pair. By this way, we can utilize as few samples as possible to explore the semantic association between the different modalities.

Furthermore, to simultaneously obtain the task-specific features and a low-dimensional joint embedding space, we adopt the proposed ranking loss to effectively unify the training of two heterogenous networks, deep residual network (ResNet) [18] over images and long short-term memory (LSTM) [1] over sentences. Through an end-to-end training, on one hand, the learned low-dimensional joint embedding can guide the learning of the visual and textual features. On the other hand, the learned visual and textual features can also give feedback to learn the better embedding. Finally, the learned model can generate the discriminative semantic expressions for images and sentences and explore the high quality correlations between different modalities.

Extensive experiments have been performed on two large-scale cross-modal datasets: MSCOCO with 80K images paired with five sentences each [22] and Flickr30K with 30K images paired with five sentences each [41]. To the best knowledge of the authors, the two datasets contain the most image-sentence pairs in the cross-modal task till now. Experimental results demonstrate that the proposed method outperforms the state-of-the-art ranking methods on the large-scale image-to-sentence and sentence-to-image retrieval.

2 RELATED WORK

Taking the applied technique into account, prior cross-modal methods model the correlations between different modalities roughly from four different aspects.

One popular technique is to learn correlations between different modalities [3, 15, 17, 29, 30, 32, 34, 35]. The motivation of these methods is to learn a joint embedding space for images and texts by maximizing the correlations between the projected vectors of different modalities. An alternative technique is to rely on manifold learning [23, 25]. These methods assume that high dimensional data are embedded in a low dimensional intrinsic space. They achieve the joint embedding representation by projecting the different modalities into a common manifold by learning an underlying manifold. Another technique models heterogenous correlations via learning to rank. Single-directional ranking methods either project the images into the text space so it cannot be applied to the task of image query texts, *e.g.*, [12, 16], or learn a joint semantic space for texts and images but only consider the single-directional ranking instances, *e.g.*, [5, 19]. To project image and text spaces into a same semantic space, several bi-directional ranking methods have been developed in recent years [39, 40, 43]. These methods adopt the bi-directional training samples to learn a semantic model which integrates the merits of both directional retrieval such that the generalization performance is improved. Finally, to exploit valuable class information [20, 38, 42], some methods directly optimize the labeling approximation error between the given multi-labeled data and the class labels. Generally, the direct linkage between different modalities is their class labels. Thus, class information can be applied more likely to learn a discriminative latent space.

Besides, it's worth mentioning that many deep models have been widely developed to bridge the heterogeneous modalities and achieved promising performance in many applications, such as image captioning [2, 6, 10, 11, 37] and visual question answering [4, 13, 21, 24, 33]. These excellent works inspire us that it is possible to exploit the intrinsic semantic association between different modalities by training deep networks.

In summary, there are two limitations for most traditional methods. On one hand, they cannot generate the specific features which are suitable for the cross-modal task. On the other hand, they are difficult to scale to large and high-dimensional datasets because their derivation processes usually involve multiplication, inverse and eigenvalue decomposition of some big matrices. Considering this, we propose a novel discriminative ranking loss which can easily adapt to large-scale datasets. Furthermore, we seamlessly integrate two different networks, one for each modality. In the proposed model, feature learning and semantic embedding depend on each other, and each part gives feedback to the other part.

3 THE PROPOSED APPROACH

In this section, we first describe the process of generating features and then present the proposed loss function in detail. Finally, we introduce the training mechanism to optimize the deep framework. In Figure 1, we show MNiL's flowchart of sentences in response to image query or vice versa.

3.1 Generating features

In this paper, we deal with the matching problem between n pairs of images and sentences $\{I_i, S_i\}_{i=1}^{n}$. Each image-sentence pair is associated with one or more class labels. We adopt deep architectures to learn optimal features which are suitable to the cross-modal task. Specially, we utilize the architectures of ResNet [18] and LSTM [1] as our basic image and sentence framework.

ResNet for images: ResNet promotes neural networks to a very deeper architecture and gains the accuracy from the considerably increased depth. Extensive works have validated that ResNet can learn the most superior deep features for images [14, 18, 21]. Therefore, we apply ResNet to deeply exploit the semantic information implied in images. For ResNet, we directly use the raw image pixels I_i as their input.

LSTM for Sentences: We use a 2-layer LSTM to learn the deep representation of sentences [1, 9, 36]. Given new information, LSTM can learn when to forget previous hidden states and when to update hidden states by incorporating memory units. By this way, LSTM solves the vanishing and exploding gradients problem of recurrent neural networks. The structure of LSTM can be simply designed such that it can be directly implemented end-to-end training together with current deep networks for images. Moreover, LSTM can model sentences with varying lengths for they do not confine to the fixed length inputs or outputs. This merit facilitates the cross-modal datasets which contains a large number of sentences with varying lengths. As to LSTM, the input of sentence S_i is transformed into a sequential one-hot-vectors $\mathcal{S}_i =< \mathbf{s}_{i1}, \mathbf{s}_{i2}, \cdots, \mathbf{s}_{iT} >\in \mathbb{R}^{D_s \times T}$, where T denotes the number of words contained in sentence S_i. $\mathbf{s}_{it} \in \mathbb{R}^{D_s}$ is a one-hot vector denoting a word of time t in sentence S_i, and the nonzero entry of $\mathbf{s}_{it}$ represents the index of the word in the vocabulary of size D_s.

Through the above two deep networks, each input (I_i, S_i) is encoded into a pair of *intermediate features* $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ denotes D_x-dimensional feature vector of the ith image I_i, and $< \mathbf{y}_{i1}, \mathbf{y}_{i2}, \cdots, \mathbf{y}_{iT} >\in \mathbb{R}^{D_y \times T}$ is a set of sequential word features of the ith sentence $\mathcal{S}_i$. It needs to be emphasized that we take the mean of $< \mathbf{y}_{i1}, \mathbf{y}_{i2}, \cdots, \mathbf{y}_{iT} >$ to obtain $\mathbf{y}_i \in \mathbb{R}^{D_y \times 1}$ as the feature vector of sentence S_i.

3.2 Loss function

Note that $\mathbf{x}_i$ and $\mathbf{y}_i$ have different feature dimensions, it is difficult to directly calculate their similarity. Hence, we insert a *feature-map layer* to replace the fully-connected layer used in ResNet, and add one to the 2-layer LSTM too. The *feature-map layers* are regarded as linear transformations for the outputs of two different sub-networks. Then the *intermediate features* $(\mathbf{x}_i, \mathbf{y}_i)$ are delivered through the *feature-map layers* and mapped into K-dimensional embedding spaces $\mathcal{H}_x$ and $\mathcal{H}_y$, where bimodal samples (I_i, S_i) can be matched by similarity function $f(\mathbf{x}_i, \mathbf{y}_i)$ (dot product).

Given a training sample from one modality, web users usually expect relevant samples from another modality appearing at the top of the ranking list. Hence, it is important to explore the class information in the learning stage. Classic methods jointly use all inter-class and intra-class samples of all query samples to model the class information. However, they are very difficult to process large-scale datasets because their calculations involve the multiplication, inverse and eigenvalue decomposition of some large matrices. This will lead to heavy computational complexity and large memory space. In order to alleviate the problem, we disintegrate the task of preserving class information into a set of sub-modules, each of which only uses one pair of image-sentence to model the semantic association. Furthermore, for each pair, we propose a novel sampling strategy to select the most representative samples, which are used to construct a discriminative ranking loss for image-query-sentences and sentence-query-images. Specifically, each directional loss includes cross-modal ranking constraint, along with within-modal discriminant constraint which can be effectively realized and easily meet the memory requirement.

Cross-modal ranking constraint: For each image $\mathbf{x}_i$, we improve weighted approximate-rank pairwise loss (WARP) [19] to select the most representative sentences, which are assigned with higher discriminance to approximate the class information. Concretely, we express the pairwise ranking loss for the ith image $\mathbf{x}_i$ with enforced margin ρ as:

$$\begin{aligned} &\max(0, \mathcal{L}(\lfloor \frac{n-1}{v_x} \rfloor) \times (\rho + f(\mathbf{x}_i, \mathbf{y}_k) - f(\mathbf{x}_i, \mathbf{y}_j))) \\ &s.t.\ \forall \mathbf{y}_j \in \mathcal{Y}_i^+,\ \forall \mathbf{y}_k \in \mathcal{Y}_i^-,\ \rho < 1 \end{aligned} \tag{1}$$

where $\mathcal{Y}_i^+$ and $\mathcal{Y}_i^-$ represent the relevant and irrelevant sentence sets of image $\mathbf{x}_i$, respectively, and $\mathbf{y}_j$ has the same class labels with $\mathbf{x}_i$. v_x is the sampling number when we discover the first negative sentence $\mathbf{y}_k$ (called violator in this paper) satisfying $\rho + f(\mathbf{x}_i, \mathbf{y}_k) > f(\mathbf{x}_i, \mathbf{y}_j)$. $\lfloor \cdot \rfloor$ denotes the floor function. $\mathcal{L}(\cdot) : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ is the mapping function that transforms the rank (*i.e.*, similarity relation among multiple samples) into a loss:

$$\mathcal{L}(k) = \sum_{i=1}^{k} \alpha_i, \alpha_1 > \alpha_2 \cdots \geq 0 \tag{2}$$

where $\alpha_i = 1/i$, which has shown good precision@k performance in image retrieval and cross-modal retrieval [7, 43].

It shall be noted that we set the margin punishment ρ less than 1 unlike WARP, in which ρ is equal to 1. Generally, the similarity of dot product between two samples is no bigger than 1. When $\rho = 1$, the condition of $1 + f(\mathbf{x}_i, \mathbf{y}_k) > f(\mathbf{x}_i, \mathbf{y}_j)$ can be easily held such that most of negative sentences can satisfy this condition. In this case, each image only requires one random sampling to find the violator so it cannot give consideration to the other samples. Hence, the selected violator is not the representative sample for constructing the ranking loss. This will degenerate the ranking loss into a pairwise classification.

Similarly, given a sentence y_i, we have:

$$\begin{aligned} &\max(0, \mathcal{L}(\lfloor \frac{n-1}{v_y} \rfloor) \times (\rho + f(\mathbf{x}_k, \mathbf{y}_i) - f(\mathbf{x}_j, \mathbf{y}_i))) \\ &s.t.\ \forall \mathbf{x}_j \in \mathcal{X}_i^+, \forall \mathbf{x}_k \in \mathcal{X}_i^- \end{aligned} \tag{3}$$

where $\mathcal{X}_i^+$ and $\mathcal{X}_i^-$ represent the relevant and irrelevant image sets of the sentence $\mathbf{y}_i$, respectively. υ_y is the sampling number when we find the first negative image $\mathbf{x}_k$ satisfying $\rho + f(\mathbf{x}_k, \mathbf{y}_i) > f(\mathbf{x}_j, \mathbf{y}_i)$.

Within-modal discriminant constraint: The ranking constraint effectively characterizes the discriminance for cross-modal samples, while ignores the discriminance among samples within the same modality. In fact, each image $\mathbf{x}_i$ usually expresses the same semantic meaning with many other images $N(\mathbf{x}_i)$, which are called neighbors of $\mathbf{x}_i$ in the original space. Undoubtedly, an optimal embedding should ensure the neighboring points in the original space close to each other in the embedded space. Likewise, this discriminant information will make more samples constructing semantic relationship with each other. Moreover, since more samples participate in modeling the semantic correlations each time, the proposed model will obtain the optimal results through fewer number of samplings.

Note that the similarities between samples from the same modality are usually much larger than those of samples from the different modalities. Hence, different from [39], we set τ greater than ρ to enlarge the distances among the different classes, by which each sample effectively preserves the discriminance. With these ideas, we enforce a margin of τ between $N(\mathbf{x}_i)$ and any other non-neighboring points:

$$\max(0, \tau + f(\mathbf{x}_i, \mathbf{x}_k) - f(\mathbf{x}_i, \mathbf{x}_j)) \; \forall \mathbf{x}_j \in N(\mathbf{x}_i), \forall \mathbf{x}_k \notin N(\mathbf{x}_i) \quad (4)$$

Analogously to the image modality, we constrain each sentence $\mathbf{y}_i$ with the margin τ as:

$$\max(0, \tau + f(\mathbf{y}_i, \mathbf{y}_k) - f(\mathbf{y}_i, \mathbf{y}_j)) \; \forall \mathbf{y}_j \in N(\mathbf{y}_i), \forall \mathbf{x}_k \notin N(\mathbf{y}_i) \quad (5)$$

where $N(\mathbf{y}_i)$ consists of the sentences describing the semantic content with $\mathbf{y}_i$, and $\mathbf{y}_k$ belongs to different classes with $\mathbf{y}_i$.

Sextuple sampling: For any one of image-sentence pair $(\mathbf{x}_i, \mathbf{y}_i)$, the proposed single-directional ranking loss involves five elements, which consist of a target sample, a positive and a negative matches from another modality as well as a positive and a negative matches within the same modality. Hence, bi-directional ranking loss needs to optimize ten samples each time, which is computationally infeasible over all such combinations of ten samples like [39]. To reduce the sampling number and improve usage of selected samples for each image-sentence pair $(\mathbf{x}_i, \mathbf{y}_i)$, we search a cross-modal positive sample $\mathbf{y}_j$ and a violator $\mathbf{y}_k$ for $\mathbf{x}_i$, and then $(\mathbf{y}_j, \mathbf{y}_k)$ directly serves as the within-modal positive and negative samples for $\mathbf{y}_i$. Similarly, $\mathbf{y}_i$ seeks a cross-modal positive sample $\mathbf{x}_j$ and a violator $\mathbf{x}_k$, which are used as the within-modal positive and negative samples for $\mathbf{x}_i$. In this way, the selected sextuple can approximate the result as the ten samples for the most representative samples are left while some redundant ones are removed.

Figure 2 provides an intuitive illustration of the sextuple ranking mechanism for an image-sentence pair. Different shapes denote different modalities (*i.e.*, images and sentences). The same color indicates the relevant semantics. The longer the length of a line segment, the more dissimilar of two instances. The black line segments are enforced with the fixed values ρ or τ. The dotted lines and dashed lines represent the distances which are adjustable by our objective function. Before ranking, the red circle with pentagram (a query sample) is close to the blue square (violator) and the green circle (within-modal negative sample) while is far away from shapes with the same color (positive instances). By enforcing margins for both ranking constraint and discriminant constraint, we force the query instance close to its within-class instances while far away from the violator with a distance larger than ρ and from the within-modal negative instance with a distance larger than τ. Similarly, the red square with pentagram can also achieve the discriminative result.

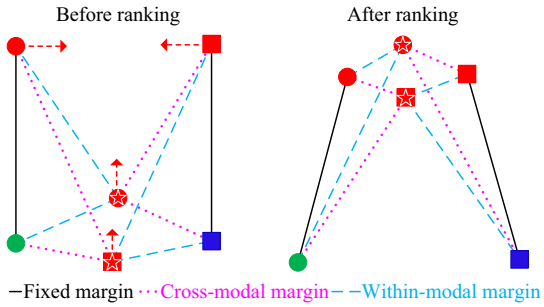


Figure 2: Demonstration on the working mechanism of the proposed discriminative ranking loss. Different shapes denote different modalities (*i.e.*, images and sentences). The same color indicates the relevant semantics.

3.3 Joint training

MNiL is a hybrid deep architecture that consists of ResNet and LSTM for learning the discriminative ranking with the accurate semantic expression. It is very difficult to directly train ResNet and LSTM together since their network structures and parameters settings are quite different. To solve this problem, we design a novel combination of ResNet and LSTM via the sextuple ranking loss. By this way, our architecture effectively unifies the joint multi-modal embedding with cross-modal ranking which enable to learn the discriminative semantic representation for images and sentences.

Concretely, we jointly train the two *feature-map layers* and fine-tune the two sub-networks by an end-to-end mechanism. Meanwhile,(1) we guarantee the precision@k by minimizing the cross-modal ranking losses; (2) we enable the robust discriminant by minimizing the within-modal hinge losses. We integrate these loss functions into a joint optimization problem, which is taken over the multi-networks and formulated as follows:

$$\begin{aligned} O(X, Y) = & \sum_{i,j,k} \max(0, \mathcal{L}(\lfloor \frac{n-1}{\upsilon_x} \rfloor) \times (\rho + f(\mathbf{x}_i, \mathbf{y}_k) - f(\mathbf{x}_i, \mathbf{y}_j))) \\ & + \sum_{i,j,k} \max(0, \mathcal{L}(\lfloor \frac{n-1}{\upsilon_y} \rfloor) \times (\rho + f(\mathbf{x}_k, \mathbf{y}_i) - f(\mathbf{x}_j, \mathbf{y}_i))) \\ & + \beta_1 \sum_{i,j,k} \max(0, \tau + f(\mathbf{x}_i, \mathbf{x}_k) - f(\mathbf{x}_i, \mathbf{x}_j)) \\ & + \beta_2 \sum_{i,j,k} \max(0, \tau + f(\mathbf{y}_i, \mathbf{y}_k) - f(\mathbf{y}_i, \mathbf{y}_j)) \end{aligned} \quad (6)$$

where the sum is over all sextuples defined in Eq. (1), (3), (4) and (5), respectively. The cross-modal margin ρ and within-modal margin τ could be different for the different forms of similarity or even different samples. But to make it easy to train, we fix ρ (0.3 in experiments) for all the cross-modal training samples and τ (0.5 in

experiments) for all the within-modal training samples. β_1 and β_2 are defined to control the importance of the discriminant terms, which can also be acted as regularizers for the bi-directional retrieval tasks.

For clarification, we give the back-propagation for one sextuple, and the remaining sextuples have the similar optimization procedure. The gradients in the back-propagation of the sextuple ranking loss are computed as:

$$\begin{aligned}
\frac{\partial O(X,Y)}{\partial \mathbf{p}} =& \mathcal{L}(\lfloor \frac{n-1}{v_x} \rfloor) \times (\mathbf{t}^- - \mathbf{t}^+) \times I_{f(\mathbf{x}_i,\mathbf{y}_k)-f(\mathbf{x}_i,\mathbf{y}_j)+\rho>0} \\
&+ \beta_1(\mathbf{p}^- - \mathbf{p}^+) \times I_{f(\mathbf{x}_i,\mathbf{x}_k)-f(\mathbf{x}_i,\mathbf{x}_j)+\tau>0} \\
\frac{\partial O(X,Y)}{\partial \mathbf{t}^+} =& -\mathcal{L}(\lfloor \frac{n-1}{v_x} \rfloor) \times \mathbf{p} \times I_{f(\mathbf{x}_i,\mathbf{y}_k)-f(\mathbf{x}_i,\mathbf{y}_j)+\rho>0} \\
&- \beta_2 \mathbf{t} \times I_{f(\mathbf{y}_i,\mathbf{y}_k)-f(\mathbf{y}_i,\mathbf{y}_j)+\tau>0} \\
\frac{\partial O(X,Y)}{\partial \mathbf{t}^-} =& \mathcal{L}(\lfloor \frac{n-1}{v_x} \rfloor) \times \mathbf{p} \times I_{f(\mathbf{x}_i,\mathbf{y}_k)-f(\mathbf{x}_i,\mathbf{y}_j)+\rho>0} \\
&+ \beta_2 \mathbf{t} \times I_{f(\mathbf{y}_i,\mathbf{y}_k)-f(\mathbf{y}_i,\mathbf{y}_j)+\tau>0}
\end{aligned} \tag{7}$$

where $\mathbf{p}$, $\mathbf{t}^+$ and $\mathbf{t}^-$ are the low-dimensional embeddings of $\mathbf{x}_i$, $\mathbf{y}_j$ and $\mathbf{y}_k$, and the embedding is realized by the *feature-map layers* from the proposed network. The indicator function $I_{condition} = 1$ if *condition* is true; otherwise $I_{condition} = 0$.

$$\begin{aligned}
\frac{\partial O(X,Y)}{\partial \mathbf{t}} =& \mathcal{L}(\lfloor \frac{n-1}{v_y} \rfloor) \times (\mathbf{p}^- - \mathbf{p}^+) \times I_{f(\mathbf{x}_k,\mathbf{y}_i)-f(\mathbf{x}_j,\mathbf{y}_i)+\rho>0} \\
&+ \beta_2(\mathbf{t}^- - t^+) \times I_{f(\mathbf{y}_i,\mathbf{y}_k)-f(\mathbf{y}_i,\mathbf{y}_j)+\tau>0} \\
\frac{\partial O(X,Y)}{\partial \mathbf{p}^+} =& -\mathcal{L}(\lfloor \frac{n-1}{v_y} \rfloor) \times \mathbf{t} \times I_{f(\mathbf{x}_k,\mathbf{y}_i)-f(\mathbf{x}_j,\mathbf{y}_i)+\rho>0} \\
&- \beta_1 \mathbf{p} \times I_{f(\mathbf{x}_i,\mathbf{x}_k)-f(\mathbf{x}_i,\mathbf{x}_j)+\tau>0} \\
\frac{\partial O(X,Y)}{\partial \mathbf{p}^-} =& \mathcal{L}(\lfloor \frac{n-1}{v_y} \rfloor) \times \mathbf{t} \times I_{f(\mathbf{x}_k,\mathbf{y}_i)-f(\mathbf{x}_j,\mathbf{y}_i)+\rho>0} \\
&+ \beta_1 \mathbf{p} \times I_{f(\mathbf{x}_i,\mathbf{x}_k)-f(\mathbf{x}_i,\mathbf{x}_j)+\tau>0}
\end{aligned} \tag{8}$$

where $\mathbf{t}$, $\mathbf{p}^+$ and $\mathbf{p}^-$ are the low-dimensional embeddings of $\mathbf{y}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$, respectively.

With this modified discriminative ranking loss function, the inputs to the proposed deep architecture are sextuples of images and sentences, *i.e.*, $\{(I_i, S_i, I_i, I_k, S_j, S_k)\}_{i=1}^n$, in which I_i is more similar to S_j than S_k while S_i is more similar to S_j than S_k. As shown in Figure 1, we propose to use a shared sub-network to automatically learn a unified representation for the input images and sentences. Through this sub-network, an input sextuple $(I_i, S_i, I_i, I_k, S_j, S_k)$ is encoded to a sextuple of *intermediate features* $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{y}_j, \mathbf{y}_k)$. In this sub-network, ResNet is shared by the three images and the 2-layer LSTM is shared by the three sentences for each input sextuple. Such a way of parameter sharing can significantly reduce the number of parameters in the whole architecture. Specially, the sextuple loss can be optimized efficiently through the standard back-propagation and easily meet with the memory requirement.

4 EXPERIMENT

In this section, we conduct extensive experiments to evaluate the efficacy of the proposed method on the tasks of image-query-sentences and sentence-query-images. We compare different methods on two benchmark datasets: MSCOCO [22] and Flickr30K [41].

4.1 Experimental setting

Datasets: After pruning images without category information, MSCOCO consists of 82,081 training images and 5,000 testing images, each of which is associated with five sentences. We randomly select 82,081 image-sentence pairs as training set, 5,000 pairs as validation set and 5,000 pairs as query set. The ground truth labels come from 81 most frequent categories. Flickr30K contains 31,783 images which mainly focus on depicting people and animals. For each image, there are 5 captions vividly describing it. Therefore, this dataset has total 158,915 captions, and 158,915 image-caption pairs can be generated for the task of training, validation and testing. We randomly select 29,783 pairs as training set, 1,000 pairs as validation set and 1,000 pairs as query set.

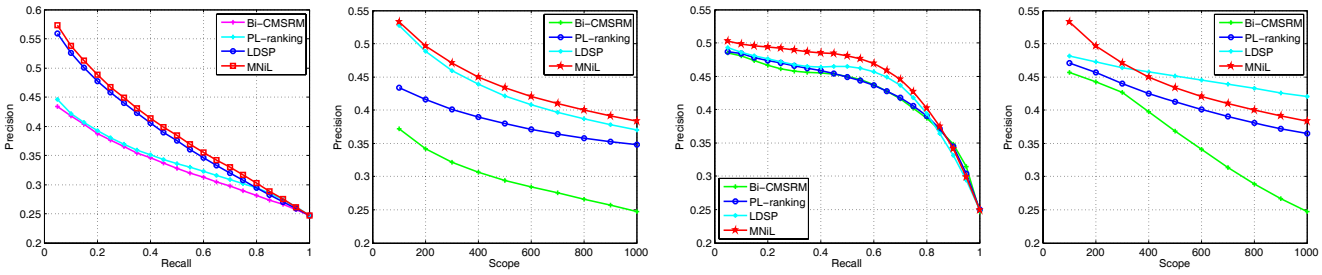
Compared methods: We compare MNiL with three state-of-the-art cross-modal ranking methods, including two shallow ranking methods: Bi-CMSRM [40] and PL-ranking [43], and one deep ranking method: LDSP [39]. We also try to run the codes of GMA [34], LCFS [38], LGCFL [20] and ml-CCA [29]. But our device cannot meet their memory requirements, thus we do not report their results. Besides, We also report results of MNiL ($I \rightarrow T$) and MNiL ($T \rightarrow I$), each of which only adopts single-directional ranking of MNiL. MNiL ($I \rightarrow T$) learns the joint embedding space for images and sentences by only ranking sentences for image queries, and MNiL ($T \rightarrow I$) only ranking images for sentence queries.

Implementation details: On MSCOCO dataset, two heterogenous samples are considered similar (dissimilar) if they share at least one (none) semantic label. Note that on Flickr30K dataset, we do not have direct ground truth labels for images. Hence, each image-sentence pair is regarded as a category. When we construct the within-modal discriminant loss, $N(\mathbf{x}_i)$ of each image $\mathbf{x}_i$ is set to the image itself. But the neighborhood of each sentence $N(\mathbf{y}_i)$ has four members for each image is paired with five captions.

For MNiL, we directly utilize the raw pixels of image as the input of ResNet. As for sentences, sequential words of each sentence are first transformed into one-hot vectors, which are the input of LSTM. For all compared methods, we use the ResNet [18] to extract a 2048-dimensional feature vector for each image. Each sentence adopts LSTM module [22] pre-trained on the MSCOCO dataset to achieve word vectors, and then we compute a mean vector of these word vectors as its feature vector.

We implement the proposed deep architecture in Torch framework. For training network, we employ the 50-layer ResNet [18] and the 2-layer LSTM [1]. Back-propagation is applied to fine-tune the ResNet and the 2-layer LSTM and train the new *feature-map layers*. Since the *feature-map layer* is trained from scratch, we set its learning rate to be 10 times that of the lower layers. We use the mini-batch stochastic gradient descent (SGD) with 0.9 momentum. We adopt 20 and 15 as the maximum number of words in each sentence for MSCOCO and Flickr30K, respectively.

Performance evaluation: We follow the same protocols as other recent works [31, 32, 39, 40]. On MSCOCO dataset, the mean average precision (MAP) [32] is used to evaluate the performance. Especially, we adopt MAP@R [40] to measure the retrieval performance at the fixed number of retrieved samples. R is set to 50 for the top 50 retrieved samples and to "*all*" for all retrieved samples. Besides, the precision-recall curve [32] and scope-precision



(a) COCO Sentence-query-images (b) COCO Image-query-sentences

Figure 3: Precision-recall curves and precision-scope curves for the image-query-sentences and sentence-query-images experiments on MSCOCO dataset.

Table 1: Performance comparison in terms of MAP@R scores on MSCOCO dataset.

Methods \ Tasks	R = 50			R = *all*		
	Sentence query	Image query	Average MAP	Sentence query	Image query	Average MAP
Bi-CMSRM [40]	0.3687	0.3653	0.3670	0.3458	0.3508	0.3483
PL-ranking [43]	0.4320	0.4020	0.4170	0.3369	0.3414	0.3392
LDSP [39]	0.5009	0.4430	0.4720	0.3471	0.3536	0.3504
MNiL ($I \rightarrow T$)	0.3654	0.3826	0.3740	0.3501	0.3617	0.3559
MNiL ($T \rightarrow I$)	0.4172	0.3643	0.3908	0.3521	0.3479	0.3500
MNiL	**0.5341**	**0.4642**	**0.4992**	**0.3563**	**0.3875**	**0.3719**

curve [31] are also displayed for all methods. The scope is specified by the number of top-ranked samples when the retrieved samples are ranked according to the similarities between them and the query. On Flickr30K dataset, consistent with [39], we report the Recall@K ($K = 1, 5, 10$), *i.e.*, the percentage of queries for which at least one correct ground truth match is ranked among the top K matches.

For all methods, we adopt 5-fold cross validation process to determine the values of parameters. The proposed method utilizes the following parameter setting: $\beta_1 = 0.1$ and $\beta_2 = 0.2$. For fair comparison, we conduct experiments 10 times by randomly selecting training/validation/testing combinations, and show the average performance for all methods [1].

4.2 Image-sentence retrieval

Table 1 and 2 report the MAP@R and Recall@K of the different methods on the test set of MSCOCO and Flickr30K, respectively. From these tables, we draw the following conclusions:

First, MNiL outperforms MNiL ($I \rightarrow T$) and MNiL ($T \rightarrow I$) by 4.5% and 6.3% on average MAP scores, respectively. One-directional ranking cannot explore the latent structure of the retrieved modality. Conversely, unifying bi-directional rankings can project texts and sentences into the same semantic space so images and sentences can easily construct semantic relation.

Second, the deep learning methods improves the performance of traditional methods. For example, the average MAP score of MNiL is improved by 13.2% compared to that of PL-ranking on R=50. This is because traditional methods design their loss functions focusing on processing small-scale datasets. When they deal with large-scale data, they very likely fall into the over-fitting problem. MNiL develops a sampling strategy to define its objective function which can adapt to large-scale data. Therefore, it can exploit a large amount of semantic association to improve retrieval performance.

Third, the performance of MNiL is superior to another deep method LDSP. For example, the average MAP score of MNiL is improved by 6.1% on R=*all*. LDSP learns the latent semantic space by using hand-crafted features which are usually not suitable for cross-modal task. However, MNiL can learn the suitable feature representation according to its objective function, thus it captures the more effective cross-modal correlations than LDSP.

Moreover, LDSP needs to optimize ten samples for each image-sentence pair, and these samples are selected only considering the pairwise relations of inter-class and intra-class. But MNiL makes use of listwise relations of the ranking list to select six most representative samples. The experimental results validate the selected samples can estimate the result as ten samples in LDSP.

Finally, MNiL achieves the best results on Recall@1, Recall@5 and Recall@10 on Flickr30K. In within-modal constraint, we use the sample itself as its neighborhood. These results prove that this operation is reasonable. Since the similarity of sample to itself is generally higher than the sample with its positive sample, the hinge loss function will punish the negative sample severely.

The precision-recall and scope-precision curves on both directional retrieval are shown in Figure 3. The scope (*i.e.*, the top K retrieved samples) of precision-scope curve varies from 100 to 1000.

[1] https://github.com/liangzhang1407/Multi-Networks-Joint-Learning-for-Large-Scale-Cross-Modal-Retrieval

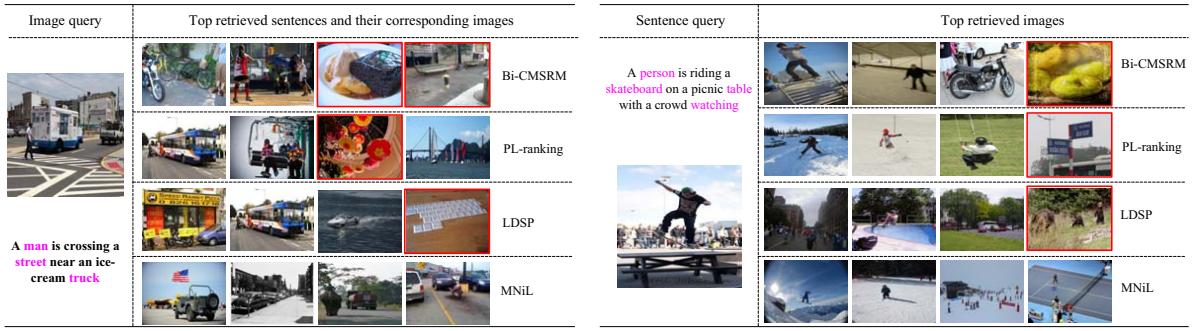


Figure 4: Two examples of image-query-sentences (in left half) and sentence-query-images (in right half) on MSCOCO dataset. For sentence-query-images, we show the query sentence with its corresponding images. The query sentence describes the semantic about"skateboard", "bench", and "person". For image-query-sentences, the top retrieved sentences are shown with their corresponding images. The query image depicts the scene about "trafficlight", "car", "person", and "truck". The incorrect retrieved results are shown in the red frame.

Table 2: Performance comparison in terms of Recall@K scores on Flickr30K dataset.

Methods \ Tasks	Image-query-sentences			Sentence-query-images		
	R@1	R@5	R@10	R@1	R@5	R@10
Bi-CMSRM [40]	0.263	0.509	0.602	0.197	0.459	0.485
PL-ranking [43]	0.299	0.554	0.621	0.214	0.470	0.548
LDSP [39]	0.357	0.629	0.744	0.251	0.539	0.665
MNiL ($I \rightarrow T$)	0.293	0.512	0.598	0.211	0.483	0.491
MNiL ($T \rightarrow I$)	0.229	0.449	0.516	0.237	0.474	0.525
MNiL	**0.369**	**0.641**	**0.769**	**0.273**	**0.557**	**0.670**

We observe that compared with the other methods, our method achieves the better results on all the curves. Hence, the curves further validate the superiority of MNiL for cross-modal retrieval.

To provide the intuitive judgement of the retrieval results, we give two retrieved instances of image-query-sentences and sentence-query-images in Figure 4. For each instance, the query and its paired samples are shown at the left, and the top four retrieved results are shown at columns 2-5. Note that MNiL finds the most relevant matches at semantic level, which is reflected by class labels. For image-query-sentences direction, we find that the top retrieved sentences of the proposed method are clearly relevant to the query images belonging to multiple classes "car", "truck", "person" and "trafficlight". For the sentence-query-images direction, given textual description about "skateboard", "bench" "person", the top retrieved images of MNiL are also relevant to the query sentence. However, the other methods produces some irrelevant results for both directional retrieval. It clearly validates that the proposed method can retrieve more relevant results comparing with the other methods.

4.3 Low-dimensional embedding

In this section, we analyze the discriminative ability of low-dimensional feature representation learned by different ranking methods. Based on MSCOCO, we construct a toy dataset using 500 image-sentence pairs from 'bicycle' class and another 500 paired samples from 'airplane' class. In Figure 5, we adopt the t-SNE [8] algorithm to project the *intermediate features* from the two different sub-networks and the embedded features of the different methods into a two-dimensional visualization spaces. The first column illustrates that the two-dimensional distributions of intermediate image and sentence features are mixed. Columns 2-5 show the embedded features of different methods. The red circles denote the distribution of the 'bicycle' class, and green circles represents the 'airplane' class.

From Figure 5, we first conclude that the proposed MNiL simultaneously unifies the same-class samples and separates the different classes for both directional retrieval, but the second best result (*i.e.*, LDSP) only unifies the same-class samples and separates the different classes for sentence query. Moreover, both image and sentence distributions of MNiL are in the same coordinate range, while coordinate ranges of other methods are different. These results validate that MNiL ensures the consistent structures between image and sentence spaces such that the low-dimensional embedding is enhanced with stronger discrimination. Finally, the separation between different classes of sentences is more obvious than that of images. In nature, class labels can be regarded as a special case of textual features because they are also language descriptions like sentences. Therefore, this observation is reasonable.

4.4 Quality of word vectors

To give more insight into the quality of learned word vectors, we provide empirical analysis by showing the nearest neighbours for

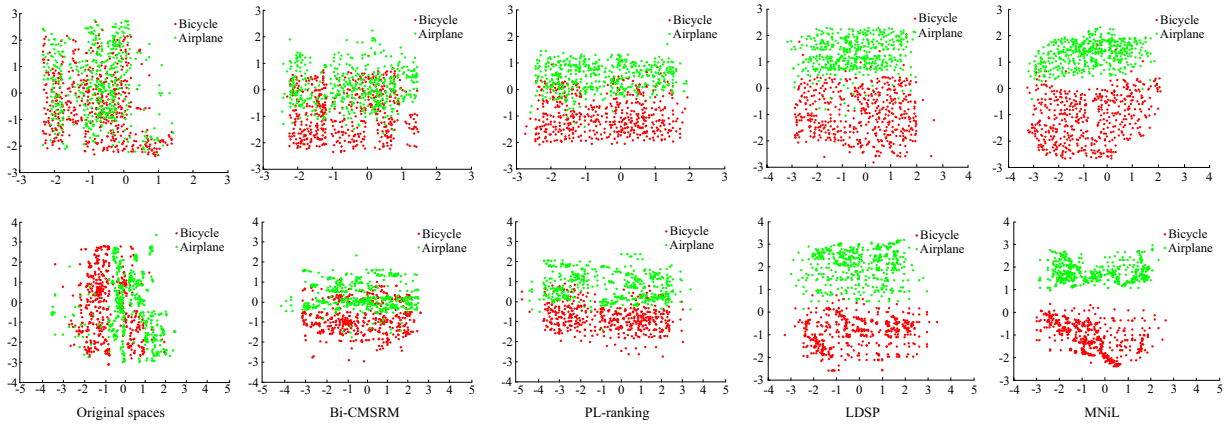


Figure 5: Low-dimensional embeddings of images and sentences from 'bicycle' and 'airplane' classes on MSCOCO dataset. The top row shows the embedding for images, and the bottom shows the embedding for sentences.

Table 3: Examples of the closest entities on the given words.

Given word	The most similar words								
art	museum	artwork	paintings	portrait	music	story	article	exhibit	gallery
sea	lake	water	river	ocean	seashore	shore	boats	rowboat	horizon
sport	players	lacrosse	cricket	athletes	rugby	football	teams	volleyball	match
music	song	jazz	musicians	band	drummer	guitars	singers	concert	spotlight
people	persons	women	men	adults	children	kids	teenagers	guys	individuals
animal	goat	alligator	pony	rabbit	frog	snail	duck	deer	pet
bicycle	bike	motorbike	bicyclist	biker	scooter	wheels	tricycle	jockey	handlebars
mountain	mountainside	mountaintop	ridge	hill	landscape	valley	hillside	hiker	climbers

each given word in Table 3. The proposed deep architecture adopts an end-to-end training mode, the outputs of second layer of LSTM are word vectors. Then we adopt cosine distance to measure the similarity between words, similar to *word2vec* model [26, 27].

From Table 3, we observe that the closest entities of each given word are clearly relevant to each other. For example, when we input the word "music", some relevant words like "song", "jazz" and "musicians" appear at the top positions of the ranked list. We think that this merit benefit from the proposed end-to-end deep training framework, which coherently combines ResNet with LSTM in an unified learning framework. It is well known that ResNet achieves the most effective deep image features [21] and LSTM explicitly takes the temporal structure starting from words of a sentence into account. By unifying the two deep architectures, the semantic information contained in images and sentences will be instructed for each other. Therefore, the word vectors will be updated toward to a more informative direction such that relevant words are more likely to be close to each other in the vector space.

5 CONCLUSION

In this paper, we propose a novel deep framework of multi-networks joint learning for large-scale cross-modal retrieval. It aims to match data from different modalities and alleviate two basic problems existing in the era of big data: scaling to large-scale data and generating task-specific features. We first design a sampling strategy to select the six most representative samples to define the cross-modal ranking loss and within-modal discriminant loss each time. Optimizing the sextuple requires less memory space so that it can easily adapt to large-scale data. Then, we apply the discriminative ranking loss to drive two heterogenous networks, ResNet for images and LSTM for sentences, by which we can simultaneously obtain task-specific features and discriminative embeddings. Extensive evaluations on two large-scale cross-modal datasets show that the proposed deep discriminative ranking model brings substantial improvements over other state-of-the-art ranking methods. Our future work will focus on improving network structure such that it can deal with more practical problems, such as image captioning and visual question and answer.

6 ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 61332016, 61620106009, 61572465, 61429201, U1636214, 61650202 and 61303153, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013. This work was also supported in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

REFERENCES

[1] K. Andrej, J. Justin, and F. Li. 2016. Visualizing and Understanding Recurrent Networks. In *International Conference on Learning Representations workshop*.

[2] K. Andrej and F. Li. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. 2013. Deep Canonical Correlation Analysis. In *International Conference on Machine Learning*.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, and D. Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*.

[5] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamsda, Y. Qi, O. Chapelle, and K. Weinberger. 2010. Learning to Rank with (a lot of) Word Features. *Information Retrieval* 13, 3 (2010), 291–314.

[6] X. Chen and C. Zitnick. 2015. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] L. Daryl and L. Gert. 2014. Efficient Learning of Mahalanobis Metrics for Ranking. In *International Conference on Machine Learning*.

[8] L. Van der Maaten and G. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[9] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[10] J. Donahue, Y. Jia, O. Vinyals, N. Zhang J. Hoffman, E. Tzeng, and T. Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning*.

[11] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P.Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig. 2015. From Captions to Visual Concepts and Back. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[12] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems*.

[13] H. Gao, J. Mao, J. Zhou, Z. Huang, and A. Yuille. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question. In *Advances in Neural Information Processing Systems*.

[14] W. Ge and Y. Yu. 2017. Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[15] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. 2014. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233.

[16] D. Grangier and S. Bengio. 2008. A Discriminative Kernel-based Approach to Rank Images from Text Queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 8 (2008), 1371–1384.

[17] D. Hardoon, S. Szedmark, and J. Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation* 16, 12 (2004), 2639–2664.

[18] K. He, X. Zhang, S. Ren, and J. sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[19] W. Jason, B. Samy, and U. Nicolas. 2010. Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. *Machine Learning* 81 (2010), 21–35.

[20] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. 2015. Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval. *IEEE Transactions on Multimedia* 17, 3 (2015), 370–381.

[21] J. Kim, S. lee, D. Kwak, and M. Heo. 2016. Multimodal Residual Learning for visual QA. In *Advances in Neural Information Processing Systems*.

[22] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

[23] V. Mahadevan, C. Wong, J. Pereira, T. Liu, N. Vasconcelos, and L. Saul. 2011. Maximum Covariance Unfolding: Manifold Learning for Bimodal Data. In *Advances in Neural Information Processing Systems*.

[24] M. Malinowski, M. Rohrbach, and M. Fritz. 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *IEEE International Conference on Computer Vision*.

[25] X. Mao, B. Lin, D. Cai, X. He, and J. Pei. 2013. Parallel Field Alignment for Cross Media Retrieval. In *ACM International Conference on MultiMedia*.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations Workshop*.

[27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*.

[28] J. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. 2014. On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 521–535.

[29] V. Ranjan, N. Rasiwasia, and C. Jawahar. 2015. Multi-Label Cross-modal Retrieval. In *IEEE International Conference on Computer Vision*.

[30] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal. 2014. Cluster Canonical Correlation Analysis. In *International Conference on Artificial Intelligence and Statistics*.

[31] N. Rasiwasia, P. Moreno, and N. Vascondelos. 2007. Bridging the Gap: Query by Semantic Example. *IEEE Transactions on Multimedia* 9, 5 (2007), 923–938.

[32] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. 2010. A New Approach to Cross-modal Multimedia Retrieval. In *ACM International Conference on MultiMedia*.

[33] M. Ren, R. Kiros, and R. Zemel. 2015. Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems*.

[34] A. Sharma, A. Kumar, D. Hal, and D. Jacobs. 2012. Generalized Multiview Analysis: A Discriminative Latent Space. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[35] N. Srivastava and R. Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. In *Advances in Neural Information Processing Systems*.

[36] I. Sutskever, O. Vinyals, and Q. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*.

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[38] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. 2013. Learning Coupled Feature Spaces for Cross-modal Matching. In *IEEE International Conference on Computer Vision*.

[39] L. Wang, Y. Li, and S. Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Enbeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[40] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. 2013. Cross-Media Semantic Representation via Bi-directional Learning to Rank. In *ACM International Conference on MultiMedia*.

[41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrices for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[42] X. Zhai, Y. Peng, and J. Xiao. 2013. Heterogeneous Metric Learning with Joint Graph Regularization for Cross-Media Retrieval. In *AAAI Conference on Artificial Intelligence*.

[43] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian. 2016. PL-ranking: A Novel Ranking Method for Cross-Modal Retrieval. In *ACM International Conference on MultiMedia*.