Learning Deep Semantic Embeddings for Cross-Modal Retrieval

Cuicui Kang KANGCUICUI@IIE.AC.CN
No.89A Minzhuang Road, Beijing, China

Shengcai Liao[*] SCLIAO@NLPR.IA.AC.CN
No.95 Zhuangguancun East Road, Beijing, China

Zhen Li, Zigang Cao, Gang Xiong LIZHEN,CAOZIGANG,XIONGGANG@IIE.AC.CN
No.89A Minzhuang Road, Beijing, China

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

Deep learning methods have been actively researched for cross-modal retrieval, with the softmax cross-entropy loss commonly applied for supervised learning. However, the softmax cross-entropy loss is known to result in large intra-class variances, which is not not very suited for cross-modal matching. In this paper, a deep architecture called Deep Semantic Embedding (DSE) is proposed, which is trained in an end-to-end manner for image-text cross-modal retrieval. With images and texts mapped to a feature embedding space, class labels are used to guide the embedding learning, so that the embedding space has a semantic meaning common for both images and texts. This way, the difference between different modalities is eliminated. Under this framework, the center loss is introduced beyond the commonly used softmax cross-entropy loss to achieve both inter-class separation and intra-class compactness. Besides, a distance based softmax cross-entropy loss is proposed to jointly consider the softmax cross-entropy and center losses in fully gradient based learning. Experiments have been done on three popular image-text cross-modal retrieval databases, showing that the proposed algorithms have achieved the best overall performances.

Keywords: Cross-Modal Retrieval, Deep Learning, Semantic Embedding Learning.

1. Introduction

With the fast development of the Internet techniques, the amount of multimedia data on the Internet has been rapidly increased. As well as the data modalities, such as images, audios, documents, etc. As a result, cross-modal learning has become one of the most difficult challenge in large-scale Internet data mining and discovering. Meanwhile, the cross-modal retrieval problem is raised to match cross-modal features directly Rasiwasia et al. (2010), which is the key technique to realize the modality transparent on the Internet. In recent years, the problem gains many researchers' attentions and has been widely studied in many applications, especially in multimedia Pan et al. (2014); Sharma et al. (2012).

To deal with this problem, many algorithms have been proposed to learn a common feature space for different modality samples, such as the classical Canonical Correlation Analysis (CCA) Hardoon et al. (2004); Gong et al. (2014); Hwang and Grauman (2010)

[*] Corresponding Author.

and Partial Least Squares regression (PLS) algorithms Rosipal and Krämer (2006). Both algorithms aim at learning a latent space where the correlations between projected vectors of two modalities could be maximized. However, these algorithms work not well in relieving the heterogeneity and diversity in different modalities, such as the "semantic gap" between images and texts Wang et al. (2012). In order to best eliminate the heterogeneity in different modalities, Rasiwasia et al. (2010) proposed a Semantic Correlation Matching (SCM) algorithm based on the CCA algorithm. They pointed out that semantic level matching combined with correlation learning methods (such as CCA and PLS) can bring much more benefit in working out the problem, rather than the correlation matching only. It is consistent with the work of Sharma et al. (2012) which shows that the class label information is very helpful to reduce the semantic gap, and so a generalized multiview analysis algorithm is proposed. Inspired by these studies, other methods have been further proposed to improve the class label guided learning, such as Kang et al. (2015), Zhuang et al. (2013), Wang et al. (2013), etc.

In recent years, deep learning methods have also been proposed to address the cross-modal retrieval problem, including deep boltzmann machines Ngiam et al. (2011); Srivastava and Salakhutdinov (2012), cross-modal auto-encoders Feng et al. (2014), deep CCA Andrew et al. (2013), deep convolutional neural networks (CNN) Wang et al. (2016), and deep metric learning Liong et al. (2017); He et al. (2016). The former three kinds of methods do not use the class label information. For deep metric learning methods, pairwise multi-modal samples are required, from which it is not easy to learn a model that generalizes well when training data is limited. For the deep CNN method Wang et al. (2016), the softmax cross-entropy loss is applied for classification. However, recent research Wen et al. (2016) in face recognition shows that though the softmax cross-entropy loss can separate inter-class embeddings efficiently, it does not explicitly reduce the intra-class variance. Therefore, it is not very suited for verification or matching problems, especially cross-modal matching.

Working with class labels commonly shared between images and texts, in this paper, a deep architecture called Deep Semantic Embedding (DSE) is proposed, which is trained in an end-to-end manner for image-text cross-modal retrieval. With a stack of CNNs, FCs and nonlinear activations, both images and texts are mapped to a feature embedding space. In addition, class labels are used to guide the embedding learning, so that the embedding space has a semantic meaning common for both images and texts. This way, the difference between different modalities is eliminated. Under this framework, the center loss Wen et al. (2016) is introduced beyond the commonly used softmax cross-entropy loss to achieve both inter-class separation and intra-class compactness. Besides, a distance based softmax cross-entropy loss is proposed to jointly consider the softmax cross-entropy and center losses in fully gradient based learning. Experiments have been done on three popular image-text cross-modal retrieval databases, showing that the proposed algorithms have achieved the best overall performances.

The rest of this paper is organized as follows. Section 2 gives a comprehensive overview of the related works. Section 3 introduces the proposed DSE framework and loss functions. The experiments and analyses are shown in Section 4. Finally, the paper is summarized in Section 5.

2. Related Work

The cross-modal matching and retrieval problem has been actively researched Bronstein et al. (2010); Pan et al. (2014); Jia et al. (2011); Zhuang and Hoi (2011); Zhuang et al. (2013); Sharma et al. (2012); Zhu et al. (2014); Gong et al. (2014). Among these related works, the CCA Hardoon et al. (2004) and PLS algorithms Rosipal and Krämer (2006); Sharma et al. (2012) are classical algorithms dealing with the cross-modal retrieval problem. In fact, the CCA and PLS both aim at learning a latent low dimensional space by maximizing the correlating relationships between two modality features, though they use different techniques to extract latent vectors Rosipal and Krämer (2006). So far, many extensions derived from CCA and PLS have been proposed for cross-modal retrieval Hardoon et al. (2004); Hwang and Grauman (2010); Rasiwasia et al. (2010); Socher and Li (2010); Li et al. (2011). Specifically, Hardoon et al. (2004) applied the kernel CCA (KCCA) to learn a latent space. Hwang and Grauman (2010) also proposed a KCCA based algorithm to discover the relationship between the tag cues and the image content. As for PLS, Sharma and Jacobs (2011) adopted the PLS algorithm for face images matching between photos and sketches. It was also applied to the cross-media retrieval problem by Sharma et al. (2012).

However, the classical CCA and PLS algorithms are unsupervised algorithms. Rasiwasia et al. (2010) showed that the Semantic Matching (SM) derived from class label information helps reducing the semantic gap between images and texts. Accordingly, they proposed a well-known cross-media retrieval algorithm called SCM, which used CCA for the first step to learn a maximally correlated subspace, and class labels were further utilized for semantic matching in the second step. Inspired by this, Sharma et al. (2012) proposed a generalized multiview analysis algorithm (GMA) utilizing the valuable class label information. In order to utilize the weakly paired samples, Lampert and Krömer (2010) proposed a Weakly-Paired Maximum Covariance Analysis (WMCA) algorithm for multi-modal dimensionality reduction. Wang et al. (2013) proposed a half-quadratic optimization based algorithm to learn a coupled feature space for two modalities. Inspired by metric learning, Kang et al. (2015) proposed a bilinear model to learn a similarity function with two different modality features, where deep CNN features are used for image representation.

In recent years, deep learning methods have been actively researched for cross-modal retrieval. Ngiam et al. (2011) proposed a deep belief network with the restricted boltzmann machine for shared feature learning of different modalities. In the work of Srivastava and Salakhutdinov (2012), the deep boltzmann machine was also applied to learn a joint feature space of images and texts in an end-to-end way. Besides, Feng et al. (2014) proposed the correspondence auto-encoder for cross-modal learning. A Deep CCA algorithm is proposed by Andrew et al. (2013) for cross-modal retrieval. However, the above deep learning methods are all unsupervised, which do not use the class label information. Considering this, Wang et al. (2016) proposed a supervised approach using deep CNN and skip-gram model for image-text retrieval, where the softmax cross-entropy loss was applied for supervised learning. However, recent research of Wen et al. (2016) shows that though the softmax cross-entropy loss can separate inter-class embeddings efficiently, it does not explicitly reduce the intra-class variance. Therefore, it is not very suited for verification or matching problems, especially cross-modal matching. He et al. (2016) proposed a deep and bidirectional representation learning model where the CNN was used for image and text modeling

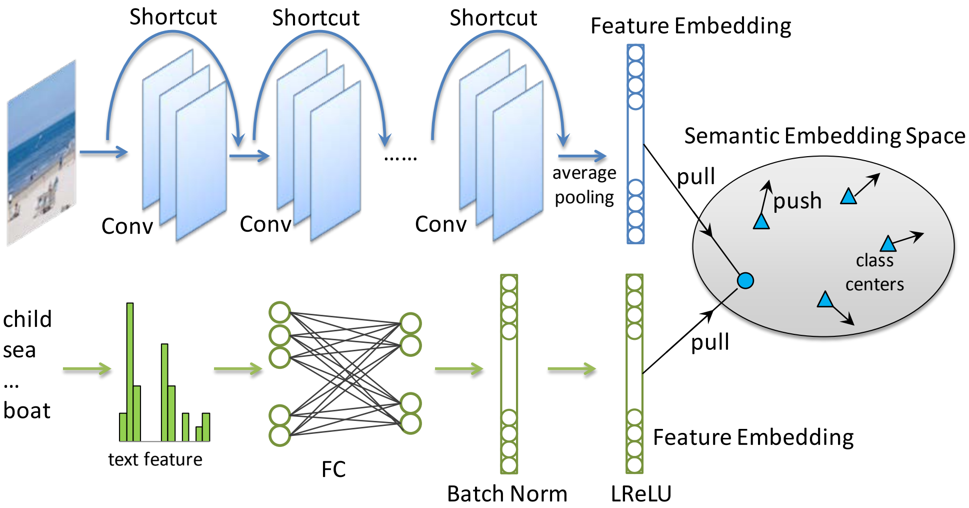


Figure 1: The overall architecture of the proposed DSE method.

with paired samples. Another kind of work is deep metric learning. Liong et al. (2017) proposed to learn a deep non-linear mapping function for metric learning, and used precomputed features. For the two methods, pairwise multi-modal samples are required, from which it is not easy to learn a model that generalizes well when training data is limited. In this paper, class labels are also used to guide the cross-modal semantic embedding learning, but are not limited to pairwise samples for the training efficiency. Besides, beyond the softmax cross-entropy loss, the center loss and distance based softmax cross-entropy loss are integrated to improve the semantic embedding learning.

3. Deep Semantic Embedding

For cross-modal matching such as between images and texts, there is a common problem known as the semantic gap, which refers to the difference between the low-level vision and the high-level semantic description. Utilizing the class label information has been found as a good way to reduce the semantic gap. Therefore, working along this direction, in this paper a deep architecture called Deep Semantic Embedding (DSE) is proposed, which is trained in an end-to-end manner for cross-modal retrieval.

3.1. Overall Architecture

The overall architecture of the proposed method is shown in Fig. 1. For images, a convolutional neural network is used for feature extraction. Specifically, the ResNet-50network is equipped for image representation learning He et al. (2015) . The ResNet-50 network contains 49 CNN layers with short-cut connections, with the final full connection (FC) layer dropped. Strided CNN is used instead of max-pooling. Finally, a vector with $d = 2048$ dimensions is obtained as the feature embedding. For texts, since text classes are relatively easy to classify, to avoid overfitting, only one FC layer is used to map the text to also a

2048-dimensional feature vector. A batch normalization layer is applied thereafter, and the leaky rectified linear unit (LReLU) is further used as the non-linear activation function.

As a result, both images and texts are mapped to a 2048-dimensional feature embedding, which is treated as a common space to eliminate the difference between different modalities. To achieve this, the class labels commonly shared between images and texts are used to guide the embedding learning, so that the embedding space has a semantic meaning common for both images and texts. This way, different from the cross-modal metric learning methods Kang et al. (2015), images and texts can be directly matched via the Euclidean distance or Cosine measure with the learned common semantic embeddings.

Under this framework, three losses are exploited to learn effective semantic embeddings. The first one is the commonly used softmax cross-entropy loss, with the resulting model denoted as DSE-S. The second one is the center loss Wen et al. (2016) recently proposed in face recognition, with the resulting model (together with the softmax cross-entropy loss) denoted as DSE-CS. And finally, a distance based softmax cross-entropy loss is proposed to jointly consider the softmax cross-entropy and center losses in fully gradient based learning. With this loss the learned model is denoted as DSE-DS.

The three losses will be described as follows.

3.2. Softmax Cross-Entropy Loss

The softmax cross-entropy loss is commonly used for the multi-class classification task. It is formulated as

$$\ell_s = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{\boldsymbol{w}_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \boldsymbol{x}_i + b_j}}, \tag{1}$$

where M is the batch size, $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th embedding (either image or text) of the batch, y_i is the class label of $\boldsymbol{x}_i$, $\boldsymbol{w}_j \in \mathbb{R}^d$ is the projection weights and b_j is the bias term for the j-th class, and C is the number of categories.

However, though the softmax cross-entropy loss is very useful for class separation and prediction, it is not optimal for the verification or matching task, such as face recognition, as observed in Wen et al. (2016). This is because the objective of softmax cross-entropy makes the samples distributed like a starburst (see examples shown in Sec. 4.5), which is pretty good for classification but with a large intra-class variation.

3.3. Center Loss

To reduce the intra-class variance, Wen et al. (2016) introduced an additional constraint called center loss in the context of face recognition. This method runs a moving average for the embedding of each class, and keeps pushing each image embedding to its corresponding class center so that the variations between image embeddings and their centers are reduced. For class label guided cross-modal matching, such as matching between images and texts, it is even more important to reduce the intra-class variance, so that different modalities of the same class will have small distances to enable direct matching. Therefor, the center loss is also exploited in this paper for cross-modal matching.

Formally, it is formulated as

$$\ell_c = \frac{1}{M}\sum_{i=1}^{M} ||\boldsymbol{x}_i - \boldsymbol{c}_{y_i}||_2^2, \tag{2}$$

where $\boldsymbol{c}_{y_i} \in \mathbb{R}^d$ is the class center of the embedding $\boldsymbol{x}_i$. Differently, unlike other weight parameters to be learned by backpropagation, the updating of the class centers $\boldsymbol{c}_j, j = 1, 2, \ldots, C$ are additionally performed as follows,

$$\begin{aligned} \Delta \boldsymbol{c}_j^t &= \frac{\sum_{i=1}^{M} \delta(y_i = j)(\boldsymbol{c}_j^t - \boldsymbol{x}_i)}{\sum_{i=1}^{M} \delta(y_i = j)}, && (3) \\ \boldsymbol{c}_j^{t+1} &= \boldsymbol{c}_j^t - \alpha \Delta \boldsymbol{c}_j^t, && (4) \end{aligned}$$

where $\alpha \in (0, 1)$ is the update rate of the centers, δ is the indicator function, and t is the batch step. In this paper, we set $\alpha = 0.5$, which is not sensitive, as in Wen et al. (2016).

The center loss is introduced as an additional constraint to the softmax cross-entropy loss. As a result, the total loss is

$$\ell_{cs} = \ell_s + \lambda \ell_c, \tag{5}$$

where $\lambda > 0$ is a parameter to balance the two loss functions.

3.4. Distance Based Softmax Cross-Entropy Loss

In the center loss implementation, it requires both the weight variables $\mathbf{W} \in \mathbb{R}^{d\times c}$ and bias variables $\boldsymbol{b} \in \mathbb{R}^c$ in the softmax cross-entropy loss, and the centers variables $\mathbf{C} \in \mathbb{R}^{d\times c}$, totally $c(d+1)$ parameters and $O\left(c(2d+1)\right)$ memory requirement. Though the centers variables are not trainable parameters in backpropagation, they still require additional update operations and nearly the same amount of memory as in softmax.

To address this, in this paper, a distance based softmax cross-entropy loss is proposed to jointly consider the softmax cross-entropy and center losses in fully gradient based learning. The DistSoftmax loss directly focuses on learning semantic centers in the embedding space. Specially, with centers variables $\mathbf{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_c] \in \mathbb{R}^{d\times c}$, it measures the Euclidean distances of each sample embedding to all the centers, uses the negative distances directly as the logits of the softmax function, and also additionally reduces the intra-class distances.

Formally, the DistSoftmax loss is formulated as

$$\ell_{ds} = \frac{1}{M}\sum_{i=1}^{M}\left(-\log \frac{e^{-||\boldsymbol{x}_i - \boldsymbol{c}_{y_i}||_2^2}}{\sum_{j=1}^{C} e^{-||\boldsymbol{x}_i - \boldsymbol{c}_j||_2^2}} + \lambda ||\boldsymbol{x}_i - \boldsymbol{c}_{y_i}||_2^2 \right), \tag{6}$$

where $\lambda > 0$ is also a balancing parameter. Intuitively, it seems that with the cross-entropy term above is enough to minimize the intra-class distances while maximizing the inter-class distances. However, the study in this paper (see Sec. 4.5) shows that it is not effective without the additional intra-class compactness constraint. This is probably because the distances have an asymmetric range $[0, \infty)$ instead of $(-\infty, \infty)$ as usual, thus limiting the optimization of the cross-entropy loss.

Note that different from the center loss, the centers variables in the DistSoftmax loss is automatically learned by gradient optimization, and it does not require another set of variables in softmax for class separation. Therefore, the DistSoftmax loss only requires cd parameters and $O(cd)$ memory.

3.5. Implementation Details

The proposed DSE model is implemented in Keras Chollet et al. (2015) with the TensorFlow[1] backend. Since the image-text datasets used in the experiments are not very large, the ResNet-50 model pre-trained in the ImageNet by He et al. (2015) is loaded for initialization, except for the top FC layer. A weight decay of 0.001 is used for regularization of the model weights during training. The negative slope coefficient of the LReLU is set to 0.2. The default Adam optimizer in Keras is used for model optimization. The batch size is set to 32. The maximal number of training epoches is 400 (usually the training is converged below 100 epoches in the experiments). The initial learning rate is 0.001, decayed by a factor of 0.1 every time a plateau is detected by monitoring the training accuracy in a tolerance of 10 epoches. If the plateau is still detected in a tolerance of 15 epoches, the training is regarded as converged and an early stopping will be triggered. All images are cropped with their square-size center parts, and resized to 224 × 224. Then, a real-time image augmentation is applied, with a random shift by a factor of 0.1 in both horizontal and vertical directions, and a random horizontal flipping.

The balance parameter λ is set to 0.01 and 0.1, respectively, in the center and DistSoftmax losses (a parameter analysis will be given in Sec. 4.5). For the training of image-text model, a batch of both image and text samples with equal number and the same class labels are randomly sampled in every batch iteration, and their corresponding losses are aggregated with a coefficient of 0.5. In the test phase, the Cosine measure is used to compute the similarity score between image and text embeddings. Note that both image and text embeddings share the same center variables, and FC weights and bias terms in softmax if applicable, so as to learn a common semantic embedding space. Note also that the image-text samples are not required to be paired in either data source or training procedure in the proposed DSE framework, in contrast to deep metric learning methods. This makes the DSE framework convenient and efficient to train.

4. Experiments

The proposed DSE-CS and DSE-DS algorithms were evaluated on three popular image-text databases, namely the Wikipedia Rasiwasia et al. (2010), PascalVOC2007 Everingham et al. (2007), and LabelMe Oliva and Torralba (2001) databases. The softmax baseline DSE-S, as well as some other popular and state-of-the-art algorithms were imported for comparison.

4.1. Experimental Setting

Compared Methods: Except the DSE-S as baseline, several famous algorithms in the cross-media retrieval field were also imported for comparision. Specifically, the classical CCA Hardoon et al. (2004) and PLS Rosipal and Krämer (2006); Sharma and Jacobs (2011) based latent space learning methods were used as baseline algorithms. In addition, state-of-the-art algorithms in the cross-media field were also compared, such as the SCM Rasiwasia

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

et al. (2010), Microsoft algorithm (MsAlg) Wu et al. (2010), Low Rank Bilinear Similarity learning (LRBS) Kang et al. (2015), and Generalized Multiview Analysis methods, including GMLDA (Generalized Multiview Linear Discriminant Analysis) and GMMFA (Generalized Multiview Marginal Fisher Analysis) Sharma et al. (2012).

For the traditional algorithms CCA, PLS, SCM, and GMA, the AlexNet Krizhevsky et al. (2012); Donahue et al. (2013) trained on the ImageNet was used to extract CNN features from images. The outputs of the sixth layer of the AlexNet were used as image features, resulting in 4096 dimensions. The results of the compared algorithms without CNNs were not included due to the low performances Sharma et al. (2012); Wang et al. (2013). Note that except the ImageNet which was the data source of the pre-trained models, no other outside data was used for training.

In the experiments, it was found that the CCA, PLS and GMA algorithms all performed better with PCA dimensional reduction than without it. Specifically, the dimensionality of the CNN image features was reduced to 1000 (about 99% energy was preserved) by PCA on the three databases.

Evaluation Metrics: For the evaluation, the mean average precision (mAP) score was used in the experiments. Besides, the precision-recall curve and precision-scope curve were applied to display the experimental results for better visualization Rasiwasia et al. (2007); Wang et al. (2013), which were popularly used in the evaluation of information retrieval systems.

The precision-scope curve shows the precision at the top N retrieved samples, where the scope denotes the number of the retrieved samples. As for the mAP, it is computed as the mean of the average precision score, which is related to the ranking of the retrieved samples. Specifically, given one query and its top N retrieved samples in a ranking list, the average precision is computed by

$$AveP = \frac{1}{T}\sum\nolimits_{r=1}^{N} P(r)rel(r), \tag{7}$$

where $P(r)$ is the precision of the top r retrieved samples. The $rel(r)$ is a binary function denoting whether the r-th retrieved sample is relevant to the query or not. T is the number of relevant samples in the retrieved results. With the AveP calculated for each query task, the mean average precision is computed as the average AveP score over all queries.

4.2. On Wikipedia Database

The Wikipedia database[2] consists of 2866 images and documents pairs from ten categories. The database is built from the the Wikipedia's "featured articles", which is selected, reviewed and continually updated by the Wikipedia's editors since 2009. To realize direct matching of different modality features, Rasiwasia et al. (2010) built the wikipedia dataset by selecting ten popular categories from the collection, including art, biology, geography, history, literature, media, music, royalty, sport, and warfare. Then, the dataset was randomly split into a training set of 2173 image-document pairs and a test set of the remaining 693 pairs by the authors. Each document associated to images is constructed of several paragraphs, resulting in at least 70 words. Since the documents contain more than 40k

[2] http://www.svcl.ucsd.edu/projects/crossmodal/

Table 1: mAP (%) results on the Wikipedia database.

Methods	Image query	Text query	Average
CNN+CCA	– (19.70)	– (17.84)	– (18.77)
CNN+SCM	– (37.13)	– (28.23)	– (32.68)
CNN+PLS	30.54 (30.55)	28.05 (28.03)	29.30 (29.29)
CNN+MsAlg	37.28 (36.07)	32.70 (30.75)	34.99 (33.41)
CNN+GMLDA	24.93 (36.77)	18.18 (29.71)	21.55 (33.24)
CNN+GMMFA	24.03 (38.74)	16.51 (31.09)	20.27 (34.91)
CNN+LRBS	44.48 (44.41)	37.70 (37.70)	41.09 (41.06)
DSE-S	33.82	35.62	34.72
DSE-CS	46.57	39.50	43.03
DSE-DS	**47.67**	**40.40**	**44.03**

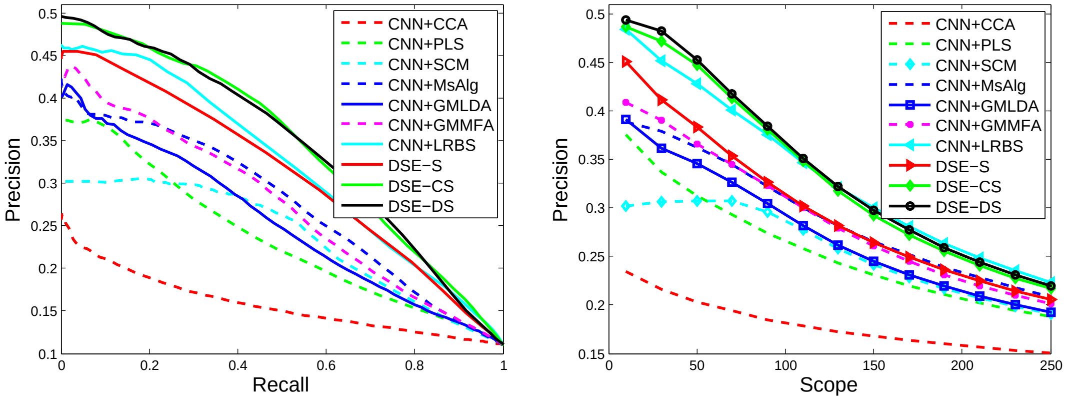


Figure 2: Precision-recall and precision-scope curves on the Wikipedia database.

unique words, the Latent Dirichlet Allocation model (LDA) Costa Pereira et al. (2014) was used to extract text feature vectors.

The mAP results are shown in Table 1, and the average precision-recall and precision-scope curves are displayed in Fig.2 for retrieval tasks of image to text and text to image. Note that the mAP results in brackets of Table 1 are the performances of the compared algorithms processed with the PCA dimensional reduction. From Table 1, we can see that the proposed DSE-DS algorithm achieves the best performance on the Wikipedia database with the average mAP of 44.03%, followed by the DSE-CS algorithm with the performance of 43.03%. The LRBS algorithm achieves the third best performance with 41.09%. The DSE-S results in only 34.72%, with a gap of more than 6% to the LRBS algorithm with deep CNN image features, and is more than 8% lower than DSE-CS and DSE-DS. As mentioned previously, the softmax cross-entropy loss applied in DSE-S is good for classification, but is not optimal for cross-modal matching due to its large intra-class variations. In contrast, the LRBS uses a bilinear similarity function to learn an adaptive similarity metric for the generalization of unseen samples, and the DSE-CS and DSE-DS introduces additional constraints to reduce intra-class variations. As for traditional methods, either with PCA or

Table 2: mAP (%) results on the Pascal VOC2007 database.

Methods	Image query	Text query	Average
CNN+CCA	– (49.06)	– (48.33)	– (48.69)
CNN+SCM	– (63.98)	– (59.73)	– (61.86)
CNN+PLS	47.55 (47.55)	45.69 (45.68)	46.62 (46.62)
CNN+MsAlg	58.43 (58.58)	59.40 (58.84)	58.91 (58.71)
CNN+GMLDA	39.95 (65.62)	35.95 (66.32)	37.95 (65.97)
CNN+GMMFA	37.97 (65.48)	33.17 (66.15)	35.57 (65.81)
CNN+LRBS	65.15 (65.10)	68.74 (68.69)	66.95 (66.90)
DSE-S	60.20	62.60	61.40
DSE-CS	71.81	73.99	72.90
DSE-DS	**74.14**	**75.23**	**74.69**

without PCA, their performances are obviously inferior than modern methods using metric learning or effective semantic embedding learning.

Fig. 2 display the precision-recall curves and the precision-scope curves of the compared algorithms on the Wikipedia database, where the precision is the average precision of the text to image and image to text retrieval tasks. The figures display the best performances of each algorithm, either with PCA or without PCA. From these figures, it is also observed that the proposed DSE-CS and DSE-DS algorithms have the best performances on both retrieval tasks, with DSE-DS being slightly better, and the DSE-S does not work very well for the cross-modal matching problem.

4.3. On Pascal VOC2007

The Pascal VOC2007 database[3] contains a total of 9963 images from 20 categories in realistic scenes. It has been divided into a training set with 5011 images and a test set with 4952 images for object classification by the Pascal VOC challenge organizers Everingham et al. (2007). Then, it was introduced for the cross-modal retrieval task by Sharma et al. (2012), where the images containing only one object were selected for experiments, discarding some multi-label images. As a result, the training set in the experiment consists of 2808 images, and the test set includes 2841 images. The 399-dimensional word counting vectors were used as raw text input.

The results are shown in Table 2 and Fig.3. From Table 2, we can find that the proposed DSE-DS and DSE-CS algorithms achieve the best and second best mAPs, respectively; both of them are much better than the best existing algorithm LRBS with deep CNN features. The proposed DSE-DS algorithm outperforms the LRBS algorithm by nearly 8% in mAP, and it is also better than DSE-CS by about 2%. This verifies that learning discriminant and class-wise compact deep semantic embbedings is effective for cross-modal retrieval. Among traditional methods, it can be observed that the supervised algorithms (GMA, SCM, LRBS and MsAlg) all perform better than the unsupervised algorithms, namely CCA and PLS, showing the importance of class label information in addressing cross-modal retrieval.

[3] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

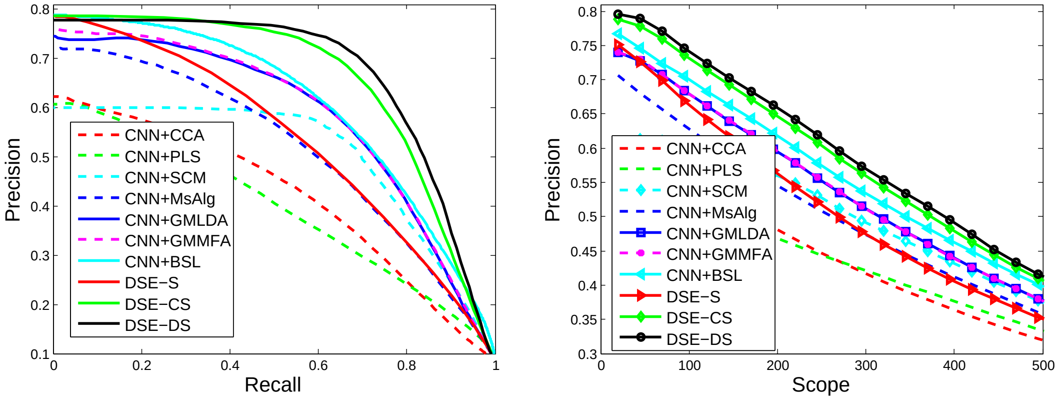


Figure 3: Precision-recall and precision-scope curves on the Pascal VOC2007 database.

From Fig.3 it can also been observed that the proposed DSE-DS and DSE-CS algorithms performance the best, and it is more obvious than in the Wikipedia database that DSE-DS is better than DSE-CS.

4.4. On LabelMe Databse

The LabelMe[4] Oliva and Torralba (2001) database contains outdoor scene images from eight categories, namely "coast", "forest", "highway", "inside city", "mountain", "open country", "street", and "tall building". The number of images for each category varies from 260 to 410, resulting in 2688 images in total. All images are color images, in the size of 256×256 pixels. In the experiment, 100 samples per class from the database were randomly selected for testing, resulting in 800 samples in total as the test set, and the remaining 1888 images were used as the training set. The LabelMe Toolbox[5] was used to generate the word counting vector for the raw textual input. In overall, 781 different tags were obtained, with frequencies varying from one time to more than 2000 times. To remove noises, tags appearing more than 3 times were selected in the experiments.

From Table 3, it can be observed that the proposed DSE-CS and DSE-DS algorithms achieve the best average mAPs, with 92.98% for DSE-DS and 92.54% for DSE-CS, outperforming the best existing algorithm LRBS by about 6%. The two top algorithms also outperform the baseline DSE-S by about 9%. The precision-recall curves and the precision-scope curves of the compared algorithms are shown in Fig.4. Obviously, the proposed DSE-CS and DSE-DS algorithms perform much better than all other algorithms, due to their particular designs in learning effective deep semantic embeddings.

4.5. Analysis and Discussion

To better understand the proposed DSE algorithms, the balance parameter λ is analyzed in this subsection, with its influence shown in Fig. 5 on the LabelMe database. Fig. 5(a)

[4] http://people.csail.mit.edu/torralba/code/spatialenvelope/

[5] http://labelme.csail.mit.edu/Release3.0/

Table 3: mAP (%) results on the LabelMe database.

Methods	Image query	Text query	Average
CNN+CCA	– (26.45)	– (26.80)	– (26.63)
CNN+SCM	– (84.59)	– (84.44)	– (84.52)
CNN+PLS	61.75 (61.76)	68.60 (68.60)	65.18 (65.18)
CNN+MsAlg	75.51 (70.70)	78.44 (72.79)	76.98 (71.74)
CNN+GMLDA	76.29 (85.38)	77.26 (87.22)	76.78 (86.30)
CNN+GMMFA	69.98 (85.33)	62.52 (87.02)	66.25 (86.18)
CNN+LRBS	86.34 (86.26)	87.52 (87.48)	86.93 (86.87)
DSE-S	83.64	83.86	83.75
DSE-CS	92.84	92.24	92.54
DSE-DS	**93.92**	**92.66**	**92.98**

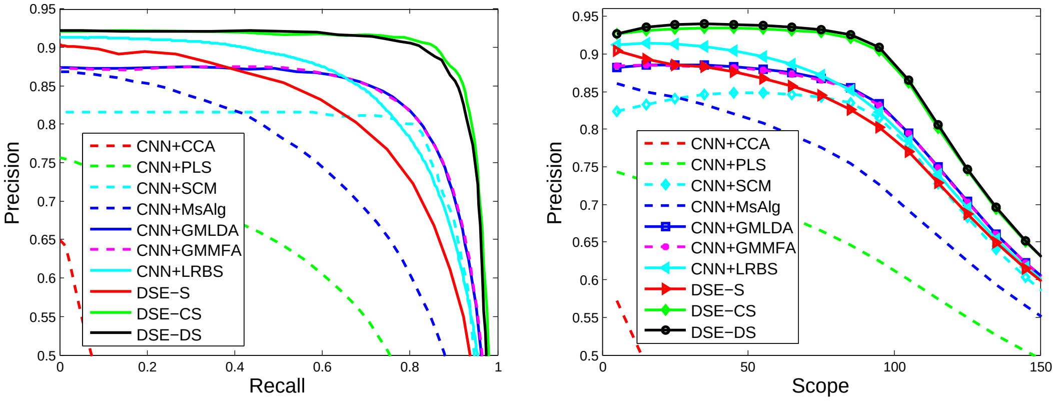
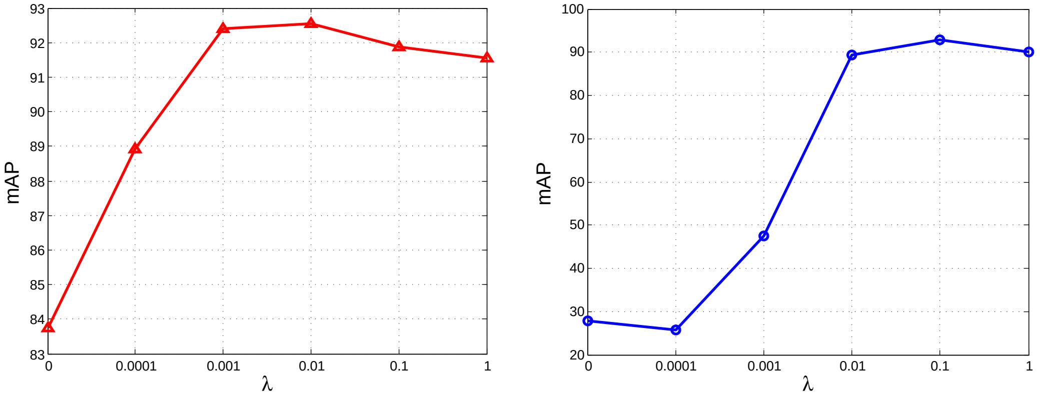


Figure 4: Precision-recall and precision-scope curves on the LabelMe database.

shows the mAPs of the proposed DSE-CS algorithm with varying parameter values of λ acrose a large scale range. Note that the DSE-CS with $\lambda = 0$ is equivalent to the DSE-S algorithm. Accordingly, it can be observed that with λ increasing from 0 to 0.001, the performance of DSE-CS is largely improving, demonstrating that intra-class compactness is very important for class-label guided cross-modal retrieval learning. After that, the performance is relatively stable, with $\lambda = 0.01$ being the best one.

As for the DSE-DS algorithm, a similar result can be observed from Fig. 5(b). Differently, the DSE-DS with small λ values perform very poor, indicating that though in the cross-entropy term in Eq. (6) is towards minimizing the intra-class distances while maximizing the inter-class distances, it is still not effective working alone. Therefore, the additional intra-class compactness constraint is also important here. The performance of DSE-DS is stable when $\lambda \geq 0.01$, with $\lambda = 0.1$ being the best choice. Compared to DSE-CS, the DSE-DS requires a relatively larger λ, because the distances in Eq. (6) have an asymmetric range $[0, \infty)$ instead of $(-\infty, \infty)$ as usual, resulting in a larger cross-entropy loss.

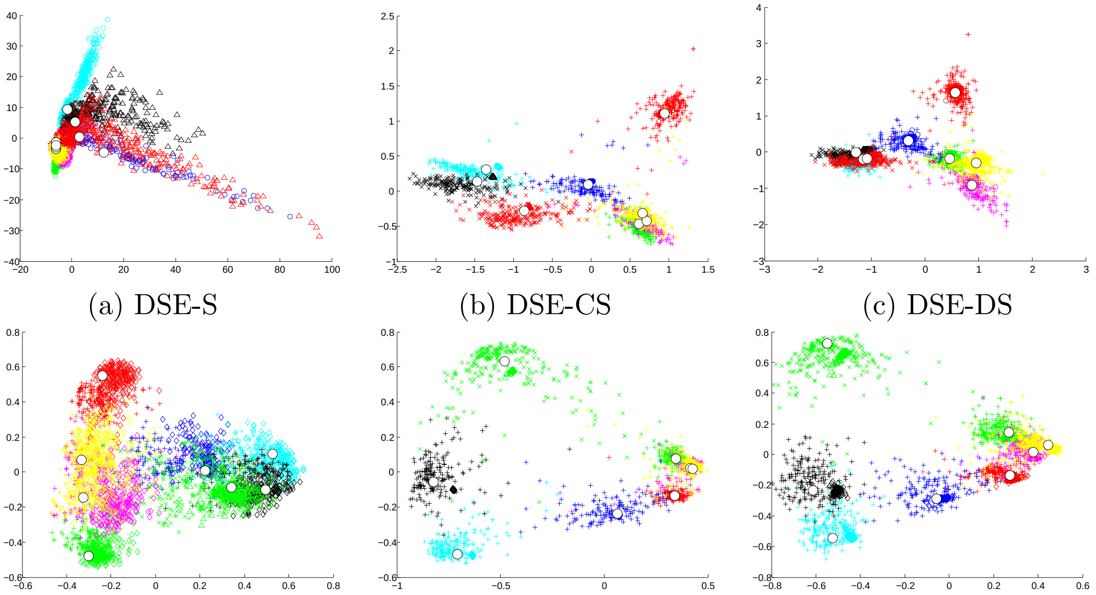
Another interest thing is the role of the Cosine measure or embedding normalization. Fig. 6 shows distributions of the learned image and text embeddings by the DSE-S, DSE-CS



(a) DSE-CS

(b) DSE-DS

Figure 5: The mAPs of the proposed DSE-CS and the DES-DS algorithms with varying parameter values of λ.



(a) DSE-S

(b) DSE-CS

(c) DSE-DS

(d) DSE-S, normalized

(e) DSE-CS, normalized

(f) DSE-DS, normalized

Figure 6: Distributions of the learned image and text embeddings by the DSE-S, DSE-CS and DSE-DS on the LabelMe database after PCA dimension reduction. The first row is with the original embeddings, and the second row is with the embeddings normalized to have unit length. Image and text embeddings of the same class are with the same color but different markers, and white circles indicate class centers.

and DSE-DS on the LabelMe database after PCA dimension reduction. With the original embeddings, it can be observed that the softmax cross-entropy loss results in a starburst

Table 4: mAP (%) results with the Euclidean distance and the Cosine measure.

Methods	Wikipedia		Pascal VOC2007		LabelMe	
	Euclidean	Consine	Euclidean	Consine	Euclidean	Consine
DSE-S	25.67	34.72	34.15	61.40	58.71	83.75
DSE-CS	39.85	43.03	68.37	72.90	91.42	92.54
DSE-DS	35.40	44.03	69.42	74.69	91.23	92.98

distribution, with a very large intra-class variation. In contrast, the DSE-CS and DSE-DS learn much better in condensing the intra-class variation. Intuitively, normalizing the learned embedding may relieve the starburst distribution, as observed in the second row of Fig. 6. However, in this case the intra-class distributions of DSE-S are still not as compact as the DSE-CS and DSE-DS.

According to the above analysis, a comparison is also done between the Euclidean distance and the Cosine measure, since the Cosine measure is equivalent to the Euclidean distance with unit-length normalized embeddings. The results are shown in Table 4. Clearly, the DSE-S is largely improved by embedding normalization, due to the effect of reducing the intra-class variation. The DSE-CS and DSE-DS are less affected since they both have an intra-class compactness constraint. Yet the performance can still be slightly improved by embedding normalization, this is because there has to be a balance between the intra-class compactness and the inter-class separation during learning.

5. Conclusion

In this paper, it is demonstrated that, guided by class labels, learning deep common semantic embeddings is effective for cross-modal retrieval. However, this is not easy to achieve by the commonly used softmax cross-entropy loss, but with the recently proposed center loss, and a new loss called distance based softmax cross-entropy introduced in this paper. This is because the softmax cross-entropy loss only considers the inter-class separation, and so usually produces large intra-class variations. In contrast, the two improved losses achieve both inter-class separation and intra-class compactness, which is important for accurate cross-modal matching. From the experiments it shows that the newly introduced distance based softmax cross-entropy loss performs better than the center loss, with reduced number of parameters and memory requirement. Yet note that as many other existing methods, the proposed method is not applicable in case where class labels are not available. In future research, it is worth investigating some other modalities and learning with large-scale data.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2016YFB0801200) and the National Natural Science Foundation of China (No. 61702501, 61602472, and U1636217).

References

Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML (3)*, volume 28, pages 1247–1255, 2013.

M.M. Bronstein, AM. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535, 2014.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results, 2007. URL http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Multimedia*, pages 7–16, 2014.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2014.

David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, 2015.

Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Trans. Multimedia*, 18(7):1363–1377, 2016.

Sung Ju Hwang and Kristen Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, pages 1–12, 2010.

Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-modality similarity for multinomial data. In *International Conference on Computer Vision*, pages 2407–2414, 2011.

Cuicui Kang, Shengcai Liao, Yonghao He, Jian Wang, Wenjia Niu, Shiming Xiang, and Chunhong Pan. Cross-modal similarity learning: A low rank bilinear formulation. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1251–1260, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

Christoph H. Lampert and Oliver Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV*, pages 566–579, 2010.

Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. Face recognition based on non-corresponding region matching. In *International Conference on Computer Vision*, pages 1060–1067, 2011.

V. E. Liong, J. Lu, Y. P. Tan, and J. Zhou. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 19(6):1234–1244, 2017.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. Click-through-based cross-view learning for image search. In *ACM conference on Research and Development in Information Retrieval (SIGIR)*, 2014.

Nikhil Rasiwasia, Pedro J. Moreno, and Nuno Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010.

Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *SLSFS*, pages 34–51. Springer, 2006.

Abhishek Sharma and David W. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR*, pages 593–600, 2011.

Abhishek Sharma, Abhishek Kumar, Hal Daum III, and David W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

Richard Socher and Fei-Fei Li. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the IEEE Conference on CVPR*, pages 966–973. IEEE, 2010.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2231–2239, 2012.

Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for large-scale search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(12):2393–2406, 2012.

Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *International Conference on Computer Vision*, pages 2088–2095, 2013.

Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *VLDB J.*, 25(1):79–101, 2016.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

Wei Wu, Jun Xu, and Hang Li. Learning similarity function between objects in heterogeneous spaces. Technical Report MSR-TR-2010-86, 2010.

Fan Zhu, Ling Shao, and Mengyang Yu. Cross-modality submodular dictionary learning for information retrieval. In *CIKM*, pages 1479–1488, 2014.

Jinfeng Zhuang and Steven C. H. Hoi. A two-view learning approach for image tag ranking. In *WSDM*, pages 625–634, 2011.

Yueting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, and Weiming Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.