Deep Cross-Modal Projection Learning for Image-Text Matching

Ying Zhang[0000−0002−6005−4989] and Huchuan Lu[0000−0002−6668−9758]

Dalian University of Technology, China
zydl0907@mail.dlut.edu.cn, lhchuan@dlut.edu.cn

Abstract. The key point of image-text matching is how to accurately measure the similarity between visual and textual inputs. Despite the great progress of associating the deep cross-modal embeddings with the bi-directional ranking loss, developing the strategies for mining useful triplets and selecting appropriate margins remains a challenge in real applications. In this paper, we propose a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning discriminative image-text embeddings. The CMPM loss minimizes the KL divergence between the projection compatibility distributions and the normalized matching distributions defined with all the positive and negative samples in a mini-batch. The CMPC loss attempts to categorize the vector projection of representations from one modality onto another with the improved norm-softmax loss, for further enhancing the feature compactness of each class. Extensive analysis and experiments on multiple datasets demonstrate the superiority of the proposed approach.

Keywords: Image-Text Matching · Cross-Modal Projection · Joint Embedding Learning · Deep Learning

1 Introduction

Exploring the relationship between image and natural language has recently attracted great interest among researchers, due to its great importance in various applications, such as bi-directional image and text retrieval [44, 22], natural language object retrieval [10], image captioning [43, 35], and visual question answering (VQA) [1, 18]. A critical task for these applications is to measure the similarity between visual data and textual descriptions. Existing deep learning approaches either attempts to learn joint embeddings [39, 44, 40, 21] for image and text in a shared latent space, or build a similarity learning network [16, 15, 22, 11, 40] to compute the matching score for image-text pairs. The joint embedding learning based methods have shown great potential in learning discriminative cross-modal representations and computation efficiency at the test stage.

Generally, the joint embedding learning framework for image-text matching adopts the two-branch [40, 39, 44, 21] architecture (as shown in Fig. 1), where one

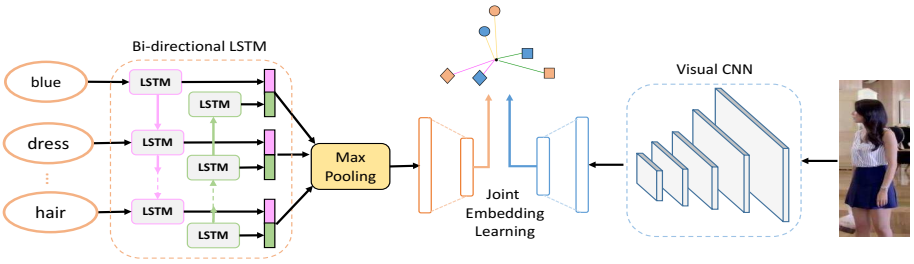


Fig. 1. Deep image-text embedding learning

branch extracts the image features and the other one encodes the text representations, and then the discriminative cross-modal embeddings are learned with designed objective functions. The most commonly used functions include canonical correlation analysis (CCA) [44], and bi-directional ranking loss [39, 40, 21]. Compared with CCA based methods, the bi-directional ranking loss produces better stability and performance [40] and is being more and more widely used in cross-modal matching [39, 21]. Nevertheless, it suffers from sampling useful triplets and selecting appropriate margins in real applications.

Despite the great success of these deep learning techniques in matching image and text with only the pair correspondence, some recent works [28, 16, 15] explore more effective cross-modal matching algorithms with identity-level annotations. These research efforts demonstrated that the discrimination ability of the learned image-text embeddings can be greatly enhanced via introducing category classification loss as either auxiliary task [28] or pre-trained initialization [16, 15]. Consider the fact that independent classification may not fully exploit the identity information for cross-modal feature learning, [15] developed the Cross-Modal Cross-Entropy (CMCE) loss which employs the cross-modal sample-to-identity affinity for category prediction, whereas this strategy requires to allocate additional identity feature buffer, which could bring large memory consumption when there are large number of subjects.

To address these problems, we propose a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss, which introduces the cross-modal feature projection operation for learning discriminative image-text embeddings. The CMPM loss attempts to minimize the KL divergence between projection compatibility distributions and the normalized matching distributions, in order to increase the variance between unmatched samples and the association between the matched ones. The CMPM loss function does not need to select specific triplets or tune the margin parameter, and exhibits great stability with various batch size. For the assistant classification task with identity labels, the CMPC loss attempts to classify the vector projection of the features from one modality onto the matched features from another modality, instead of independently categorizing the original features. Extensive experiments and analysis demonstrate the superiority of the proposed approach for efficiently learning discriminative image-text embeddings.

2 Related Work

2.1 Deep Image-Text Matching

Most existing approaches for matching image and text based on deep learning can be roughly divided into two categories: 1) joint embedding learning [39, 15, 44, 40, 21] and 2) pairwise similarity learning [15, 28, 22, 11, 40].

Joint embedding learning aims to find a joint latent space under which the embeddings of images and texts can be directly compared. This type of approaches usually associate features from two modalities with correlation loss [44], and the bi-directional ranking loss [39, 40, 21]. The deep canonical correlation analysis (DCCA) [44] aims to learn nonlinear transformations of two views of data with the deep networks such that the resulting representations are highly linearly correlated, while the major caveat of DCCA is the eigenvalue problem brought by unstable covariance estimation in each mini-batch [23, 40]. The bi-directional ranking loss [39, 40, 21] extends the triplet loss [29], which requires the distance between matched samples to be smaller than unmatched ones by a margin for image-to-text and text-to-image ranking. Whereas the bi-directional ranking loss inherits the disadvantage of selecting negative samples and margins from the triplet loss.

Pairwise similarity learning focus on designing a similarity network which predicts the matching score for image-text pairs. Apart from the efforts [40] to measure the global similarity between image and text, many of the research works [15, 28, 22, 11, 26] attempt to maximize the alignments between image regions and textual fragments. However, this strategy may lack efficiency involving preparing all the image-text pairs to predict the matching score at the test stage.

For image-text matching with identity-level annotations, Reed *et al.* [28] proposed to learn discriminative image-text joint embeddings with the indication of class labels, and collected two datasets of fine-grained visual descriptions, while [16] attempted to search persons with language description under the assistance of identity classification. As an improvement, Li *et al.* [15] developed a two-stage learning strategy for textual-visual matching. Stage-1 pre-trains the network with the cross-modal cross-entropy (CMCE) loss under the supervision of identity labels, and stage-2 retrains the network with latent co-attention restriction under the supervision of pairwise labels.

2.2 Discriminative Feature Learning

Recent years have witnessed the advance of deep neural networks for learning discriminative features, which has great importance in many visual tasks, such as face recognition [32, 29, 41, 20, 19], face verification [33, 37], and person re-identification [42, 8, 2]. Intuitively, discriminative features should be able to maximize both the inter-class separability and the intra-class compactness.

As the most widely used supervision loss for learning strong representations, cross-entropy loss (or softmax loss) [32, 33, 42] has achieved significant success in various applications. Nevertheless, many research works have been focusing on

improvements to generate more discriminative features. Wen *et al.* [41] proposed the center loss to assist the softmax loss for face recognition, where the distance between samples and the corresponding class centres are minimized to improve the intra-class compactness. Liu *et al.* developed the L-softmax [20] which introduces the angular margin into softmax loss for further increasing the feature separability, and refined it to A-softmax [19] by adding the normalization of the classification weights. It is notable that the A/L-softmax imposes feature discriminativeness by incorporating the angular margin to achieve remarkable results in face recognition. However, the strong restriction of angular and weights makes models difficult to converge [36, 3, 38] in real applications, especially when the training data has too many subjects. Ranjan *et al.* [27] proposed to normalize the features to strengthen the verification signal and better model the difficult samples. Wang *et al.* [37] modified the softmax loss by normalizing both the features and the classification weights, which achieves performance improvements with much easier implementation.

On the other hand, deep metric learning gains increasing popularity by learning general distance metrics, under which the distance between relevant samples are smaller than that of irrelevant ones. Hadsell *et al.* [5] proposed the contrastive loss to minimize the distance between similar points and restrict the distance between dissimilar points to be smaller than a margin. Schroff *et al.* [29] designed the triplet loss to encourage a relative distance constraint between matched face pairs and unmatched ones, and it has proved effective for matching pedestrians from different cameras in [8]. Recently, quadruplet loss [2] added a negative pair constrain to the triplet loss such that the intra-class variations and inter-class similarities are further reduced. It also introduced the adaptive margin to compute distance penalization and select negative samples.

Unfortunately, there are two main challenges when applying the above loss functions: sampling useful data units (i.e. positive and negative pairs, triplets, or quadruplets) and determining appropriate margins. Generating all possible triplets would result in heavy computation and slower convergence [29] while sampling the hardest negatives may cause the network to converge to a bad local optimum [29, 31]. [29] proposed to choose semi-hard negative samples from within a mini-batch online, while this strategy requires large batch size to select useful negative samples. Song *et al.* [31] optimized the smoothed upper bound of the original triplet loss and utilized all the negative samples within a mini-batch, and Sohn *et al.* [30] proposed the N-pair loss in the form of multi-class softmax loss with the request of carefully selected imposter examples. To avoid highly-sensitive parameters, the Histogram loss [34] is developed to estimate the similarity distributions of all the positive and negative pairs in a mini-batch and then minimize the probability that a random negative pair has a higher similarity than a random positive pair, under which the large batch size is preferred to achieve better performance. Nevertheless, these modifications for learning embeddings to preserve the association relationship of samples are specifically designed for single-modal applications, and may not readily adapt to the cross-modal matching problems.

3 The Proposed Algorithm

3.1 Network Architecture

The framework of our proposed method is shown in Fig. 1. We can see that the image-text matching architecture consists of three components: a visual CNN to extract image features, a bi-directional LSTM (Bi-LSTM) to encode text features, and a joint learning module for associating the cross-modal representations.

Given a sentence, we apply basic tokenizing and split it into words, and then sequentially process them with a Bi-LSTM. The hidden states of forward and backward directions are concatenated, and the initial text representations are obtained with a max-pooling strategy. For an image, we employ MobileNet [9] and extract its initial feature from the last pooling layer. In the association module, the extracted image and text features are embedded into a shared latent space, where the compatibility between matched features and the variance between unmatched samples are maximized.

In this paper, we focus on learning the discriminative features in the association module, and describe the proposed cross-modal projection matching (CMPM) and cross-modal projection classification (CMPC) loss function in the following sections.

3.2 Cross-Modal Projection Matching

We introduce a novel image-text matching loss termed as Cross-Modal Projection Matching (CMPM), which incorporates the cross-modal projection into KL divergence to associate the representations across different modalities.

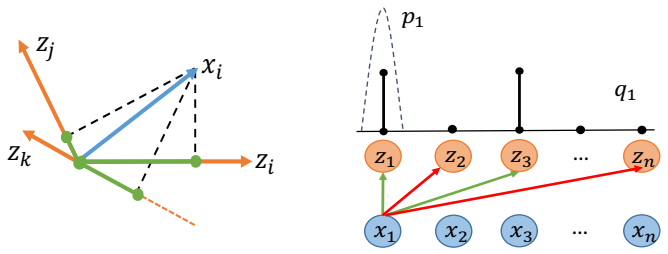
Given a mini-batch with n image and text samples, for each image $\boldsymbol{x}_i$ the image-text pairs are constructed as $\{(\boldsymbol{x}_i, \boldsymbol{z}_j), y_{i,j}\}_{j=1}^{n}$, where $y_{i,j} = 1$ means that $(\boldsymbol{x}_i, \boldsymbol{z}_j)$ is a matched pair, while $y_{i,j} = 0$ indicates the unmatched ones. The probability of matching $\boldsymbol{x}_i$ to $\boldsymbol{z}_j$ is defined as

$$p_{i,j} = \frac{\exp(\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_j)}{\sum_{k=1}^{n} \exp(\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_k)} \quad s.t.\ \bar{\boldsymbol{z}}_j = \frac{\boldsymbol{z}_j}{\|\boldsymbol{z}_j\|} \tag{1}$$

where $\bar{\boldsymbol{z}}_j$ denotes the normalized text feature. Geometrically $\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_j$ represents the scalar projection image feature $\boldsymbol{x}_i$ onto text feature $\boldsymbol{z}_j$ and $p_{i,j}$ can be viewed as the percent of scalar projection of $(\boldsymbol{x}_i, \boldsymbol{z}_j)$ among all pairs $\{(\boldsymbol{x}_i, \boldsymbol{z}_j)\}_{j=1}^{n}$ in a mini batch. Fig. 2 (a) shows the geometrical explanation of the cross-modal projection. We can see that the more similar image feature to text feature, the larger the scalar projection would be. Note that the scalar projection can be negative if the two vectors lie in opposite directions, such as $\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_k$ shown in the figure.

Considering the fact that there might be more than one matched text samples for $\boldsymbol{x}_i$ in a mini-batch, we normalize the true matching probability of $(\boldsymbol{x}_i, \boldsymbol{z}_j)$ as

$$q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{n} y_{i,k}} \tag{2}$$



(a) cross-modal projection (b) cross-modal matching

Fig. 2. Interpretation of cross-modal projection and matching. (a) The image feature $\boldsymbol{x}_i$ is projected onto different text directions, and the scalar projection of $\boldsymbol{x}_i$ onto the matched text $\boldsymbol{z}_i$ is larger than that of unmatched text $\boldsymbol{z}_j$ and $\boldsymbol{z}_k$. (b) For the image $\boldsymbol{x}_1$ with $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$ as matched candidates (green arrowed line) in a mini-batch, and the other texts as unmatched samples (red arrowed line), the CMPM loss attempts to find a distribution $\boldsymbol{p}_1$ having low probability where the true matching distribution $\boldsymbol{q}_1$ has low probability

The matching loss of associating $\boldsymbol{x}_i$ with correctly matched text samples is defined as

$$\mathcal{L}_i = \sum_{j=1}^{n} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}+\epsilon} \tag{3}$$

where ϵ is a small number to avoid numerical problems, and the matching loss from image to text in a mini-batch is computed by

$$\mathcal{L}_{i2t} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i \tag{4}$$

Note that Eq. 3 actually represents the KL divergence from distribution $\boldsymbol{q}_i$ to $\boldsymbol{p}_i$, and minimizing $KL(\boldsymbol{p}_i \| \boldsymbol{q}_i)$ attempts to select a $\boldsymbol{p}_i$ that has low probability where $\boldsymbol{q}_i$ has low probability [4]. Fig. 2 (b) illustrates the proposed matching loss with a mini-batch data, we can see that the true matching distribution $\boldsymbol{q}_1$ for image $\boldsymbol{x}_1$ has multiple modes with more than one matched text candidates in the mini batch, and the proposed matching loss attempts to select a single mode distribution $\boldsymbol{p}_1$ to avoid putting probability mass in the low-probability areas between modes of $\boldsymbol{q}_1$, such that the compatibility of the unmatched image-text pairs are minimized while the relevance of the matched pairs are maximized. Note that given an image, all the positive and negative text candidates in a mini-batch are taken into consideration for computing the matching loss, getting rid of the dedicated sampling procedures in traditional bi-directional ranking loss.

It might raise the concerns about using $KL(\boldsymbol{q}_i \| \boldsymbol{p}_i)$ to maximize the compatibility of matched pairs for learning discriminative embeddings. As explained in [4], $KL(\boldsymbol{q}_i \| \boldsymbol{p}_i)$ would try to find $\boldsymbol{p}_i$ as a blur mode, towards generating high probability where $\boldsymbol{q}_i$ has high probability. This may cause difficulties for distinguishing matched and unmatched pairs when there are multiple positive pairs in a mini-batch. The advantages of $KL(\boldsymbol{p}_i \| \boldsymbol{q}_i)$ over $KL(\boldsymbol{q}_i \| \boldsymbol{p}_i)$ will be further demonstrated in experiments.

In image-text embedding learning, the matching loss is often computed in two directions [39, 40, 21]: the image-to-text matching loss requires the matched text to be closer to the image than unmatched ones, and in verse the text-to-image matching loss constrains the related text to rank before unrelated ones. Similarly, the matching loss $\mathcal{L}_{t2i}$ from text to image can be formulated by exchanging $\boldsymbol{x}$ and $\boldsymbol{z}$ in Eq. 1−4, and the bi-directional CMPM loss is calculated by

$$\mathcal{L}_{cmpm} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \tag{5}$$

3.3 Cross-Modal Projection Classification

For image-text matching with identity-level annotations, the classification loss applied to each modality helps to learn more discriminative features. However, the matching relationships of image-text pairs may not be sufficiently exploited in separate classification tasks. In this section, we develop a novel classification function where the cross-modal projection is integrated into the norm-softmax loss to further enhance the compactness of the matched embeddings.

Norm-softmax First we revisit the traditional softmax loss by looking into the decision criteria of softmax classifiers. Given the extracted image features $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ from visual CNN, text features $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ from Bi-LSTM, and the label set $\mathcal{Y} = \{y_i\}_{i=1}^N$ from M classes, the original softmax loss for classifying images can be computed as

$$\mathcal{L}_{softmax} = \frac{1}{N}\sum_i -log(\frac{exp(\boldsymbol{W}_{y_i}^\top \boldsymbol{x}_i + b_{y_i})}{\sum_j exp(\boldsymbol{W}_j^\top \boldsymbol{x}_i + b_j)}) \tag{6}$$

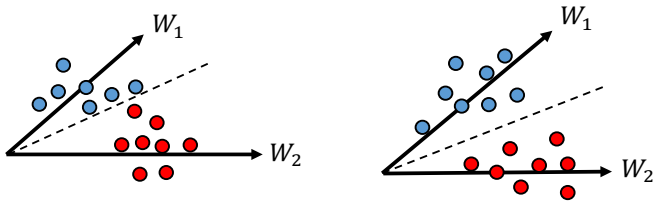
where y_i indicates the label of $\boldsymbol{x}_i$, $\boldsymbol{W}_{y_i}$ and $\boldsymbol{W}_j$ represent the y_i-th and j-th column of weight matrix $\boldsymbol{W}$, and b_{y_i} and b_j respectively denote the y_i-th and j-th element of bias vector $\boldsymbol{b}$.

To improve the discriminative ability of the image feature $\boldsymbol{x}_i$ during classification, we impose weight normalization on the softmax loss as with [37, 19], and reformulate Eq. 6 as

$$\mathcal{L}_{image} = \frac{1}{N}\sum_i -log(\frac{exp(\boldsymbol{W}_{y_i}^\top \boldsymbol{x}_i)}{\sum_j exp(\boldsymbol{W}_j^\top \boldsymbol{x}_i)}) \quad s.t.\ \|\boldsymbol{W}_j\| = 1 \tag{7}$$

Compared with the original softmax loss, the norm-softmax loss normalizes all the weight vectors into the same length in order to reduce the impact of weight magnitude in distinguishing different samples. Here we omit the bias $\boldsymbol{b}$ for simplifying analysis and in fact found it makes no difference as with [20, 19].

The intuitive explanation of the norm-softmax loss is shown in Fig. 3. We can see that, for the original softmax, the classification results depends on $\|\boldsymbol{W}_k\|\|\boldsymbol{x}\|\cos(\theta_k), (k = 1, 2)$, where θ_k indicates the angle between $\boldsymbol{x}$ and $\boldsymbol{W}_k$. For the norm-softmax, all the weight vectors are normalized into the same length, and the classification results can be only depended on $\|\boldsymbol{x}\|\cos(\theta_k)$. This restriction encourages the feature $\boldsymbol{x}$ to distribute more compactly along the weight vector in order to be correctly classified.



(a) softmax loss (b)norm-softmax loss

Fig. 3. Geometric interpretation of softmax and norm-softmax

Cross-Modal Projection In this paper, we attempt to classify the projection of image features onto the corresponding text features instead of categorizing the original feature representations. The cross-modal projection integrates the image-text similarity into classification and thus strengthens the association within matched pairs.

By incorporating the cross-modal projection into the norm-softmax, we can reformulated Eq. 7 as

$$\mathcal{L}_{ipt} = \frac{1}{N}\sum_i -log(\frac{exp(\boldsymbol{W}_{y_i}^\top \hat{\boldsymbol{x}}_i)}{\sum_j exp(\boldsymbol{W}_j^\top \hat{\boldsymbol{x}}_i)}) \quad s.t.\ \|\boldsymbol{W}_j\| = r,\ \hat{\boldsymbol{x}}_i = \boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_i \cdot \bar{\boldsymbol{z}}_i \tag{8}$$

where $\hat{\boldsymbol{x}}_i$ denotes the vector projection of image feature $\boldsymbol{x}_i$ onto normalized text feature $\bar{\boldsymbol{z}}_i$. Intuitively, all the matched text samples needs to lie in the direction of $\boldsymbol{W}_{y_i}$ for the image feature $\boldsymbol{x}_i$ to project onto, in order to promote correct categorization. The text classification loss function can be written as

$$\mathcal{L}_{tpi} = \frac{1}{N}\sum_i -log(\frac{exp(\boldsymbol{W}_{y_i}^\top \hat{\boldsymbol{z}}_i)}{\sum_j exp(\boldsymbol{W}_j^\top \hat{\boldsymbol{z}}_i)}) \quad s.t.\ \|\boldsymbol{W}_j\| = r,\ \hat{\boldsymbol{z}}_i = \boldsymbol{z}_i^\top \bar{\boldsymbol{x}}_i \cdot \bar{\boldsymbol{x}}_i \tag{9}$$

The final CMPC loss can be calculated with

$$\mathcal{L}_{cmpc} = \mathcal{L}_{ipt} + \mathcal{L}_{tpi} \tag{10}$$

3.4 Objective Functions

For matching tasks with only pairwise correspondence, we can utilize the proposed CMPM loss for learning discriminative image-text embeddings. If identity labels are available, we adopt the joint of the proposed CMPM loss and CMPC loss for more accurately associating the cross-modal representations. The overall objective function is formulated as

$$\mathcal{L} = \mathcal{L}_{cmpm} + \mathcal{L}_{cmpc} \tag{11}$$

At the test stage, given an image and text, we first extract the image feature $\boldsymbol{x}$ and text feature $\boldsymbol{z}$ with the visual CNN and Bi-LSTM network, respectively. Then the cosine distance between $\boldsymbol{x}$ and $\boldsymbol{z}$ is computed for image-to-text and text-to-image retrieval evaluation.

4 Experiments

4.1 Datasets and Settings

Datasets Five datasets are used in our experiments. The *Flickr30K* [45] dataset contains 31,783 images with each one annotated by five text descriptions. We adopt the data split in [12] to use 29,783 images for training, 1,000 images for validation, and 1,000 images for testing. The *MSCOCO* [17] dataset consists of 12,3287 images and each one is also described by five sentences. Following the protocol of [12], we split the data into 82,783 training, 30,504 validation, and 5,000 test images, and report the evaluation results on both 5K and 1K (5 fold) test images. The *CUHK-PEDES* [16] dataset contains 40,206 pedestrian images of 13,003 identities, with each image described by two textual descriptions. The dataset is split into 11,003 training identities with 34,054 images, 1000 validation persons with 3,078 images and 1000 test individuals with 3,074 images. The *Caltech-UCSD Birds (CUB)* [28] dataset consists of 11,788 bird images from 200 different categories. Each image is labelled with 10 visual descriptions. The dataset is split into 100 training, 50 validation and 50 test categories. The *Oxford-102 Flowers (Flowers)* [28] dataset contains 8,189 flower images of 102 different categories, and each image has 10 textual descriptions. The data splits provide 62 training, 20 validation, and 20 test categories.

Evaluation Metrics We adopt Recall@K (K=1, 5, 10) [12] and AP@50 [28] for retrieval evaluation. Recall@K (or R@K) indicates the percentage of the queries where at least one ground-truth is retrieved among the top-K results, and AP@50 represents the percent of top-50 scoring images whose class matches that of the text query, averaged over all the test classes.

Implementation Details All the models are implemented in TensorFlow with a NVIDIA GEFORCE GTX 1080 GPU. For all the datasets, we use MobileNet [9] and Bi-LSTM for learning visual and textual features, respectively. The adam optimizer [13] is employed for optimization with $lr = 0.0002$. For Flickr30K and MSCOCO, we also report the results with ResNet-152 [7] as image feature extractor, where we start training with $lr = 0.0002$ for 15 epochs with fixed image encoder and then training the whole model with $lr = 0.00002$ for 30 epochs.

4.2 Results on the Flickr30K dataset

We summarize the comparison of retrieval results on the Filckr30K dataset in Table 1. We can see that with MobileNet as image encoder, the proposed CMPM loss achieves competitive results of R@1=37.1% for image-to-text retrieval, and R@1=29.1% for text-to-image retrieval. The performance can be improved to 48.3% and 35.7% respectively by employing ResNet-152 as with RRF-Net [21] and DAN [26]. We also explore the assistant effect of the CMPC loss by training the classifiers single category per image, and we observe that the retrieval results can be further improved by around 1.3%, demonstrating the effectiveness of cross-modal projection learning for image-text matching.

Table 1. Comparison of bi-directional retrieval results (R@K(%)) on Flickr30K

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
DCCA [44]	16.7	39.3	52.9	12.6	31.0	43.0
DVSA [12]	22.2	48.2	61.4	15.2	37.7	50.5
m-CNN [22]	33.6	64.1	74.9	26.2	56.3	69.6
VQA-A [18]	33.9	62.5	74.5	24.9	52.6	64.8
DSPE [39]	40.3	68.9	79.9	29.7	60.1	72.1
sm-LSTM [11]	42.5	71.9	81.5	30.2	60.4	72.3
RRF-Net [21]	47.6	77.4	87.1	35.4	68.3	79.9
DAN [26]	**55.0**	**81.8**	**89.0**	**39.4**	**69.2**	**79.1**
CMPM (MobileNet)	37.1	65.8	76.3	29.1	56.3	67.7
CMPM+CMPC (MobileNet)	40.3	66.9	76.7	30.4	58.2	68.5
CMPM (ResNet-152)	48.3	75.6	84.5	35.7	63.6	74.1
CMPM +CMPC (ResNet-152)	49.6	76.8	86.1	37.3	65.7	75.5

4.3 Results on the MSCOCO dataset

We compare the proposed appproach with state-of-the-art methods on the MS-COCO dataset in Table 2. We can see that for 1K test images the proposed CMPM loss achieves R@1=56.1% and 44.6% with image and text as quires, respectively. For 5K test images the algorithm achieves R@1=31.1% and 22.9%, outperforming the second best by 7.0% and 5.3%, which further verifies the superiority of the proposed loss functions.

4.4 Results on the CUHK-PEDES dataset

Table 3 compares the proposed method against existing approaches on the CUHK-PEDES dataset. We can see that the proposed CMPM loss achieves 44.02% of R@1 and 77.00% of R@10, outperforming the second best performer [15] by a large margin. When we add the CMPC loss supervised by the identity-level annotations, the text-to-image retrieval performance is further improved to 49.37% for R@1 and 79.27% for R@10. This illustrates the effectiveness of the CMPM loss for person search applications, and the promotion effect of the CMPC loss when the category labels are available in real applications.

4.5 Results on the CUB and Flowers dataset

The comparison of image-to-text and text-to-image retrieval results on the CUB and Flowers dataset is shown in Table 4. Consider that the bi-directional losses are implemented in our approach, we choose the symmetric results [15] of the existing methods for fair comparison. We can see that the proposed algorithm outperforms the state-of-the-art, achieving 64.3% of R@1 for image-to-text retrieval and 67.9% of AP@50 for text-to-image retrieval on CUB, and reporting the best R@1 of 68.90% for image-to-text retrieval and the second best AP@50 of 69.70% for text-to-image retrieval on Flowers.

Table 2. Comparison of bi-directional retrieval results (R@K(%)) on MSCOCO

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
1K test images						
DVSA [12]	38.4	69.9	80.5	27.4	60.2	74.8
GMM-FV [14]	39.4	67.9	80.9	25.1	59.8	76.6
m-CNN [22]	42.8	73.1	84.1	32.6	68.6	82.8
VQA-A [18]	50.5	80.1	89.7	37.0	70.9	82.9
DSPE [39]	50.1	79.7	89.2	39.6	75.2	86.9
sm-LSTM [11]	53.2	83.1	91.5	40.7	75.8	87.4
RRF-Net [21]	**56.4**	85.3	91.5	43.9	78.1	88.6
CMPM (MobileNet)	51.4	80.8	89.8	40.9	73.9	85.2
CMPM+CMPC (MobileNet)	52.9	83.8	92.1	41.3	74.6	85.9
CMPM (ResNet-152)	56.1	**86.3**	**92.9**	**44.6**	**78.8**	**89.0**
5K test images						
DVSA [12]	16.5	39.2	52.0	10.7	29.6	42.2
GMM-FV [14]	17.3	39.0	50.2	10.8	28.3	40.1
VQA-A [18]	23.5	50.7	63.6	16.7	40.5	53.8
CMPM (MobileNet)	23.9	51.5	65.4	18.9	43.8	56.9
CMPM+CMPC (MobileNet)	24.6	52.3	66.4	19.1	44.6	58.4
CMPM (ResNet-152)	**31.1**	**60.7**	**73.9**	**22.9**	**50.2**	**63.8**

Table 3. Comparison of text-to-image retrieval results (R@K(%)) on CUHK-PEDES

Method	Text-to-Image	
	R@1	R@10
deeper LSTM Q+norm I [1]	17.19	57.82
iBOWIMG [46]	8.00	30.56
NeuralTalk [35]	13.66	41.72
Word CNN-RNN [28]	10.48	36.66
GNA-RNN [16]	19.05	53.64
GMM+HGLMM [14]	15.03	42.27
Latent Co-attention [15]	25.94	60.48
CMPM	**44.02**	**77.00**
CMPM+CMPC	**49.37**	**79.27**

5 Ablation Studies

To investigate the effect of each component of the proposed CMPM and CMPC loss, we perform a series of ablation studies on the CUHK-PEDES dataset. We conduct further comparative experiments in three aspects: comparison of the CMPM loss with other matching losses under various batch size, impact of cross-modal projection and weight normalization for the CMPC loss, and the cross-modal feature distribution learned with different losses.

Table 4. Comparison of image-to-text (R@K(%)) and text-to-image (AP@K(%)) retrieval results on the CUB and Flowers dataset

Method	CUB		Flowers	
	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image
	R@1	AP@50	R@1	AP@50
Bow [6]	44.1	39.6	57.7	57.3
Word2Vec [25]	38.6	33.5	54.2	52.1
Word CNN [28]	51.0	43.3	60.7	56.3
Word CNN-RNN [28]	56.8	48.7	65.6	59.6
GMM+HGLMM [14]	36.5	35.6	54.8	52.8
Triplet [15]	52.5	52.4	64.3	64.9
Latent Co-attention [15]	61.5	57.6	68.4	**70.1**
CMPM	**62.1**	**64.6**	66.1	67.7
CMPM+CMPC	**64.3**	**67.9**	**68.9**	69.7

5.1 Analysis of Cross-Modal Matching

Table 5 compares the proposed CMPM loss with the commonly used bi-directional ranking (Bi-rank) loss [39, 40, 21], the most similar N-pair loss [30], and Histogram Loss [34] with different batch size on the CUHK-PEDES dataset. We add the image-to-text retrieval evaluation for more comprehensive analysis of learned embeddings, since good cross-modal embeddings should be able to perform bi-directional matching tasks. Note that all the loss functions are implemented in the bi-directional mode and the triplets are online sampled.

Table 5. R@1 (%) comparison of cross-modal matching functions with different batch size on the CUHK-PEDES dataset

Matching Loss	Text-to-Image				Image-to-Text			
	16	32	64	128	16	32	64	128
Bi-rank [21]	31.11	37.85	42.11	41.42	32.56	41.28	47.46	46.88
Histogram [34]	14.68	19.20	21.70	21.31	4.78	13.53	13.04	2.88
N-pair [30]	34.57	45.55	45.68	39.33	17.66	13.66	12.07	10.83
$KL(\boldsymbol{q}_i \| \boldsymbol{p}_i)$	42.58	43.81	41.89	36.06	41.87	38.81	22.35	19.97
CMPM	42.28	43.42	44.02	42.43	51.95	52.09	51.98	48.67

From the table we can see that the previous matching loss fluctuates greatly when the batch size varies between 16 and 128. The bi-directional ranking loss depends on larger batch size to generate comparative matching accuracies, due to the negative sampling requirements [29]. The Histogram loss [34] performs much worse than other methods for cross-modal matching. The N-pair loss [30] produce better text-to-image retrieval results with moderate batch size, while the image-to-text matching performance are much worse. This might due to the

scalar gap of image and text embeddings from different networks. The $KL(\boldsymbol{q}_i \| \boldsymbol{p}_i)$ discussed in Section 3.2 generates satisfying results when the batch size is small, while deteriorates with larger batch size of 128. This further verifies the analysis that, when there are more positive pairs in larger mini batches, the inappropriate KL direction blurring the multiple modes could cause ambiguities for image-text matching. In contrast, the proposed CMPM loss produces much more stable matching results with different batch size (R@1 remains above 42% for text-to-image retrieval), and the advantages are more obvious when the batch size are too small or too large, exhibiting great superiority and broad applicability.

5.2 Analysis of Cross-Modal Classification

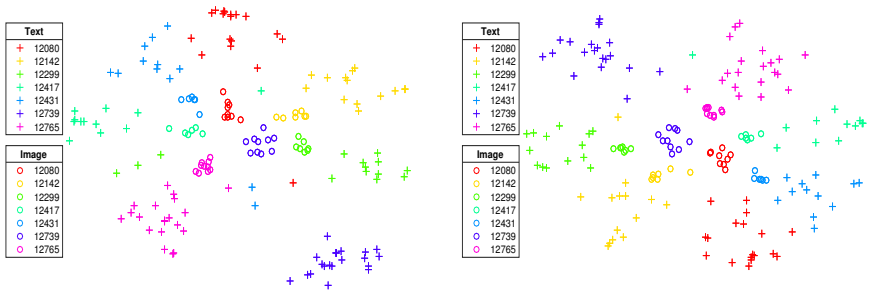
Table 6. R@1 (%) comparison of different components of the cross-modal projection learning on the CUHK-PEDES dataset

Matching	Classification			Text-to-Image		Image-to-Text	
CMPM	softmax	normW	CMP	R@1	R@10	R@1	R@10
✓	✗	✗	✗	44.02	77.00	51.98	87.02
✓	✓	✗	✗	45.38	78.43	55.14	89.30
✓	✓	✓	✗	47.12	78.38	56.51	90.50
✓	✓	✗	✓	46.95	79.40	55.82	89.17
✓	✓	✓	✓	49.37	79.45	57.71	91.28
✗	✓	✗	✓	16.93	40.90	17.63	43.98
✗	✓	✓	✓	42.25	73.29	50.72	85.95

Table 6 illustrates the impact of the softmax loss, weight normalization (normW) and cross-modal projection (CMP) in image-text embedding learning on the CUHK-PEDES dataset. We can see that adding the supervision loss indeed improves the matching performance, while the original softmax loss offers limited assistance. By adding the weight normalization, the R@1 rates are increased from 45.38% to 47.12% for image-to-text retrieval, and 55.14% to 56.51% for text-to-image retrieval. The cross-modal projection further improves the bi-directional retrieval results by 2.25% and 1.20%. We also notice that the CMPC loss alone achieves competitive results for image-text matching and weight normalization brings significant improvements. This indicates the effectiveness of weight normalization and cross-modal projection in learning discriminative cross-modal representations.

5.3 Feature Visualization

To better understand the effect of the proposed cross-modal matching loss and cross-modal classification loss for learning discriminative image-text embeddings, we show the t-SNE [24] visualization the test feature distribution learned using



(a) CMPM (b) CMPM + CMPC

Fig. 4. Comparison of feature distribution learned with the proposed approach

the CMPM loss and CMPM+CMPC loss on the CUHK-PEDES dataset. From Fig. 4 (a) we can see that the CMPM loss learns image-text embeddings distributed along radial spokes, where the image and text features from the same class approximately lie in the same direction. This type of angular distribution is consistent with the traditional softmax loss [19], and therefore the added CMPC loss naturally improves the compactness of the features along each spoke as shown in Fig. 4 (b). We can also observe that the radius of image feature areas is smaller than text features, which indicates the scalar gap brought by different networks (i.e., the CNN network for image and Bi-LSTM for text). In experiments we obtain the average length (value of ℓ_2 norm) of 52.62 for image features and 128.92 for text features. The cross-modal distribution shows the importance of feature normalization in cross-modal projection for bridging the scalar gap in image-text embedding learning.

6 Conclusions

In this paper, we proposed a novel cross-modal projection matching loss (CMPM) and cross-modal projection classification (CMPC) loss, for learning deep discriminative image-text embeddings. The CMPM loss utilize the KL divergence to minimize the compatibility score of the unmatched image-text pairs while maximizing the relevance between the matched ones. It shows great stability and superiority for associating image and text under various batch size, without triplet sampling and margin selection that hampers the traditional bi-directional ranking loss. The CMPC loss incorporates the matching relationship into the auxiliary classification task, which further enhances the representation compactness of each category. In the future, we will work on how to better interact the matching task and classification task in identity-aware matching problems.

Acknowledgements. This work was supported by the Natural Science Foundation of China under Grant 61725202, 61751212, 61771088, 61632006 and 91538201.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: ICCV. pp. 2425–2433 (2015)
2. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. In: CVPR. pp. 1320–1329 (2017)
3. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. arXiv: 1801.07698 (2018)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
5. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR. pp. 1735–1742 (2006)
6. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv: 1703.07737 (2017)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861 (2017)
10. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: CVPR. pp. 4555–4564 (2016)
11. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: CVPR. pp. 7254–7262 (2017)
12. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137 (2015)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv: 1412.6980 (2014)
14. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: CVPR. pp. 4437–4446 (2015)
15. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: ICCV. pp. 1908–1917 (2017)
16. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: CVPR. pp. 5187–5196 (2017)
17. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
18. Lin, X., Parikh, D.: Leveraging visual question answering for image-caption ranking. In: ECCV. pp. 261–277 (2016)
19. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR. pp. 6738–6746 (2017)
20. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. pp. 507–516 (2016)
21. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: ICCV. pp. 4127–4136 (2017)
22. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV. pp. 2623–2631 (2015)
23. Ma, Z., Lu, Y., Foster, D.P.: Finding linear structure in large datasets with scalable canonical correlation analysis. In: ICML. pp. 169–178 (2015)

24. van der Maaten, L.: Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research **15**(1), 3221–3245 (2014)
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
26. Nam, H., Ha, J., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: CVPR. pp. 2156–2164 (2017)
27. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv: 1703.09507 (2017)
28. Reed, S.E., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR. pp. 49–58 (2016)
29. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
30. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NIPS. pp. 1849–1857 (2016)
31. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR. pp. 4004–4012 (2016)
32. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10, 000 classes. In: CVPR. pp. 1891–1898 (2014)
33. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. pp. 1701–1708 (2014)
34. Ustinova, E., Lempitsky, V.S.: Learning deep embeddings with histogram loss. In: NIPS. pp. 4170–4178 (2016)
35. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. PAMI **39**(4), 652–663 (2017)
36. Wang, F., Liu, W., Liu, H., Cheng, J.: Additive margin softmax for face verification. arXiv: 1801.05599 (2018)
37. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L_2 hypersphere embedding for face verification. arXiv: 1704.06369 (2017)
38. Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. arXiv: 1801.09414 (2018)
39. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR. pp. 5005–5013 (2016)
40. Wang, L., Li, Y., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. arXiv: 1704.03470 (2017)
41. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. pp. 499–515 (2016)
42. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. pp. 1249–1258 (2016)
43. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
44. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: CVPR. pp. 3441–3450 (2015)
45. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
46. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv: 1512.02167 (2015)