Multiplicative Angular Margin Loss for Text-Based Person Search

Peng Zhang[1], Deqiang Ouyang[1], Feiyu Chen[1,2], and Jie Shao[1,2*]
[1]Center for Future Media, School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, China
[2]Sichuan Artificial Intelligence Research Institute, Yibin, China
{pengzhang,ouyangdeqiang,chenfeiyu}@std.uestc.edu.cn,shaojie@uestc.edu.cn

ABSTRACT

Text-based person search aims at retrieving the most relevant pedestrian images from database in response to a query in form of natural language description. Existing algorithms mainly focus on embedding textual and visual features into a common semantic space so that the similarity score of features from different modalities can be computed directly. Softmax loss is widely adopted to classify textual and visual features into a correct category in the joint embedding space. However, softmax loss can only help classify features but not increase the intra-class compactness and inter-class discrepancy. To this end, we propose multiplicative angular margin (MAM) loss to learn angularly discriminative features for each identity. The multiplicative angular margin loss penalizes the angle between feature vector and its corresponding classifier vector to learn more discriminative feature. Moreover, to focus more on informative image-text pair, we propose pairwise similarity weighting (PSW) loss to assign higher weight to informative pairs. Extensive experimental evaluations have been conducted on the CUHK-PEDES dataset over our proposed losses. The results show the superiority of our proposed method. Code is available at https://github.com/pengzhanguestc/MAM_loss.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

person search, multiplicative angular margin, pair weighting

1 INTRODUCTION

Text-based person search, as a newly emerging task, is originated from person re-identification [4, 14, 19]. Given a query of fine-grained description of pedestrian appearance in the form of natural language, text-based person search aims to retrieve the most relevant person images from a large-scale image database, as illustrated in Figure 1. Text-based person search is believed to be closer to practical applications in contrast to person re-identification, considering the fact that it is normally easier to have access to language description than picture information of a person. Recent years have witnessed the rapid development of person re-identification. Owing to the remarkable feature extraction ability of deep convolution neural network (CNN) and discriminative learning algorithms, the performance of person re-identification has been improved to an unprecedent level. However, progress that has been made in text-based person search is barely satisfactory.

Figure 1: Given a textual description, language person search aims at retrieving the most relevant person images from an image dataset.

The main challenge of text-based person search lies in the heterogenous semantic gap between textual feature and visual feature. During deployment, we need to compute the distance between the probe text feature and the gallery image feature. However, it makes no sense for us to directly compute the feature distance of image and text because they are from two totally different semantic spaces which means that the feature similarity may be not related to whether they match or not. Many algorithms have been devoted to learning a modality-invariant and discriminative feature for each identity in a shared feature space [28–30].

*Corresponding author: Jie Shao.

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446314

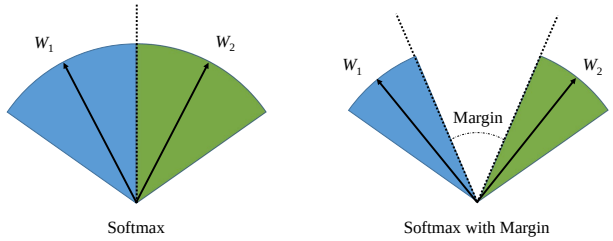

Figure 2: A geometrical interpretation of the effect that angular margin has on softmax loss. W_1 and W_2 are two weight vectors of the last fully connected layer corresponding to two different classes. Different color areas indicate feature space for different classes. Obviously, feature region in softmax with margin is relatively compact.

Among those algorithms, category classification loss is often adopted as auxiliary task to learn more discriminative joint image-text embeddings. Through classifying image feature and text feature of the same identity into the same category, we can indirectly increase the similarity of matched image-text pair. Intuitively, simultaneously maximizing the intra-class compactness and inter-class discrepancy in the shared latent space would lead to better learned features. Looking into softmax loss, we can observe that it is composed of a fully connected layer, a softmax function and a cross-entropy loss and does not explicitly increase intra-class compactness and inter-class discrepancy. The fully connected layer in softmax loss actually plays the role of a linear classifier and the final probability distribution of class for each feature depends on its inner product with each class weight vector in the last fully connected layer. It is worth noting that the value of inner product can be factorized into magnitude of vector and angular cosine. In the light of this, many face recognition algorithms [5, 17, 18, 25] propose to impose angular margin on learned features as shown in Figure 2. However, in the task of person search, there are no relevant methods to impose angular margin on the softmax loss. Compared with face recognition, how to incorporate margin into softmax loss while taking both visual and textual features into consideration is where the challenge lies. In this paper, we propose multiplicative angular margin loss to deal with this problem.

In the filed of pedestrian-related task, deep metric learning has been widely applied and in some areas such as face recognition and person re-identification, contrastive loss [6] and triplet loss [20] have demonstrated the impressive capability of improving model performance. However, triplet loss appears to have little effect on text-based person search. By examining the formulas of pair-based metric loss such as contrastive loss and triplet loss, we observe that the coefficients of positive pair or negative pair are equal and this seems to be unreasonable. For example, there may exist some abnormal image-text pairs which are matched but have low cosine similarity scores or are unmatched but have high cosine similarity scores. These abnormal pairs are always more informative and valuable. Obviously, abnormal pairs are supposed to be paid more attention to than those normal pairs. Motivated by these analyses, we consider improving pair-based loss and setting higher weights for those informative pairs.

In short, our contributions of this paper can be summarized into three folds:

- For the task of text-based person search, we specially design a novel multiplicative angular margin loss which aims to enlarge the inter-class angular margin and condense the intra-class angular margin.
- We propose pairwise similarity weighting loss in text-based person search by setting different weights to image-text pairs. The more informative a pair is, a larger weight it will be assigned to.
- Through conducting extensive experiments, we show the effectiveness of our proposed method on the CUHK-PEDES dataset.

2 RELATED WORK

2.1 Text-based Person Search

Compared with other tasks of person search in terms of query type, text-based person search is the most similar to attribute-based person search [23, 24, 26]. Attribute-based person search predefines a set of attributes that would limit the capacity of describing person appearance, which makes it less practical than text-based person search. Textual descriptions are more accessible because it can describe person of interest in a more natural way. A critical procedure in text-based person search is to measure the similarity between person image and textual description. Existing algorithms either attempt to design a similarity learning network or focus on embedding both visual and textual features into a common latent space [29]. Similarity learning network puts emphasis on predicting a similarity score of the input text and image while joint embedding learning tries to eliminate the semantic gap caused by different data distributions existed in image and text data. In [12], Li *et al.* developed a method to measure the relevancy degree of each word with different visual feature units and eventually predict the similarity score between the whole sentence and image. In [11], Li *et al.* proposed an identity-aware two-stage framework and designed a novel cross-modal cross-entropy (CMCE) loss to fully exploit identity information. In [30], Zheng *et al.* for the first time employed a CNN to extract discriminative textual feature and proposed an instance loss under the assumption that each identity can be viewed as one class. In [29], Zhang *et al.* proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss to learn discriminative image-text embeddings. In the above mentioned methods, most of them adopted identity classification loss as an effective auxiliary task but ignored adjusting intra-class distance and inter-class distance. In this paper, we propose multiplicative angular margin loss to enlarge inter-class angular margin while condensing the intra-class one so as to learn more discriminative image-text embeddings.

2.2 Metric Learning

In many retrieval-related tasks such as person re-identification, face recognition and person search, it is widely acknowledged that preserving an appropriate distance relationship between pairs of samples plays an indispensable role in improving model performance. Generally speaking, deep metric loss can be roughly divided into two categories, namely softmax losses and pairwise

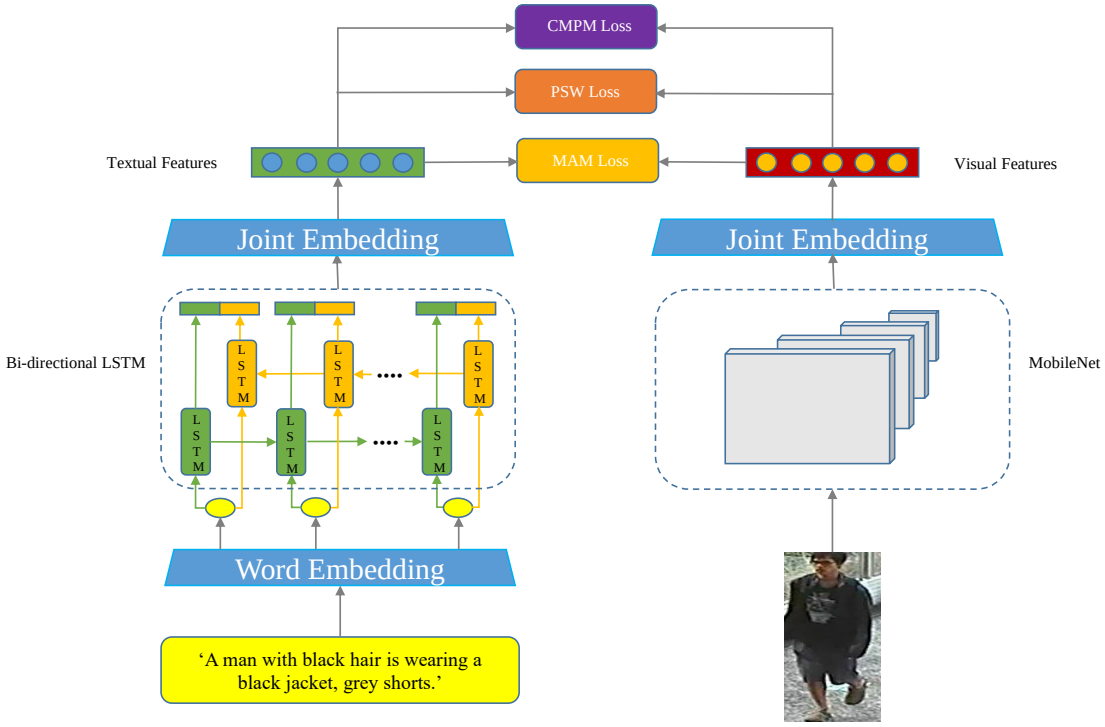

Figure 3: Overall architecture of our proposed model. It consists of a visual feature extractor, a textual feature extractor and joint embedding learning modules for visual feature and textual feature respectively. In addition, we specially present where our proposed loss functions work in the overall architecture.

metric losses [15]. Softmax loss, as mentioned before, consists of a fully connected layer, a softmax function and a cross-entropy loss. For the purpose of learning discriminative features, many researchers have been working on reformulating softmax loss to minimize intra-class variance and maximize inter-class variance. In [18], Liu *et al.* proposed a large margin softmax (L-softmax) to improve feature discrimination by introducing an angular margin to each identity. Angular softmax (A-softmax) [17], as an improved version of L-softmax, normalizes the weights of fully connected layer and achieves better performance with less training time for model convergence. In [25], Wang *et al.* refined softmax loss by normalizing both feature vectors and weight vectors, based on which a cosine margin is introduced. In [5], Deng *et al.* proposed an additive angular margin (ArcFace) loss to learn highly discriminative features by adding an additive angular margin to the target angle. We can observe that all the above softmax-based losses are proposed to solve single-modality task. However, text-based person search is concerned with data of two different modalities. In this paper, we specially design a multiplicative angular margin loss for text-based person search.

Pairwise metric loss aims at preserving an reasonable distance structure for pair samples including matched pair and unmatched ones. Some of the most well-known pairwise metric losses include contrastive loss [6], triplet loss [20] and triplet-center loss [7]. Both contrastive loss and triplet loss require a hard example mining strategy, which is both time-consuming and performance-sensitive. Multi-similarity (MS) [27] loss pioneered the concept of pair weighting. Both lifted structure loss [22] and N-pair loss [21] introduced effective pair weighting schemes. In this paper, a pairwise similarity weighting scheme is introduced specifically for text-based person search, which assigns a larger weight to a more informative image-text pair.

3 OUR METHOD

In this section, we elaborate on our proposed method as shown in Figure 3. First, we formulate the task of text-based person search. Then, we introduce multiplicative margin loss and our proposed pairwise cosine similarity weighting scheme. Finally, we give the overall objection function of our model.

3.1 Network Architecture

Figure 3 depicts the framework of our proposed method. From the figure, we observe that our model can be divided into three components including a visual feature extractor, a textual feature extractor and a joint embedding learning module for eliminating the semantic gap between textual and visual feature. Textual feature extractor will do some data preprocessing work including sentence tokenization and splitting and then extract the initial textual feature with a bi-directional long short-term memory (Bi-LSTM) model [31]. As for image, we adopt a pre-trained MobileNet model [8] as our visual extractor and finally obtain its initial feature from the last

pooling layer. To embed textual and visual features into a shared latent space, we employ a fully connected layer to project the initial feature into a joint embedding one. In this work, we mainly focus on joint embedding learning module and are dedicated to designing effective loss function so as to learn a discriminative feature embedding for each identity. In the following, we will describe our proposed multiplicative angular margin (MAM) loss and pairwise similarity weighting (PSW) loss.

3.2 Multiplicative Angular Margin Loss

In a mini-batch of N person images and the corresponding textual descriptions, we indicate the image-text feature pairs which are obtained from joint embedding learning module as $\left\{\left(\boldsymbol{x}_i, \boldsymbol{z}_j\right), y_{i,j}\right\}_{j=1}^{n}$ where $y_{i,j} = 1$ means whether this pair match or not, with 1 for matched pair and 0 for unmatched one. Before introducing our proposed multiplicative angular margin loss, we start by presenting the frequently used softmax loss in our text-based person search task. Given a visual feature $\boldsymbol{x}_i$ and its corresponding label y_i, we formulate the softmax loss as:

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right), \tag{1}$$

where $\boldsymbol{W}_j$ is the weight vector of class j in fully connected layer of softmax loss, y_i indicates the ground-truth class of $\boldsymbol{x}_i$, and θ_{y_i} is the angle between $\boldsymbol{W}_{y_i}$ and visual feature $\boldsymbol{x}_i$. The softmax loss attempts to learn discriminative features by maximizing the posterior probability of ground-truth class. Looking back on experience of face recognition, we find that weight and feature normalization has been proven effective in improving model performance. Thus, we fix $\boldsymbol{W}_j$ by L_2 normalization. Therefore, cosine of θ_j turns to be one of the only two variables of the posterior probability of feature $\boldsymbol{x}_i$. We formulate the modified softmax loss as:

$$L_{ms} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\|\boldsymbol{x}_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right). \tag{2}$$

Unlike face recognition which only involves visual features, text-based person search is concerned with not only visual but also textual features. In our proposed multiplicative angular margin loss, we integrate the textual and visual features by projecting one onto another normalized one. Consequently, we get a new feature vector with its magnitude and direction respectively from textual feature and visual feature. In this way, we strengthen the association within the matched pairs. In addition, we propose to incorporate angular margin into our designed projection vector via multiplying θ_{y_i} by a hyper-parameter m. To sum up, we reformulate the softmax loss and eventually get the image classification loss as:

$$L_{\text{ipt}} = \frac{1}{N}\sum_{i} -\log\left(\frac{e^{\|\hat{\boldsymbol{x}}_i\|\cos(m\theta_{y_i,i})}}{e^{\|\hat{\boldsymbol{x}}_i\|\cos(m\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|\hat{\boldsymbol{x}}_i\|\cos(\theta_{j,i})}}\right), \tag{3}$$

where $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i^{\top}\overline{\boldsymbol{z}}_i \cdot \overline{\boldsymbol{z}}_i$, $\overline{\boldsymbol{z}}_i$ indicates the normalized text feature of $\boldsymbol{z}_i$ and $\hat{\boldsymbol{x}}_i$ denotes the vector projection of $\boldsymbol{x}_i$ onto $\overline{\boldsymbol{z}}_i$. Similar to the image classification loss, the text classification loss can be written as:

$$L_{\text{tpi}} = \frac{1}{N}\sum_{i} -\log\left(\frac{e^{\|\hat{\boldsymbol{z}}_i\|\cos(m\theta_{y_i,i})}}{e^{\|\hat{\boldsymbol{z}}_i\|\cos(m\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|\hat{\boldsymbol{z}}_i\|\cos(\theta_{j,i})}}\right). \tag{4}$$

Finally, the MAM loss can be written as:

$$L_{\text{mam}} = L_{\text{ipt}} + L_{\text{tpi}}. \tag{5}$$

3.3 Pairwise Similarity Weighting Loss

It is worth noticing that in text-based person search, the number of matched image-text pairs for each identity is much smaller than unmatched ones. However, in the existing algorithms such as triplet loss, all matched and unmatched pairs are assigned with same weight resulting in slow convergence and poor performance. To better take advantage of informative pairs, hard sample mining strategies and pair weighting methods have been proposed in many deep metric learning literatures. In particular, it is necessary to design a pair weighting mechanism to assign larger weight for informative image-text pairs in our task. According to the focal loss [13] and the MS loss [27], it is believed that the weight of positive pair is inversely proportional to its similarity value while the weight of negative pair is in proportion to its similarity value. In our proposed pairwise similarity weighting loss, we define weight of an image-text pair as the derivative of pairwise similarity weighting loss with respect to its cosine similarity. For convenience and simplicity, we specially design our pairwise similarity weighting loss based on quadratic function. In the following, we will first give the formulation of our loss and then further give a mathematical proof of its validity.

In this loss, we refer to S_{ij} as the cosine similarity of $(\boldsymbol{x}_i, \boldsymbol{z}_j)$. $N_{x_i} = \left\{S_{ij,j\neq i}\right\}$ indicates a set of similarity scores for all negative pairs with $\boldsymbol{x}_i$ as anchor and $N_{z_j} = \left\{S_{iz,z\neq i}\right\}$ indicates a set of similarity scores for all negative pairs with $\boldsymbol{z}_j$ as anchor. Finally, we formulate our proposed PSW loss as:

$$\begin{aligned} L_{\text{psw}} = &\frac{1}{N}\sum_{i=1}^{i=N}\left[\sum_{p=0}^{2} a_p S_{ii}^{p} + \sum_{q=0}^{2} b_q \operatorname{Max}\left\{N_{x_i}\right\}^{q}\right] + \\ &\frac{1}{N}\sum_{j=1}^{j=N}\left[\sum_{p=0}^{2} a_p S_{jj}^{p} + \sum_{q=0}^{2} b_q \operatorname{Max}\left\{N_{z_j}\right\}^{q}\right], \end{aligned} \tag{6}$$

where $\operatorname{Max}\left\{N_{x_i}\right\}$ and $\operatorname{Max}\left\{N_{z_i}\right\}$ mean the cosine similarity of hardest negative pair of image feature $\boldsymbol{x}_i$ and text feature $\boldsymbol{z}_j$ separately, $\left\{a_p\right\}_{p=0}^{p=2}$ and $\left\{b_q\right\}_{q=0}^{q=2}$ are hyper-parameters which are set in advance. To further explain how our proposed pairwise similarity weighting loss works, we plot about the relation between cosine similarity of negative pairs and the pairwise similarity weighting loss. First, we denote the weight of a negative pair $(\boldsymbol{x}_i, \boldsymbol{z}_{j,i\neq j})$ as:

$$w_{ij} = \frac{\partial L_{\text{psw}}}{\partial S_{ij}}. \tag{7}$$

It can be clearly inferred from Figure 4 that as the similarity of negative pair increase, the loss value increases. Moreover, the derivative of the loss value with respect to negative pair cosine similarity (i.e, the pair weight w_{ij}) increases as its similarity value increases. Thus, our proposed pairwise similarity weighting loss has the property that the weight of a negative pair is in proportion to its similarity value, ensuring that a larger weight is assigned to more informative pair.

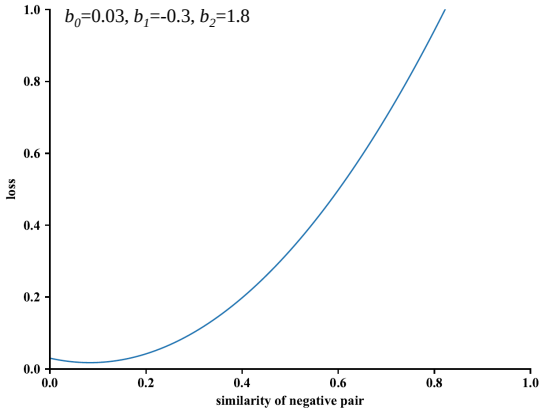

Figure 4: As similarity of negative pair increases, both its related pairwise similarity weighting loss and the pair weight w_{ij} increase.

3.4 Objective Functions

In order to reduce the model convergence time and further improve performance, we introduce cross-modal projection matching (CMPM) loss which uses KL divergence to increase the association between the representations across different modalities. In the CMPM loss, the posterior probability of matching $\boldsymbol{x}_i$ to $\boldsymbol{z}_j$ is formulated as:

$$p_{i,j} = \frac{\exp\left(\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_j\right)}{\sum_{k=1}^{n} \exp\left(\boldsymbol{x}_i^\top \bar{\boldsymbol{z}}_k\right)} \quad \text{s.t. } \bar{\boldsymbol{z}}_j = \frac{\boldsymbol{z}_j}{\left\|\boldsymbol{z}_j\right\|}. \tag{8}$$

The true matching probability of $\left(\boldsymbol{x}_i, \boldsymbol{z}_j\right)$ in a mini-batch is calculated as:

$$q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{n} y_{i,k}}. \tag{9}$$

Then, the matching loss from image to text is calculated as:

$$L_i = \sum_{j=1}^{n} p_{i,j} \log \frac{p_{i,j}}{q_{i,j} + \epsilon}, \tag{10}$$

$$L_{\text{i2t}} = \frac{1}{n} \sum_{i=1}^{n} L_i, \tag{11}$$

where ϵ is a small value set to avoid numerical problems. Lastly, the cross-modal projection matching loss is defined as:

$$L_{\text{cmpm}} = L_{\text{i2t}} + L_{\text{t2i}}. \tag{12}$$

L_{t2i} indicates the matching loss from text to image with the similar derivation procedure to L_{i2t}. To sum up, the overall objection function is written as:

$$L = L_{\text{mam}} + L_{\text{psw}} + L_{\text{cmpm}}. \tag{13}$$

4 EXPERIMENTS

4.1 Dataset and Implementation Details

Up to now, CUHK-PEDES [12] is the only public dataset for the task of text-based person search. There are totally 40,206 pedestrian images of 13,003 identities. Each pedestrian image in this dataset is described by about two textual descriptions with 80,440 sentences in total. We split the dataset into training set, validation set and test set according to the protocol in [12]. The training set includes 11,003 identities with 34,054 images and 68,108 sentence descriptions. The validation set and test set both contain 1,000 identities and 3,074 images with 6,158 and 6,156 textual descriptions, respectively. Our proposed method is implemented by TensorFlow. All the images are resized to 224×224. We employ Adam optimizer to train our model and set the learning rate to 0.0002. The batch size is 16 and the multiplicative coefficient m in Eq. 3 is set at 4. As for pairwise similarity weighting loss, we set $\{a_0 = 0.5, a_1 = -0.7, a_2 = 0.2, b_0 = 0.03, b_1 = -0.3, b_2 = 1.8\}$. In the inference stage, we first extract image features from MobileNet and textual features from Bi-LSTM and then obtain target features from joint embedding learning module. After normalizing the target feature vectors, we calculate the cosine similarity between visual and textual features. Recall@K, which means the percentage of probes where at least one ground-truth is retrieved among the top-K results, is adopted as our evaluation metric. The code of MAM loss is available at https://github.com/pengzhanguestc/MAM_loss.

Table 1: Comparison of person search results(R@K (%)) on the CUHK-PEDES dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
LSTM Q+norm [1]	17.19	–	57.82
GNA-RNN [12]	19.05	–	53.64
IATVM [11]	25.94	–	60.48
PWM-ATH [3]	27.14	49.45	61.02
GLA [2]	43.58	66.93	76.26
Dual Path [30]	44.40	66.26	75.07
CAN [9]	45.52	67.12	76.98
CMPM-CMPC [29]	49.37	–	79.27
MCCL [28]	50.58	–	79.06
A-GANet [16]	53.14	74.03	81.95
PMA [10]	53.81	73.54	81.23
Ours	**54.24**	**74.82**	**82.39**

4.2 Performance Comparison

In Table 1, we make a performance comparison of our proposed method and other existing algorithms on the CUHK-PEDES dataset. The compared algorithms can be classified into two categories. One is building a similarity score learning network with attention mechanism, including LSTM Q+norm [1], GNA-RNN [12], IATVM [11], PWM-ATH [3], GLA [2], and PMA [10]. The other is focusing on joint embedding learning methods, including dual path [30], CAN [9], MCCL [28], CMPM-CMPC [29] and A-GANet [16]. Compared with all these above dedicated and sophisticated methods, clearly our proposed method is more efficient with lightweight model and without applying human segmentation, attribute alignment and other complicated technologies.

4.3 Ablation Study

We have conducted a series of ablation experiments on the CUHK-PEDES dataset to validate the effectiveness of our proposed loss function. Through comparing model performance with and without

(a) Text queries (b) Full model (c) w/o MAM (d) w/o PSW

Figure 5: Examples of top-5 language person search results by different algorithms, including (b) full model, (c) full model w/o MAM, and (d) full model w/o PSW. Corresponding images are marked by green bounding boxes.

Table 2: Ablation study of the effect of each proposed loss within our model on the CUHK-PEDES dataset.

CMPM	MAM	PSW	Top-1	Top-5	Top-10
√	×	×	44.13	67.11	77.35
√	×	√	52.76	73.96	82.36
√	√	×	52.94	73.49	81.95
√	√	√	**54.24**	**74.82**	**82.39**

a specific component, we can assess how much contribution each proposed component makes to the model performance. In this section, we mainly evaluate the effects of MAM loss and PSW loss.

Table 2 demonstrates the ablation results of our proposed method. We can observe from the table that our model obtains 44.13%, 67.11%, 77.35% of rank-1, rank-5 and rank-10 accuracy respectively under the circumstance of keeping only the CMPM loss. Through absorbing either the MAM loss or PSW loss, our model achieves a considerable improvement on model performance that the rank-1 accuracy increases by approximately 8.0%, which proves the effectiveness of our proposed MAM loss and PSW loss. Moreover, by comparing the performance of CMPM+MAM+PSW with CMPM+MAM and CMPM+PSW, we come to a conclusion that our model yields the best rank-1 accuracy performance of 54.24% with the presence of both MAM and PSW losses, which further validate the effectiveness of our proposed method.

4.4 Visualization of Retrieval Results

In addition to quantitative analysis, we have conducted qualitative experiments. In Figure 5, we present the visualization of person retrieval results in order to provide a qualitative examination. For each given textual description on the CUHK-PEDES dataset, we show the top-5 results in Figure 5. Figure 5(b) shows the results of our proposed method, while Figure 5(c) represents the retrieval results without MAM loss, and Figure 5(d) represents the retrieval results without PSW loss. From the visualized retrieval results, we can observe that MAM loss and PSW loss have a promotive effect on the retrieval performance.

5 CONCLUSION

Text-based person search is essentially a cross-modal retrieval task and a promising line of research is to build a joint embedding learning module which aims at eliminating the semantic gap between data from different modalities. In this paper, we focus on designing effective loss function to encourage the joint embedding learning process. MAM loss is a modified version of softmax loss and it can promote learning a more discriminative embedding for each pedestrian identity by adding a multiplicative angular margin. Besides, we exploit pairwise similarity weighting loss to reasonably assign weight to image-text pair in terms of its quantity of information. Both quantitative and qualitative experiments have demonstrated the effectiveness of our proposed method. Limited by the lack of public dataset resources in text-based person search, all of our experiments are conducted on the CUHK-PEDES dataset. In the future, we plan to do some works about building or enlarging dataset about text-based person search and further validate our proposed method on the new dataset.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 61832001 and No. 61672133) and Sichuan Science and Technology Program (No. 2019YFG0535).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 2425–2433.

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving Deep Visual Representation for Person Re-identification by Global and Local Image-language Association. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*. 56–73.

[3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. 1879–1887.

[4] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. 2019. Video Person Re-Identification by Temporal Residual Learning. *IEEE Trans. Image Process.* 28, 3 (2019), 1366–1377.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 4690–4699.

[6] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. 1735–1742.

[7] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. 2018. Triplet-Center Loss for Multi-View 3D Object Retrieval. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 1945–1954.

[8] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* abs/1704.04861 (2017).

[9] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Cascade Attention Network for Person Search: Both Image and Text-Image Similarity Selection. *CoRR* abs/1809.08440 (2018).

[10] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*. 11189–11196.

[11] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-Aware Textual-Visual Matching with Latent Co-attention. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 1908–1917.

[12] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 5187–5196.

[13] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2999–3007.

[14] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* 95 (2019), 151–161.

[15] Haijun Liu, Jian Cheng, Wen Wang, and Yanzhou Su. 2019. The General Pair-based Weighting Loss for Deep Metric Learning. *CoRR* abs/1905.12837 (2019).

[16] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. 665–673.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 6738–6746.

[18] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. 507–516.

[19] Zimo Liu, Dong Wang, and Huchuan Lu. 2017. Stepwise Metric Promotion for Unsupervised Video Person Re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2448–2457.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 815–823.

[21] Kihyuk Sohn. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems NIPS 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 1849–1857.

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4004–4012.

[23] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2016. Deep Attributes Driven Multi-camera Person Re-identification. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. 475–491.

[24] Daniel A. Vaquero, Rogério Schmidt Feris, Duan Tran, Lisa M. Brown, Arun Hampapur, and Matthew A. Turk. 2009. Attribute-based people search in surveillance environments. In *IEEE Workshop on Applications of Computer Vision (WACV 2009), 7-8 December, 2009, Snowbird, UT, USA*. 1–8.

[25] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive Margin Softmax for Face Verification. *IEEE Signal Process. Lett.* 25, 7 (2018), 926–930.

[26] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2275–2284.

[27] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5022–5030.

[28] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. 2019. Language Person Search with Mutually Connected Classification Loss. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2057–2061.

[29] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*. 707–723.

[30] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path Convolutional Image-Text Embeddings with Instance Loss. *ACM Trans. Multim. Comput. Commun. Appl.* 16, 2 (2020), 51:1–51:23.

[31] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.