A Novel Soft Margin Loss Function for Deep Discriminative Embedding Learning

ZHAO YANG, (Member, IEEE), TIE LIU, JIEHAO LIU, LI WANG, AND SAI ZHAO

School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China
Huangpu Research and Graduate School, Guangzhou University, Guangzhou 510006, China

Corresponding author: Zhao Yang (yangdxng100@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501177 and Grant 61902084, in part by Guangzhou University’s Training Program for Excellent New-Recruited Doctors under Grant YB201712, in part by the Science and Technology Planning Project of Guangzhou Municipal Government under Grant 201904010466, and in part by the Featured Innovation Project of Guangdong Education Department under Grant 2018KTSCX174.

ABSTRACT Deep embedding learning aims to learn discriminative feature representations through a deep convolutional neural network model. Commonly, such a model contains a network architecture and a loss function. The architecture is responsible for hierarchical feature extraction, while the loss function supervises the training procedure with the purpose of maximizing inter-class separability and intra-class compactness. By considering that loss function is crucial for the feature performance, in this article we propose a new loss function called soft margin loss (SML) based on a classification framework for deep embedding learning. Specifically, we first normalize the learned features and the classification weights to map them into the hypersphere. After that, we construct our loss with the difference between the maximum intra-class distance and minimum inter-class distance. By constraining the distance difference with a soft margin that is inherent in the proposed loss, both the inter-class discrepancy and intra-class compactness of learned features can be effectively improved. Finally, under the joint training with an improved softmax loss, the model can learn features with strong discriminability. Toy experiments on MNIST dataset are conducted to show the effectiveness of the proposed method. Additionally, experiments on re-identification tasks are also provided to demonstrate the superior performance of embedding learning. Specifically, 65.48% / 62.68% mAP on CUHK03 labeled / detected dataset (person re-id) and 74.36% mAP on VeRi-776 dataset (vehicle re-id) are achieved respectively.

INDEX TERMS Soft margin loss, deep embedding learning, feature representation, person re-identification, vehicle re-identification.

I. INTRODUCTION

Deep embedding learning focuses on learning discriminative representations from input data, whose fundamental purpose is to pull similar samples close and push dissimilar samples away. This intuitive but practical principle enables embedding learning to be widely applied in various fields including person re-identification [1]–[3], vehicle re-identification [4], [5], face recognition [6]–[9], etc. In general, a deep embedding learning framework comprises two basic components: network architecture and loss function. The network architecture usually consists of cascaded deep convolutional neural networks which can extract highly abstract representations of the input images through its strong non-linear transformation ability and map them into an embedding space. Then the loss function is used to enhance the discriminability capacity of learned features in the embedding space by constraining their intra-class and inter-class relationships. Since existing deep models have enough power to extract informative features of input images, the loss functions play a critical role in discriminative embedding learning.

Most of loss functions used in embedding learning directly constrain the distance between samples [3], [7], [10], [11]. An intuitive and typical loss function is contrastive loss [10] which pulls a pair of samples together if they belong to the same class and pushes them away by a margin if they come from different classes. Another extensively used loss function is triplet loss [3], [7]. It also adopts a margin to decrease
the distance between an anchor and a positive sample and increase the distance between the anchor and a negative sample. As a predefined margin is involved in the expression of loss function, it is tricky to select an optimum margin parameter for both contrastive loss and triplet loss in the training procedure. Recently, loss functions that used for classification tasks are gradually applied to guide embedding learning via the form of classification. The most representative loss functions are the several variants of softmax, such as L-Softmax [12], A-Softmax [9], ArcFace [6], and so on. They have been proved to be practical and effective for embedding learning. However, the softmax based loss functions only try to separate the features of different classes from the training set instead of learning discriminative features directly. Thus large intra-class variations cannot be handled well in the training process.

Therefore, to alleviate these problems in the discriminative embedding learning, we present a novel loss function named soft margin loss (SML) in this article. We first normalize the learned features and classification weights thus these vectors are mapped into the hypersphere. By regarding the classification weights as class centers, the intra-class distance can be calculated as the distance between the center and the feature, and the inter-class distance can be calculated as the distance between different class centers. A simple diagrammatic illustration is shown in the left of Figure 1, where the solid dots denote the normalized features and the hollow dots denote the normalized classification weights (a.k.a., the class centers). Different classes are represented with different colors. Theoretically, the learned features are well separable once the maximal intra-class distance is smaller than the minimal inter-class distance, and the discriminability power of learned features will be enhanced as the difference of these two distances decreases. Therefore, we select the difference of the maximal intra-class distance and the minimal inter-class distance as a constraint objective in our loss. Then we constrain the distance difference with a soft margin in our proposed loss whose general formulation and the curve characteristic are shown in the right of Figure 1.

From the loss curve, we can find that the proposed soft margin loss has some superiorities for embedding learning. On the one hand, the proposed loss strictly penalizes the embedding distances when the difference of the maximal intra-class distance and the minimal inter-class distance is larger than zero. On the other hand, our loss still provides the constraint power with a soft margin when the difference is smaller than zero. Compared with some fixed margin-based losses, our proposed loss is more flexible in feature learning and its parameter setting of the model training is less empirical. Under the constraint of the soft margin loss, the within-class compactness and between-class discrepancy of learned features are effectively improved, which is favorable to discriminative embedding learning.

We organize the rest paper as follows: Section II gives an introduction of some related works about loss function in the field of deep embedding learning. Section III describes the proposed soft margin loss in detail and introduces the joint training scheme with the improved softmax. Section IV demonstrates the effectiveness and superiority of the soft margin loss via MNIST experiments and some re-identification experiments. After that, Section V compares our loss with some similar works. Finally, Section VI draws some conclusions about our works.

II. RELATED WORKS

In recent years, embedding learning has attracted a great interest in the computer vision community and has been extensively applied in many popular fields including person re-identification, vehicle re-identification, image retrieval and face recognition, etc. The key idea of embedding learning is to learn a satisfactory embedding space where the intra-class compactness and inter-class separability are as large as possible. In the discriminative embedding learning, many works [3], [7], [9], [10], [12]–[14] focus on the design of loss function which can provide a powerful and clear supervision for discriminative feature learning. In this section, we will give a view of some loss functions that are frequently used for embedding learning.

A. METRIC LOSS

In deep learning framework, the metric loss is usually used to constrain the distances between learned features. It accords with the aim of embedding learning which keeps semantically related images close and unrelated images far away in the embedding space. Therefore, various metric losses [1], [3], [7], [10], [11], [15] are extensively applied in embedding learning.

One of the most concise and intuitive metric losses is the contrastive loss [10]. It takes a couple of images as the inputs and pulls the distance between the images if they belong to the identical class or pushes them away by a margin if they come from different classes. For example, Varior et al. [11] applied the contrastive loss in a gated siamese convolutional neural network for person re-id task. Similarly, Taigman et al. [16] used the contrastive loss in the face verification with a siamese network. Although the contrastive loss is verified to be effective in many tasks, the number of pairwise
introduced a framework for person re-identification where the triplet loss and softmax loss are used together to improve the discriminability of learned features. Wen et al. [14] learned discriminative face representations with the joint optimization of a center loss and softmax loss in which the center loss is applied to further shrink the intra-class variation of each class. By considering the within-class variation and between-class relationship simultaneously, He et al. [13] put forward a triplet-center loss which combines the center loss with triplet loss, and used a joint training scheme with softmax loss for 3D object retrieval.

### III. PROPOSED METHODS
In this section, we firstly review the original softmax loss function and its improved version which is beneficial to discriminative feature learning. After that, we detail our proposed soft margin loss. Finally, we present the joint training scheme of the improved softmax loss and the soft margin loss in our method.

#### A. PRELIMINARIES

A typical softmax loss comprises a softmax activation and cross-entropy loss. Softmax loss converts the model output into the class predictions by the softmax activation and calculates the loss via the cross-entropy. The original softmax loss can be expressed as:

$$
L_S = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{W_{yi}^{T}f_i + b_y}}{\sum_{j=1}^{C} e^{W_{yi}^{T}f_i + b_j}}
$$

where $i$ is the index of a sample in a batch of training data, and $f_i$ indicates the deep feature of $i$-th sample whose class label is $y_i$. $W_j$ and $b_j$ represent the $j$-th weight column vector in the last fully connected (FC) layer and corresponding bias respectively. $C$ is the class number, and $n$ is the size of the batch.

Some works [6], [8] have shown that better model performance can be obtained by eliminating the bias item and the magnitude influences of both the features and classification weights. Specifically, the logit (19) item $W_{yi}^{T}f_i + b_j$ is transformed as $\|W_j\| \cdot \|f_i\| \cdot \cos(\theta_j)$ with ignoring the bias $b_j$. $\theta_j$ is the angle between $W_j$ and $f_i$. By fixing the weight $\|W_j\| = 1$ and the embedding feature $\|f_i\| = s$ with L2 normalizations respectively, the logit item can be simplified as $s \cdot \cos(\theta_j)$, where $s$ serves as a scale factor that controls the range of the feature space. Finally, the improved softmax can be expressed as follows:

$$
L_{NSL} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{W_{yi}^{T}f_i \cdot \cos(\theta_j)}}{\sum_{j=1}^{C} e^{W_{yi}^{T}f_i \cdot \cos(\theta_j)}}
= -\frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{s \cdot \cos(\theta_j)}}{\sum_{j=1}^{C} e^{s \cdot \cos(\theta_j)}}.
$$

With the normalizations of classification weights and features, the classification scores only depend on the angle between the feature and corresponding classification weight. Thus the learned features are angularly separable in the
hypersphere. In this article, we name the improved softmax expressed in equation (2) as normalized softmax loss (NSL) to distinguish the original softmax loss.

**B. THE PROPOSED SOFT MARGIN LOSS**

Softmax loss can learn a fine embedding space for input images. However, once the images are well classified in the embedding space, softmax loss lacks a distinct and strong supervision to continuously pull the similar images close and push the dissimilar images away. So it is hard for the model to mine enough discriminative information of input images. Therefore, we propose the soft margin loss for discriminative embedding learning. Concretely, we first normalize the features and classification weights as the same processing in the NSL. After that, we regard the normalized classification weights in the last FC layer as the center of each class. In this way, the intra-class distance could be represented as the cosine distance between the center and the feature coming from the same class, and the inter-class distance can be represented as the cosine distance between the different centers. To constrain the intra-class and inter-class distance strictly, we use the difference of the maximal intra-class distance and the minimal inter-class distance as the constraint objective of the proposed loss. Finally, the specific formulation of the soft margin loss is given by:

\[
L_{SML} = \frac{1}{C} \sum_{j=1}^{C} \log \left( e^{\max_{y \neq j}(\langle f_i, W_j \rangle - \min_{y \neq j}(\langle W_k, W_j \rangle)} \right),
\]

where \( \varepsilon \) is a moderating factor to adjust the strength of the soft margin.

Our proposed soft margin loss has three superior properties. First, since the features and classification weights are normalized into the hypersphere, the intra-class and inter-class distances are irrelevant to the magnitudes of both features and classification weights. Thus the soft margin loss calculated by these distances can efficiently guide the embedding learning. Second, the soft margin loss can generate a strict penalty when the difference of the maximal intra-class distance and the minimal inter-class distance is larger than zero. It means that the soft margin loss can help the model learn a correct classification quickly at the initial training stage. Third, even if the distance difference is smaller than zero, the proposed loss still provides a soft margin to help the model mine the discriminative information from the input data. On the one hand, this soft margin can help discriminative embedding learning via further improving the within-class compactness and between-class discrepancy. On the other hand, the softness of the margin can provide flexibility during the model training procedure compared with a fixed margin. Therefore, the proposed soft margin loss could be used for discriminative embedding learning.

**C. JOINT TRAINING**

The classification loss can learn a fine embedding space but lacks a distinct and strong supervision signal to continuously enlarge between-class discrepancy and shrink within-class compactness. The metric loss can be used for discriminative embedding learning but encounters the low convergence problem. Therefore, in this article, we introduce a joint training scheme to combine the normalized softmax loss and the proposed soft margin loss. Thus the final loss representation is:

\[
L = L_{NSL} + \lambda L_{SML},
\]

where \( \lambda \) is a parameter to balance the normalized softmax loss and the soft margin loss. Different from most joint training schemes that directly use softmax loss, our method applies the normalized softmax loss in the joint training. Since the features and classification weights are regulated by the L2 normalization, the normalized softmax loss is only relevant to the angles between the features and classification weights. Therefore, the optimization targets of these two losses are consistent during the training procedure.

**IV. EXPERIMENTS**

In this section, we first carry out some toy experiments based on MNIST dataset to intuitively show the superiority of the proposed loss by comparing with the original softmax loss and the normalized softmax loss. Subsequently, we further demonstrate the effectiveness of our method on re-identification tasks, including person re-identification and vehicle re-identification. Both of them can be regarded as an image retrieval problem [20], which retrieves similar images to a query image among a large dataset.

**A. TOY EXAMPLES ON MNIST**

MNIST dataset [21] is the most popular hand-written digit benchmark dataset. It includes a total of 70000 images of 10 types of numbers from 0 to 9, where 60000 images are used for training and the rest 10000 images are used for testing.

For convenience, we use a concise CNN model to conduct the MNIST experiments with the original softmax, the normalized softmax and the normalized softmax with soft margin loss respectively. All model parameter settings are the same except the loss function used in the network. To intuitively demonstrate the effects of three different losses, we visualize the learned features by setting their dimension as 2 and projecting them into 2-D embedding space. The results are illustrated in Figure 2. The first row denotes the features without normalization and the second row represents the corresponding features with normalization.

From the feature distributions in Figure 2, we can see that the features learned by the original softmax are well separable in the embedding space while they have poor intra-class compactness. For the features learned by the normalized softmax, we can find that the intra-class compactness has been improved greatly since the features and classification weights are normalized into hypersphere. In spite of the satisfactory feature embedding characteristics, preferable intra-class compactness can be obtained by adding the soft margin loss...
FIGURE 2. Experimental results based on MNIST dataset with original softmax, normalized softmax and normalized softmax with soft margin loss (SML) respectively. The first row represents the data distributions of original features in 2-D embedding space and the second row represents the corresponding normalized features. From the results, we can see that the features learned by original softmax have less within-class compactness while the normalized softmax can effectively improve the within-class compactness by normalizing the features and classification weights. On the basis of the normalized softmax, the soft margin loss constrains both the within-class compactness and between-class discrepancy of features simultaneously thus the learned features are more discriminative. Best viewed in color.

to the normalized softmax. From the results trained by the normalized softmax with soft margin loss, we can obviously observe that the intra-class variance becomes smaller than that of the normalized softmax, which means that the feature discriminations are further strengthened.

From our observations in the MNIST experiments, some conclusions can be drawn. First of all, it is practicable to use the classification loss for the embedding learning. However, the features learned by softmax loss are less discriminative because the original softmax mainly focuses on the separability of learned features. Second, the normalized softmax improves the discriminability of feature embedding by regulating the magnitudes of both features and classification weights as constants. In this way, the model only focuses on the cosine distances between features during the training procedure. Thus the learned features are more discriminative than that learned from the original softmax. Third, on the basis of the normalized softmax, the proposed soft margin loss explicitly constrains the inter-class and intra-class distances between features. Therefore, with the joint training of the normalized softmax and the proposed soft margin loss, the learned features can obtain powerful discriminability.

Based on the experimental results on MNIST as well as above analysis, the proposed soft margin loss shows its superiority and latent capacity for discriminative embedding learning.

B. EXPERIMENTS ON RE-IDENTIFICATION TASKS

1) DATASET DESCRIPTIONS

CUHK03 [22] is an extensively used person re-id dataset which is collected by 5 pairs of cameras in CUHK campus. CUHK03 contains 14096 images of 1467 person identities. For practical applications, CUHK03 dataset provides not only manually cropped pedestrian images, but also automatically detected bounding boxes. However, the dataset is originally designed for a single-shot situation which cannot comprehensively evaluate the performance of person re-id tasks. Therefore Zhong et al. [23] introduced a new protocol for training/testing. Specifically, 767 person identities are allocated to the training set and the rest 700 person identities are used as the testing set. We use the new protocol of CUHK03 in our experiments. During the training procedure, we rescale the input images to $288 \times 144$ and randomly crop them to $256 \times 128$. Then the images are horizontally flipped with the probability 0.5. As most re-id tasks do, we normalize the image RGB channels by subtracting $(0.485, 0.456, 0.406)$ and dividing $(0.229, 0.224, 0.225)$, respectively. Moreover, a random erasing operation [24] is used on the training images as a kind of data augmentation trick to enhance the robustness of the model. In the testing phase, the images are resized to $288 \times 144$. The same normalization operation is done before the images are fed into the testing network. It is worth noting that the final feature embedding is the average of features from the original image and its horizontally flipped version.

VeRi-776 [25] is a large scale publicly available vehicle dataset built from the VeRi dataset [26], which is collected in a real-world traffic scene by 2 to 18 cameras at different viewpoints, illuminations, resolutions and occlusion conditions. After the expansion from the VeRi dataset, VeRi-776 contains 776 different vehicles, including 37781 pictures of 576 vehicles in the training set and 11579 pictures of 200 vehicles in the testing set. Different from the preprocessing of the person
In many re-identification datasets, the available images of each class are usually quite different in the number. Therefore, the model may tend to overfit the class with abundant images and ignore the class with few images. To alleviate this problem, we propose a balanced sampling scheme. In this scheme, we randomly select a fixed number of images from each class. Specifically, in each epoch, we randomly select \( P \) classes and corresponding \( K \) images for training. The images are then cropped to \( 256 \times 256 \) randomly. Then the sampled classes and corresponding images are stationary in each mini batch. Specifically, \( P \) classes are randomly selected without replacement in each epoch. Then for every class, we randomly select \( K \) images for training. The images will be replaced if the number of images is less than \( K \). So, there are always \( P^*K \) images in a mini batch, and we set \( P \) and \( K \) as 16 and 4 respectively in the experiments.

Besides, we adopt Adam [30] optimizer to update model parameters during the training. We set the total epoch as 150 and use a warm-up strategy [2] during the initial 20 epochs. It means that the learning rate will increase continuously from a small value. Concretely, the learning rate linearly increases from \( 10^{-5} \) to \( 10^{-3} \) in the first 20 epochs. Since the model may tackle different tasks (e.g., person or vehicle re-identification tasks), a relatively small learning rate is beneficial for the model to obtain a well initial state in different tasks. After the warm-up stage, the learning rate remains at \( 10^{-3} \) until 90th. Then the learning rate is decayed by 0.1 at 90th and 130th separately to fine-tune the parameters.

Moreover, we apply a hard mining strategy in our method which can enhance the generalization ability of the learned model. More specifically, we sort a batch of samples according to the predictions of the normalized softmax in descending order, and take the first 80% of the samples to update the parameters of our model. Besides, the scale parameter \( s \) in normalized softmax is 14, the moderating factor \( \epsilon \) in the proposed soft margin loss is 0.2 and the balance weight \( \lambda \) for soft margin loss is set as 1.0.

### 3) EXPERIMENTAL SETTINGS

All experiments are conducted in the Pytorch [29] framework with an NVIDIA GTX 1080 Ti GPU. Except for the tiny difference in the data preprocessing, we keep all the network parameters same in the person and vehicle re-identification tasks, which can illustrate the generalization ability of our proposed method.

In many re-identification datasets, the available images of each class are usually quite different in the number. Therefore, the model may tend to overfit the class with abundant images and ignore the class with few images. Therefore, we use a balanced sampling scheme [2] in which the sampled classes and corresponding images are stationary in each mini batch. Specifically, \( P \) classes are randomly selected without replacement in each epoch. Then for every class, we randomly select \( K \) images for training. The images will be replaced if the number of images is less than \( K \). So, there are always \( P^*K \) images in a mini batch, and we set \( P \) and \( K \) as 16 and 4 respectively in the experiments.

Besides, we adopt Adam [30] optimizer to update model parameters during the training. We set the total epoch as 150 and use a warm-up strategy [2] during the initial 20 epochs. It means that the learning rate will increase continuously from a small value. Concretely, the learning rate linearly increases from \( 10^{-5} \) to \( 10^{-3} \) in the first 20 epochs. Since the model may tackle different tasks (e.g., person or vehicle re-identification tasks), a relatively small learning rate is beneficial for the model to obtain a well initial state in different tasks. After the warm-up stage, the learning rate remains at \( 10^{-3} \) until 90th. Then the learning rate is decayed by 0.1 at 90th and 130th separately to fine-tune the parameters.

Moreover, we apply a hard mining strategy in our method which can enhance the generalization ability of the learned model. More specifically, we sort a batch of samples according to the predictions of the normalized softmax in descending order, and take the first 80% of the samples to update the parameters of our model. Besides, the scale parameter \( s \) in normalized softmax is 14, the moderating factor \( \epsilon \) in the proposed soft margin loss is 0.2 and the balance weight \( \lambda \) for soft margin loss is set as 1.0.

### 4) EXPERIMENTAL RESULTS

In re-identification tasks, cumulative match characteristic (CMC) and mean average precision (mAP) are two widely used evaluation metrics. CMC gives the average probability that the image which matches with a specific query image arises in the first-\( k \) candidates of the gallery set. However, when there are multiple matching images in gallery (e.g., person re-identification task), CMC metric will be deficient and cannot evaluate the method comprehensively. Therefore, many re-id tasks use mAP evaluation which considers all
true matches and returns the mean average precision. In our experiment, we report values of CMC at Rank-1 and mAP for the person and vehicle re-identification tasks.

The experimental results and comparisons with corresponding state-of-the-art works on person and vehicle re-identification tasks are given in Table 1 and Table 2 respectively. In both re-id tasks, the models trained by the normalized softmax are treated as the baseline model.

**Table 1. The results of our method and some state-of-the-art works on CUHK03 dataset.**

| Methods | Labeled | Detected |
|---------|---------|----------|
|         | mAP     | Rank-1   | mAP     | Rank-1   |
| DaF [31] | 31.5    | 27.5     | 30.0    | 26.4     |
| PAN [32]  | 35.0    | 36.9     | 34.0    | 36.3     |
| SVDNet [33] | 37.83   | 40.93    | 37.3    | 41.5     |
| DPFL [34] | 40.5    | 43.0     | 37.0    | 40.7     |
| HA-CNN [35] | 41.0    | 44.4     | 38.6    | 41.7     |
| MGCAM-Siamese [36] | 50.21   | 50.14    | 46.87   | 46.71    |
| MLFN [37] | 49.2    | 54.7     | 47.8    | 52.8     |
| PCB+RPP [38] | -       | -        | 56.7    | 62.8     |
| Ours     | 65.48   | 67.86    | 62.68   | 64.21    |

**Table 2. The results of our method and some state-of-the-art works on VeRi-776 dataset.**

| Methods            | VeRi-776 |
|--------------------|----------|
|                    | mAP     | Rank-1   |
| XVGAN [39]         | 24.65   | 60.20    |
| VAMI [40]          | 50.13   | 77.03    |
| PROVID [41]        | 53.42   | 81.56    |
| SDC-CNN [42]       | 53.45   | 83.49    |
| JFSDL [5]          | 53.53   | 82.90    |
| Hard-View-EALN [4] | 57.44   | 84.39    |
| RAM [43]           | 61.5    | 88.6     |
| QD-DLF [44]        | 61.83   | 88.50    |
| Ours               | 73.00   | 94.87    |

| Methods | VeRi-776 |
|---------|----------|
|         | mAP     | Rank-1   |
|         | 74.36   | 94.99    |

In the results of person re-identification task, we can find that the proposed method outperforms the baseline in CUHK03 labeled version and detected version. For example, in the labeled version, our proposed method has improved performance by +2.14% and +3.43% on mAP and Rank-1. In detected version, our method brings +2.32% and +2.78% improvements on mAP and Rank-1. Besides, we also make a comparison between the proposed method and some state-of-the-art works such as SVDNet [33] and PCB+RPP [38], etc. Compared with PCB+RPP, our results increase by +5.98% on PCB+RPP mAP (56.7%) in detected version.

For the vehicle re-id task, we compare the current state-of-the-art methods with our proposed approach, and the total results are recorded in Table 2. Obviously, compared to the baseline, the proposed soft margin loss brings improvements on both mAP and Rank-1. In specific, there are +1.36% and +0.12% increase on mAP and Rank-1 respectively. Besides, our method surpasses the most competitive method QD-DLF by +12.53% and +6.49% on mAP and Rank-1. The results show that the proposed soft margin loss is also effective in the vehicle re-identification.

5) COMPARISONS WITH FIXED MARGIN

In the proposed loss, we use the “soft” margin to guide the training procedure. Here, we try to compare our method with fixed margin based loss functions. We choose the triplet loss and LMCL [8] for comparative experiments, since the two losses are both margin based methods and they represent the typical metric loss and classification loss in recent researches.

For the triplet loss, its formulation is given by:

\[
L_{TRI} = \frac{1}{C} \sum_{j=1}^{C} \max \left[ 0, \ max_{i \neq j} (f_i, W_j) - \min_{k \neq j} (W_k, W_j) + m \right].
\]

The definition of intra-class and inter-class distances is same with that in the soft margin loss. For fair comparison, we jointly optimize the normalized softmax and the triplet loss for model training. The total loss can be expressed as

\[
L = L_{NSL} + \lambda L_{TRI},
\]

where \( \lambda \) is the weight parameter and is set as 1 for simplicity. For LMCL with margin \( m \), it can be represented as:

\[
L_{LMCL} = \frac{1}{n} \sum_{i=1}^{n} \log e^{s \cdot \cos(\theta_i - m)} + \sum_{j \neq i}^{C} e^{s \cdot \cos(\theta_j)}.
\]

We keep all the previous experimental settings unchanged and substitute the soft margin loss with the triplet loss and LMCL. The comparative results with different margins are recorded in Table 3. From the results, we can see that the mAP values of “NSL+Triplet” on CUHK03 and VeRi-776 datasets change a lot along with margin \( m \). Besides it exhibits a saturation pheromone when \( m \) exceeds a large value, e.g. \( m > 1.0 \). LMCL has not obvious improvements on CHHK03 while obtains better accuracy on VeRi-776 with different margins. By comparison, our proposed soft margin loss can achieve preferable results without a margin parameter, which is close to the best performance from the margin based methods.

6) PARAMETER ANALYSIS

The parameter \( \lambda \) denotes the weight of the soft margin loss in the joint training scheme. To observe how \( \lambda \) impacts the model performance, we keep \( \varepsilon \) as 0.2 and vary \( \lambda \) from \{0.1, 0.2, 0.5, 1.0, 1.5, 2.0, 5.0\} for both CUHK03 and VeRi-776. The corresponding results are plotted in Figure 4. For CUHK03 labeled version, the value of mAP fluctuates with the increasing of \( \lambda \), and mAP achieves the largest value when \( \lambda = 1.0 \). While for detected version, the mAP keeps a steady and slight increase. In the results of VeRi-776 dataset, we can see that the mAP increases gradually as \( \lambda \) increases from 0.1 to 2.0 and declines when \( \lambda \) larger than 2.0.
We check the influences of the parameter $\varepsilon$ on the model performance with a similar manner. We set $\lambda$ as 1.0 and vary $\varepsilon$ from {0.01, 0.1, 0.2, 1.0, 1.5, 2.0} respectively. The corresponding mAP curves are shown in Figure 5. For CUHK03 labeled version, the best result is obtained when $\varepsilon$ is 0.2. While in CUHK03 detected version, the mAP curve roughly shows a decline tendency except a surge where $\varepsilon$ is 1.0. For VeRi-776, the mAP raises when $\varepsilon$ increases from 0.01 to 0.05 but has an obvious decrease when $\varepsilon$ is larger than 0.5.

V. DISCUSSION

Here, we first discuss the relations between our proposed method with two similar loss functions, including ArcFace [6] and triplet-center loss [13]. ArcFace maps the features and classification weights into the hypersphere with the normalization operation. Then it constrains the angles between features and their corresponding weight by adding a fixed margin for discriminative embedding learning. The triplet-center loss combines the triplet loss and center loss to maximize the intra-class compactness and inter-class separability simultaneously, and it performs discriminative embedding learning with a joint training scheme of softmax loss.

Compared to the abovementioned approaches, our proposed method has three following advantages. First, since the difference of maximal intra-class distance and the minimal inter-class distance is taken as one of the optimization objectives, the proposed soft margin loss can learn discriminative features without an empirically fixed margin. It is convenient and practicable for the soft margin loss to generalize itself in

### TABLE 3. Comparisons between our method and two fix margin based methods on CUHK03 and VeRi-776 datasets.

| Methods     | $m$ | CUHK03 labeled |       | CUHK03 detected |       | VeRi-776 |       |
|-------------|-----|---------------|-------|-----------------|-------|----------|-------|
|             |     | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 |
| NSL+Triplet | 0.01| 63.29 | 64.86 | 60.91 | 61.79 | 73.63 | 94.76 |
|             | 0.1 | 63.32 | 65.29 | 59.81 | 62.07 | 73.31 | 94.58 |
|             | 0.3 | 63.53 | 65.57 | 60.29 | 60.86 | 73.41 | 94.40 |
|             | 0.5 | 63.74 | 65.43 | 61.67 | 62.93 | 73.74 | 93.98 |
|             | 1.0 | 65.76 | 67.71 | 62.77 | 64.14 | 74.43 | 94.28 |
|             | 1.5 | 65.76 | 67.71 | 62.77 | 64.14 | 74.43 | 94.28 |
| LMCL        | 0.01| 63.12 | 63.71 | 61.42 | 63.36 | 73.04 | 94.40 |
|             | 0.1 | 63.55 | 66.50 | 62.54 | 63.64 | 74.92 | 95.11 |
|             | 0.3 | 63.32 | 64.64 | 60.71 | 62.00 | 75.32 | 94.87 |
|             | 0.5 | 62.15 | 64.21 | 58.49 | 60.43 | 75.13 | 94.93 |
|             | 1.0 | 61.86 | 63.50 | 58.46 | 59.79 | 74.54 | 95.17 |
|             | 1.5 | 61.92 | 63.36 | 60.22 | 62.21 | 74.62 | 94.76 |

For CUHK03 labeled version, the best result is obtained when $\varepsilon$ is 0.2. While in CUHK03 detected version, the value of mAP changes within a narrow range.
various applications. Second, in the soft margin loss, we treat the classification weights as the feature centers when we calculate the inter-class distance and intra-class distance. This operation is reasonable in the methodology and costless in the calculation. Third, since the features and classification weights are normalized in the hypersphere, the soft margin loss and the normalized softmax have identical optimization goals in the embedding space, i.e., the angle between learned features and weights. Therefore the joint training with these two losses can effectively alleviate the problem of low convergence and benefit discriminative embedding learning.

In this work, we design our loss function with the difference between the maximum intra-class distance and minimum inter-class distance for embedding learning. I think the discriminative power would be further enhanced if we consider an additional constraint on the intra-class compactness alone. So in the future research, we plan to explore new types of loss functions to improve the performance of deep embedding learning, and validate the methods on other related vision tasks, such as face recognition [8], texture classification [45], and so on.

VI. CONCLUSION

In this work, we propose a new loss function called soft margin loss for discriminative embedding learning. Specially, we first normalize the learned features and classification weights to map them into the hypersphere. With the normalization operation, the model prediction scores only depend on the angle between the feature and the weight, which is beneficial to the model convergence. Then the proposed soft margin loss is used to increase the between-class discrepancy and shrink the within-class compactness by constraining the difference of the maximal intra-class distance and the minimal inter-class distance. Finally, the soft margin loss and the normalized softmax are joined together to supervise the model for achieving discriminative and robust feature embedding. The proposed method can efficiently optimize the intra-class and inter-class distances of learned features with a soft margin and help for discriminative embedding learning. Extensive experiments on toy examples and re-identification tasks (e.g., person and vehicle re-identification) are conducted to illustrate the effectiveness of our method.

REFERENCES

[1] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 21–26.
[2] X. Fan, W. Jiang, H. Luo, and M. Fei, “SphereReID: Deep hypersphere manifold embedding for person re-identification,” J. Vis. Commun. Image Represent., vol. 60, pp. 51–58, Apr. 2019.
[3] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, arXiv:1703.07737. [Online]. Available: http://arxiv.org/abs/1703.07737
[4] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, “Embedding adversarial learning for vehicle re-identification,” IEEE Trans. Image Process., vol. 28, no. 8, pp. 3794–3807, Aug. 2019.
[5] J. Zhu, H. Zeng, Y. Du, Z. Lei, L. Zheng, and C. Cai, “Joint feature and similarity deep learning for vehicle re-identification,” IEEE Access, vol. 6, pp. 43724–43731, Aug. 2018.
[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 15–20.
[7] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 7–12.
[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “CosFace: Large margin cosface loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.
[9] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 21–26.
[10] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), New York, NY, USA, Jun. 2006, pp. 17–22.
[11] R. Varior, M. Haloi, and G. Wang, “Gated siamese convolutional neural network architecture for human re-identification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 11–14.
[12] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), New York, NY, USA, Jun. 2016, pp. 19–24.
[13] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, “Triplet-center loss for multi-view 3D object retrieval,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.
[14] Y. Wen, Z. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 11–14.
[15] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 27–30.
[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, Jun. 2014, pp. 23–28.
[17] H. Choi, A. Som, and P. Turaga, “AMC-loss: Angular margin contrastive loss for improved explainability in image classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Seattle, WA, USA, Jun. 2020, pp. 14–19.
[18] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, “A strong baseline and batch normalization neck for deep person re-identification,” IEEE Trans. Multimedia, vol. 22, no. 10, pp. 2597–2609, Oct. 2020, doi: 10.1109/TMM.2019.2958756.
[19] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. Hinton, “Regularizing neural networks by penalizing confident output distributions,” 2017, arXiv:1701.06548. [Online]. Available: http://arxiv.org/abs/1701.06548
[20] N. Hor and S. Fekri-Eshad, “Image retrieval approach based on local texture information derived from predefined patterns and spatial domain information,” Int. J. Comput. Sci. Eng., vol. 8, no. 6, pp. 246–254, Nov./Dec. 2019.
[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[22] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Jun. 2014, pp. 23–28.
[23] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 21–26.
[24] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in Proc. AAAI Conf. Artif. Intell. New York, NY, USA: Hilton New York Midtown, Feb. 2020, pp. 13001–13008.
[25] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 11–14.
[26] X. Liu, W. Liu, H. Ma, and H. Fu, “Large-scale vehicle re-identification in urban surveillance videos,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Seattle, WA, USA, Jul. 2016, pp. 11–15.
[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 27–30.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, Jun. 2009, pp. 20–25.

[29] Pytorch. Accessed: Dec. 23, 2019. [Online]. Available: https://pytorch.org/

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015, pp. 7–9.

[31] R. Yu, Z. Zhou, S. Bai, and X. Bai, “Divide and fuse: A re-ranking approach for person re-identification,” in Proc. Brit. Mach. Vis. Conf. (BMVC), London, U.K., Sep. 2017, pp. 4–7.

[32] Z. Zheng, L. Zheng, and Y. Yang, “Pedestrian alignment network for large-scale person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[33] Y. Sun, L. Zheng, W. Deng, and S. Wang, “SVDDNet for pedestrian retrieval,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 22–29.

[34] Y. Chen, X. Zhu, and S. Gong, “Person re-identification by deep learning multi-scale representations,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Venice, Italy, Oct. 2017, pp. 22–29.

[35] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.

[36] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Mask-guided contrastive attention model for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.

[37] X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorization net for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.

[38] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 8–14.

[39] Y. Zhou and L. Shao, “Cross-view GAN based vehicle generation for re-identification,” in Proc. Brit. Mach. Vis. Conf. (BMVC), London, U.K., Sep. 2017, pp. 4–7.

[40] Y. Zhou and L. Shao, “Viewpoint-aware attentive multi-view inference for vehicle re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.

[41] X. Liu, W. Liu, T. Mei, and H. Ma, “PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance,” IEEE Trans. Multimedia, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[42] J. Zhu, H. Zeng, Z. Lei, S. Liao, L. Zheng, and C. Cai, “A shortly and densely connected convolutional neural network for vehicle re-identification,” in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), Beijing, China, Aug. 2018, pp. 20–24.

[43] X. Liu, S. Zhang, Q. Huang, and W. Gao, “RAM: A region-aware deep model for vehicle re-identification,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), San Diego, CA, USA, Jul. 2018, pp. 23–27.

[44] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, “Vehicle re-identification using quadruple directional deep learning features,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 1, pp. 410–420, Jan. 2020.

[45] S. Fekri-Ershad, “Bark texture classification using improved local ternary patterns and multilayer neural network,” Expert Syst. Appl., vol. 158, Nov. 2020, Art. no. 113509.