Pedestrian Re-ID Based on Improved Triplet Loss

Zhengjun Xu* and Yunfei Wang
School of Electronic and Information Engineering, Anhui University, Hefei 230601, China

*Corresponding author. Email: 1203997536@qq.com

Abstract. Person re-identification technology is an important foundation for security, pedestrian tracking and other fields, and is the key to building a safe and smart city. In recent years, a large number of researchers have trained pedestrian re-identification networks through triplet loss, especially the triplet loss with batch hard mining (BHTri loss) has greatly improved person re-identification networks on accuracy. However, triplet loss with batch hard mining for an anchor sample only select the hardest positive sample and the hardest negative sample to calculate the loss, ignoring the influence of other samples on network parameters. In response to the above issues, this paper proposes a variant of triplet loss with batch hard mining, which is called adaptive weight triplet loss with batch hard mining. After the training dataset extracts features from the backbone network, in the phase of calculating loss, it takes the average of the sum of the distances between an anchor sample and all corresponding positive samples as a threshold, and both positive samples with an anchor point distance greater than the threshold and negative samples smaller than the threshold are retained, then based on The distance of the anchor sample is given the corresponding sample weight for calculating the loss. Compared with BHTri, mAP have improved 1.79%, 2.04%, and 1.25% respectively on the Market1501, DukeMTMC-reID and CUHK03 datasets, indicating that the proposed algorithm is effective.

Keywords: triplet loss, person re-identification, sample weights.

1. Introduction
Person re-ID (person re-Identification) is an important research topic in the field of computer vision, which refers to correlating images of the same pedestrian under different camera perspectives. Person re-identification is still a challenging task due to the effects of lighting, background, obstructions, differences in camera parameters, and changes in pedestrian poses.

Person re-ID tasks usually have two phases: feature extraction and distance measurement. With the rapid development of deep learning, more and more excellent models based on convolutional neural network (CNN) have appeared. These models can automatically extract features from images based on tasks, which is just suitable for person re-ID Feature extraction phase. The method of using deep learning network to extract strong discriminative features and then calculating the distance between person is called the person re-ID method based on representation [1-4]. Zheng et al. [5] classified the representation-based person re-ID method into classification and verification models. The classification model uses pedestrian identity information as a category and generally uses cross entropy loss to train the network, the verification model uses the network to determine whether the input paired pictures is the same person.
Metric learning is different from representation learning. It is a method widely used in the field of image retrieval. It aims to learn the similarity of two images through the network. On the problem of person re-ID, the similarity of different pictures of the same person is greater than that of different pedestrians. Finally, the loss function of the network makes the distance between pictures of the same pedestrian as small as possible and the distance between different pedestrian pictures as large as possible. Varior et al. [6] used Contrastive Loss to train the Siamese network to obtain the similarity between the two images; Schroff et al. [7] used deep learning networks to extract features in the face recognition task and instead of using SVM and other methods for classification, it uses the proposed triplet loss optimization feature to achieve better performance than traditional methods; Cheng et al. [8] only required the inner class distance in the initial triplet loss It is improved based on less than the inter-distance, requiring the inner-class distance to be less than a value at the same time; Chen et al. [9] was inspired by the loss of triples and proposed adding a negative sample with different identity information to form a quadruplet. Compared with the triple loss, which only considers the relative distance between the positive and negative samples, the second item added by the quadruplet loss does not share identity information, which takes into account the absolute distance between the positive and negative samples. Hermans et al. [10] randomly select three pictures from the training data based on the traditional triplet loss. The method is simple, but most of the samples are simply divided into regions, a large number of simple training pairs are not conducive to better learning of the network, and triplet loss with batch hard mining, BHTri Loss, is proposed; Zhang et al. [11] combined the quadruplet loss with BHTri Loss, both the relative distance and the absolute distance are taken into account, and the marginal sample mining loss (MSML) of difficult sample sampling is introduced. Only the most difficult positive sample pair and the most difficult negative sample pair are selected for calculating the loss. Ristani et al. [12] inspired the BHTri Loss and proposed an adaptive weighted triplet loss.

This paper is inspired by the ideas of Ristani et al. Combined with the mean of the distance sum of all positive sample pairs in the quadruple loss as a constraint, both positive pairs with a distance greater than \( d \) and negative pairs with a distance less than \( d \) are reserved for calculation loss, as shown in Figure 2. The positive samples in the red box are the positive samples whose distance from the anchor is greater than \( d \). The negative samples in the green box are the negative samples whose distance from the anchor is less than \( d \).

\[
d = \frac{1}{k} \sum_{a \in P(a)} d(x_a, x_i)
\]

(1)

In the end, the paper mainly contributes to the following two points: (1) Considering the absolute distance between positive and negative samples in the paper [9], ignoring most simple samples in the paper [10], and calculating only the triples of difficult positive and negative samples The loss and the advantage of the paper [12] assigning an adaptive weight to each positive and negative sample, proposed the average value of the sum of the distances of all positive samples in a batch as a threshold, and the distance between the positive sample and the anchor is greater than the threshold, the samples whose distance between the negative and the anchor is less than the threshold are retained, and adaptive weights are obtained according to formulas (3) and (4) for calculating the improved triple loss that called AW-BHTri Loss; (2) the proposed method is in three a large number of experiments have been performed on the main person re-ID dataset. Rank-1 has improved by 1.79%, 2.42% and 0.85% on Market1501, DukeMTMC-reID and CUHK03 datasets respectively.

2. Improved triplet loss function

2.1. Algorithm source

Hermans et al.’s BHTri Loss method, for each training batch, randomly select P person with different identities, and each person chooses K different pictures to form one batch. Then for each image in the batch, a positive sample with the furthest distance and a negative sample with the closest distance are
selected to form a triplet. Ristani et al.'s idea, the distance between an anchor and K positive samples and (P-1) K negative samples are respectively calculated by formula (3) and formula (4) in the text to obtain K and (P-1) K weight coefficient, the sum of the product of the distance of the K positive samples and the corresponding weighting coefficient and (P-1) the sum of the product of the K negative samples and the corresponding weight coefficient for the calculation of the triplet loss. This paper combines the ideas of the above two algorithms, and proposes a new adaptive weight triplet loss AW-BHTri. After the training set extracts features from the backbone network, during the loss calculation phase, for each anchor, the average value of the sum of the distances is used as a threshold value, and positive samples with a distance from the anchor point sample larger than the threshold value and negative samples smaller than the threshold value are retained, and then a corresponding sample weight is given to calculate the loss according to the distance from the anchor.

2.2. Selecting positive and negative samples
For a given batch of P × K size, first obtain the features of each image through the backbone network, then use one sample in P as the anchor, calculating between the anchor and the remaining K positive samples, taking the mean value d of the sum of the distances of K positive sample pairs as the threshold, selecting positive sample with anchor distance greater than d and negative samples with a distance less than d for the calculation of the triplet loss, \( P(a) \) in formula (1) represents the data set that belongs to the same pedestrian as the anchor.

![Figure 1. Sampling strategy of TriHard Loss.](image1)

![Figure 2. Our sample strategy.](image2)
2.3. Adaptive weights

As shown in Figure 3, the blue histogram and its height in the figure represent the distance between the corresponding four positive samples and the anchor, and the gray histogram represents the weight coefficient assigned by the positive samples. (a) the probability that each positive sample in the traditional triplet is the same; (b) the BHTri represents the positive sample with the largest distance from the anchor to form a triple; (C) representing each positive sample in [12], and assigning weight coefficients to the four positive samples according to formula (3); (d) The graph represents the positive samples selected in this paper and the weight factor.

Figure 3. Assigning weights for positive samples.

Figure 4 represents the selection of negative samples and the assignment of weight coefficients. (a): Each sample with a different identity from the anchor can be used as a negative sample; (B): BHTri selects the samples with different identities closest to the anchor as negative samples; (c): All negative samples are used, and different weight coefficients are assigned according to formula (4) according to their distance from the anchor; (d): The selection method of negative samples in this paper, the weight coefficient allocation method is similar to (c).
2.4. AW-BHHard loss

For a person re-ID dataset with labels, learning its representative features through improved triple loss. For an anchor \( x_a \), positive sample \( x_i \in P(a) \), negative sample \( x_j \in N(a) \), the triple loss function redefined in this paper is as follows:

\[
L_{AW-TriHard} = \frac{1}{P \times K} \sum_{a \in \text{batch}} \left[ \sum_{i \in P(a)} d(x_a, x_i) - \sum_{j \in N(a)} d(x_a, x_j) + \alpha \right],
\]

In Eq.2, \( \alpha \) represents the given inter-class distance, \( d(\cdot, \cdot) \) represents the Euclidean distance between the two images, and \([\cdot]\), represents a non-negative calculation result. The value is 0. The newly defined triplet loss function in this paper has the following two advantages: (1) Compared with the traditional triplet loss function, this article discards a large number of simple samples that will affect the model's generalization, compared with the BHTri, AW-BHTri loss function only uses the same kind of positive samples that are farthest from the anchor point samples and negative samples of different labels nearest to the anchor. As a result, the samples between the difficult samples and the simple samples are ignored. This article focuses on each anchor, for samples, the average value \( d \) of the sum of the distances of all positive samples and anchor is used as the threshold, and the positive samples with a distance from the anchor greater than the threshold \( d \) and the negative samples with a distance less than \( d \) from the anchor are used as the triplet loss. At the same time, assign more weight to positive and negative samples that are more difficult to distinguish in order to expect better feature characterization. (2) Positive and negative samples can achieve class balance through weight distribution.

\[
W_{i \in P(u)} = \frac{e^{d(x_a, x_i)}}{\sum_{s \in P(u)} e^{d(x_a, s)}} \quad (3)
\]

\[
W_{j \in N(u)} = \frac{e^{-d(x_a, x_j)}}{\sum_{s \in N(u)} e^{-d(x_a, s)}} \quad (4)
\]

3. Experiment

In order to verify the effectiveness of the adaptive weighted triplet loss proposed in this paper, the experiments are performed on three common person re-ID datasets Market-1501[13], CUHK03[14] and DukeMTMC-reID [15].

3.1. Evaluation criteria

Commonly used evaluation indicators in person re-ID tasks are cumulative matching characteristic curve (CMC) and mean average precision (mAP). The average accuracy of each query is from its accuracy rate-recall, and mAP is the average of the average accuracy of all queries, that is, CMC reflects the retrieval accuracy, and mAP reflects the recall rate. Rank-n in CMC refers to the first n pictures of the matching result. Accuracy, the above three data sets are evaluated by the Single-Person (Single Query) query in the search database for matching person. In the experiment, Rank-1, Rank-5, Rank-10 and mAP are given in three different datasets.

3.2. Experimental details

In the training phase, this article sets \( P = 16 \) and \( K = 4 \) for the batch size, and sets it to \( 256 \times 128 \) before inputting the data to the network, and flips it randomly horizontally. The network uses the Resnet50 network and initializes the Resnet50 network with the ImageNet pre-trained network parameters in the dataset. The number of training iterations is 70. During the training process, the
initial learning rate of the backbone network was set to 0.001, the initial learning rate of the fully connected layer and the classification layer was set to 0.01, and the learning rate was set to 0.0001 after 40 iterations of training. Stochastic Gradient Descent (SGD) was used to update the parameters for each batch. The momentum was set to 0.9 and the weight decay was set to 0.0005. The experimental platform is 64-bit Ubuntu-16.04 operating system, NVIDIA GTX 1060 GPU, deep learning framework is Pytorch, and Python version is 3.6.

3.3. Experimental results

In this section, the experimental results of the proposed algorithm are compared with the baseline and the experimental results of current mainstream algorithms on three commonly used person re-ID data.

3.3.1. Comparison of experimental results with baseline. On the Market-1501 dataset, the rank-1 and mAP of BHTri as the baseline are 89.25% and 75.92%, respectively. The AW-BHTri improved 1.34% and 1.79% on the baseline, as shown in Table 1.

<table>
<thead>
<tr>
<th>method</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>rank-10</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHTri</td>
<td>89.25</td>
<td>96.17</td>
<td>97.48</td>
<td>75.92</td>
</tr>
<tr>
<td>AW-BHTri</td>
<td>90.59</td>
<td>96.02</td>
<td>97.51</td>
<td>77.71</td>
</tr>
</tbody>
</table>

On the DukeMTMC-reID dataset, the results are shown in Table 2. The experimental results in this article have achieved a greater improvement in rank-1 and mAP on the DukeMTMC-reID dataset, compared with rank-1 of 80.43% and mAP of 65.27%. The benchmark of this paper is 2.42% and 2.04% respectively.

<table>
<thead>
<tr>
<th>method</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHTri</td>
<td>80.43</td>
<td>89.72</td>
<td>92.68</td>
<td>65.27</td>
</tr>
<tr>
<td>AW-BHTri</td>
<td>82.85</td>
<td>90.48</td>
<td>93.40</td>
<td>67.31</td>
</tr>
</tbody>
</table>

On the CUHK03 dataset, compared with benchmarks with rank-1 of 68.29% and mAP of 62.64%, this paper has achieved 0.85% and 1.25% improvement on rank-1 and mAP, as shown in Table 3.

<table>
<thead>
<tr>
<th>method</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHTri</td>
<td>68.29</td>
<td>84.79</td>
<td>90.79</td>
<td>62.64</td>
</tr>
<tr>
<td>AW-BHTri</td>
<td>69.14</td>
<td>84.86</td>
<td>91.29</td>
<td>63.89</td>
</tr>
</tbody>
</table>

3.3.2. Comparison with existing methods. Table 4 shows the comparison between the improved adaptive weight triplet loss and the existing methods on the Market-1501 dataset. The results are public results of the Single Query evaluation method on the dataset homepage or in the literature. The results show that the improved triple loss proposed in this paper is effective for the improvement of person re-ID tasks.

218
Table 4. Comparison with existing methods on Market-1501 datasets %.

<table>
<thead>
<tr>
<th>method</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verif_Identif + LSRO[16]</td>
<td>83.97</td>
<td>66.07</td>
</tr>
<tr>
<td>Basel+LSRO[17]</td>
<td>78.06</td>
<td>56.23</td>
</tr>
<tr>
<td>Fusion[18]</td>
<td>80.31</td>
<td>57.53</td>
</tr>
<tr>
<td>SSM[17]</td>
<td>82.21</td>
<td>68.80</td>
</tr>
<tr>
<td>Chen et al.[19]</td>
<td>87.9</td>
<td>95.3</td>
</tr>
<tr>
<td>BHHard</td>
<td>89.25</td>
<td>75.92</td>
</tr>
<tr>
<td>AW-BHTri</td>
<td>90.59</td>
<td>77.71</td>
</tr>
</tbody>
</table>

4. Conclusion

This article proposes to use the average of the distance sum of all positive sample pairs of each anchor as the threshold of the corresponding anchor within a batch, and the positive samples with a distance greater than the threshold from the anchor and the negative samples with a distance less than the anchor are all retained, then assigning weight coefficient according to the distance, which is used to improve the loss of the triples, and a more discriminative feature is obtained. Experiments were performed on the Market-1501, DukeMTMC-reID, and CUHK03 datasets. Finally the experimental results show that the improvement has performed well.

Reference

[12] E Ristani and C Tomasi, Features for Multi-target Multi-camera Tracking and


