MHEF-TripNet: Mixed Triplet Loss with Hard Example Feedback Network for Image Retrieval

Xuebin Yang[1], Shouhong Wan[1,2](✉), Peiquan Jin[1,2], Chang Zou[1], and Xingyue Li[1]

[1] School of Computer Science and Technology,
University of Science and Technology of China, Hefei 230027, China
yangxb@mail.ustc.edu.cn

[2] Key Laboratory of Electromagnetic Space Information,
Chinese Academy of Science, Hefei 230027, China

Abstract. Image retrieval has made significant advances, fueled mainly by deep convolutional neural networks, but their training procedure is not efficient enough. Because of the large imbalance between easy examples and hard examples, networks lack direct guidance information from hard examples. In this paper, we solve the problem by developing an effective and efficient method, called mixed triplet loss with hard example feedback network (MHEF-TripNet). Since the proportion of hard examples is small, a sample selection probability matrix is introduced to select hard examples, which assists a network to focus more on enlarging the gap between the confusing categories in triplet loss. And it will be adjusted according to the feedback of test results after each training iteration. Furthermore, a mixed triplet loss function is proposed, which combines triplet loss with category loss to take advantage of association information between images and category information. The effectiveness of MHEF-TripNet is confirmed by experimentation on UC Merced Land Use and Kdelab Airplane datasets. Compared with previous image retrieval approaches, our approach obtains superior performance.

Keywords: Triplet loss · Probability matrix of sample selection · Image retrieval

1 Introduction

Image retrieval is one of the greatly worthwhile computer vision tasks. It pays more attention to the image similarity, and is suitable for retrieving images from a massive image database, Content based Image Retrieval (CBIR) [5–7] is the mainstream method for image retrieval at present. The idea of CBIR is as follows. First, extract features from the query image. After that, use features to calculate the similarity between the query images and the images in database. And then sort the images in descending order by similarity. Finally, regard the result as the feedback to further improve the performance of feature extraction. Particularly, feature extraction and

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 35–46, 2019.
https://doi.org/10.1007/978-3-030-34113-8_4

similarity measurement are two important processes in CBIR, which determine the accuracy and efficiency of image retrieval methods.

CBIR approaches are often trained through a reduction that converts image retrieval into an image classification problem, and then take an intermediate bottleneck layer as extracted feature representation used to retrieving images. Although classification loss could increase the difference between classes, it cannot narrow the inner-class distance effectively. To solve this dilemma, FaceNet [1] proposes a novel triplet loss which calculates loss by learning a Euclidean embedding, it makes a great progress especially in face recognition. This method uses the squared L2 distance according to the image similarity: the larger the distance between two images, the less the similarity. The key of the method is to enlarge the distance between the images from different categories and narrow the distance between the images from the same category. However, new problem has arisen: the training set is distinguished by a large imbalance between easy examples and hard examples, where easy examples are the images with distinguishable deep features from different categories, while hard examples are the images with similar deep features from confusing categories.

Many studies [2, 3] have shown that hard examples (samples which are difficult to be distinguished by models) are beneficial to network convergence, since network frequent misclassify hard examples, which propagates back more loss. Thus, mining hard examples play an important role in model training. However, [1] selects training dataset in a random way, which might be inefficient to get hard examples, because the proportion of hard examples is small. The straightforward way to choose hard examples is to traverse the entire dataset, but the complexity of this method is too high to be directly applied. To tackle this issue, Hermans et al. [4] propose a variant triplet loss that provides a new hard examples selection method, which select hard examples by traversing a batch of data. Although it reduces the randomness of sample selection, it only considers the images in each batch and the complexity of this method is still too high. Wherefore mining hard examples is still a challenge for network training in triplet loss. In addition, triplet loss only utilizes the associated information between images, which does not make full use of the classification information.

To circumvent the limitations embedded in the existing triplet loss networks, we propose a novel network for image retrieval called mixed triplet loss with hard examples feedback network (MHEF-TripNet). Different from triplet networks, our method introduces a sample selection probability matrix to select hard examples. The matrix is used to select a different category that has the maximum similarity of the known category. After each iteration, we adjust the sample selection probability matrix according to the feedback of test results, and then the matrix can select hard examples more accurately. Also, we propose a mixed loss function [17], which combines triplet loss with category loss to extract discriminative features. The two main contributions of proposed method can be summarized as follows.

- A probability matrix of sample selection is introduced to choose hard example pairs. Additionally, the probability matrix will be updated according to the test results of the model after each iteration.

- The proposed mixed triple loss takes advantage of association information between images and category information simultaneously, so that more distinctive features can be learned.

The remainder of this paper is organized as follows: we summarize the related work of image retrieval in Sect. 2. The formulation of proposed MHEF-TripNet is described in Sect. 3. Section 4 shows our experimental results and their corresponding analysis. At last, we give the conclusion of this work in Sect. 5.

2 Related Work

Image feature extraction is very important for image retrieval. The ability of extracting features has made significant advances riding on the wave of convolutional neural network (CNN), which has achieved great success in target recognition [8], target detection, image segmentation [9], natural language understanding [10] and other fields. Previous image retrieval approaches based on deep networks use a classification layer [11–13] trained over a set of known categories and then take an intermediate bottleneck layer as a representation used to retrieving images. The downsides of this approach are its indirectness and its inefficiency: the bottleneck representation cannot generalize well to new categories. Triplet loss is proposed to solve this dilemma.

Triplet loss [1] is a metric learning method which is first introduced by Google along with FaceNet. The sketch of triple loss learning target is illustrated in Fig. 1. Triplet consists of the following components: a sample which is randomly selected from the training set is regarded as an anchor, other two samples which have the same class as the anchor denotes a positive and with different class represents a negative. Before training, the relationship between the three may be similar to the left part of the figure, where the negative is closer to the anchor than the positive. After training, the positive becomes much closer to the anchor, like the right part of the figure. In a word, the triplet loss aims to narrow the distance between an anchor and a positive and enlarge the distance between the anchor and a negative. The traditional triplet model randomly selects three samples from the training set, which is simple but too random. The key point in triple training is to find out the hard triplets, that is, an anchor together with a remote positive (hard positive) and a close negative (hard negative). Since the proportion of hard triplets is low, randomly selection might not effectively get the hard triplet, causing poor performance.

Hermans et al. [4] propose a variant of triplet loss that provides a new hard examples selection based on batch training. The main idea of this method is to select hard triplet from batches. First, randomly select P classes, and then randomly select K images from each class, thus forming a batch of $P * K$ images. As for each sample in the batch, a selected triplet can be consisted of the sample and its hardest negative and hardest positive within the batch. In this way, the selection randomness of the traditional method can be reduced to a certain extent, making it more conducive to model training. Although this approach avoids the randomness of triple sampling to a certain extent, it only enlarges the sampling range locally, and cannot guarantee that the difficult sample pair is optimal.

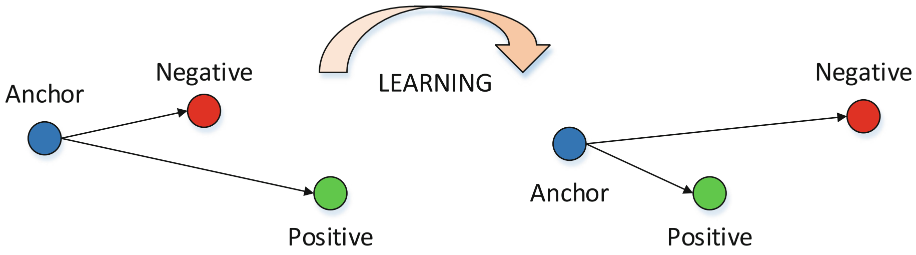

Fig. 1. Sketch of triple loss learning target

3 MHEF-TripNet

We propose MHEF-TripNet for effective image retrieval. We argue that single triplet loss is inefficient and the current way of triple sampling is suboptimal. Our method introduces a sample selection probability matrix to select hard examples and a mixed loss function combines triplet loss with category loss to extract discriminative features.

The framework of the proposed method is shown in Fig. 2, and the network is called MHEF-TripNet. In the training stage, similar to the traditional triple training, three images are transmitted at the same time, but the selection of the three images is different. Specifically, an image is randomly selected from the training data as an anchor. The category of the positive is consistent with the anchor, while the category of the negative is selected according to the probability matrix. The probability matrix is an $N \times N$ matrix, where N denotes the number of categories, the element V_{ij} denotes the probability of choosing j as a category of a negative when the anchor category is i. And each row adds up to 1. The selected three images are simultaneously fed into the same convolution neural network for feature extraction. After that, the network parameters are optimized by the mixed loss which consists of triplet loss and classification loss.

In order to better demonstrate the experimental results of our method, Table 1 presents details of a simple network architecture which is the backbone of MHEF-TripNet.

3.1 Sample Selection Probability Matrix

Although there exist some methods for hard example sampling in recent years, most of these methods focus on expanding the sampling range locally which only partly improve the sampling performance. Instead, in this paper, the selection is guided by f sample selection probability matrix globally to get hard examples with high generalization and pertinent. The training set and validation set are regarded as the input, while the network parameters for feature extraction are considered as the output.

Firstly, the images are preprocessed, including size clipping and normalization. In addition, the sample selection probability matrix is initialized as follows:

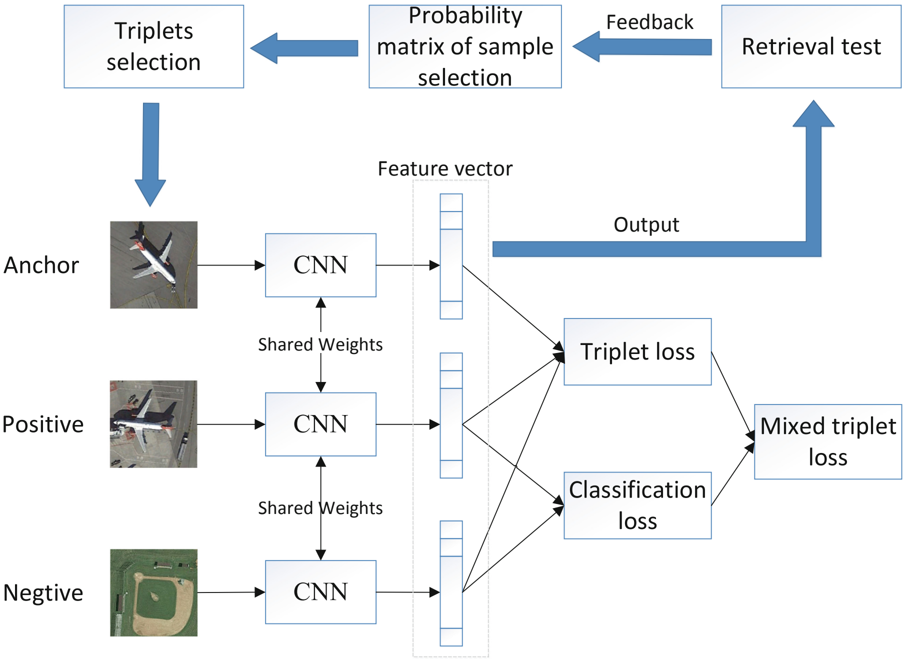

Fig. 2. The framework of MHEF-TripNet

Table 1. Network architecture: the backbone of MHEF-TripNet.

Layer name	Input size	Kernel	Stride	Padding	Output size
Conv_1	128 × 128 × 3	4 × 4	2	1	64 × 64 × 32
BN_1	64 × 64 × 32	–	–	–	64 × 64 × 32
Conv_2	64 × 64 × 32	4 × 4	2	1	32 × 32 × 64
BN_2	32 × 32 × 64	–	–	–	32 × 32 × 64
Conv_3	32 × 32 × 64	4 × 4	2	1	16 × 16 × 128
BN_3	16 × 16 × 128	–	–	–	16 × 16 × 128
Conv_4	16 × 16 × 128	4 × 4	2	1	8 × 8 × 256
BN_4	8 × 8 × 256	–	–	–	8 × 8 × 256
Conv_5	8 × 8 × 256	4 × 4	2	1	4 × 4 × 512
BN_5	4 × 4 × 512	–	–	–	4 × 4 × 512
Conv_6	4 × 4 × 512	4 × 4	2	1	1 × 1 × 1024
BN_6	1 × 1 × 1024	–	–	–	1 × 1 × 1024
Feature layer	1 × 1 × 1024	1 × 1	1	0	128
Classification layer	128	1 × 1	1	0	N

$$\begin{bmatrix} 0 & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & 0 & \cdots & \frac{1}{N-1} \\ \cdots & \cdots & 0 & \cdots \\ \frac{1}{N-1} & \cdots & \cdots & 0 \end{bmatrix} \tag{1}$$

Where N denotes the number of categories of training images,V_{ij} denotes the probability of choosing j as a category of a negative when the anchor category is i. $V_{ii} = 0,\ i \in (1, N),\ V_{ij} = \frac{1}{N-1},\ i \neq j,\ i, j \in (1, N)$, since the probability is uniformly distributed.

After that is the iterative process of training. In each iteration, the triples are sampled by the current probability matrix. More specifically, a sample is randomly selected from the training data as an anchor. Suppose the category of the anchor is i, the probability of sampling a negative with j category is V_{ij}. A positive is selected from images with i category. This three images form a triple. The number of selected triples in each iteration is consistent to the size of training batch.

The triples are fed into the feature extraction network to optimize the parameters, and the feature extraction network parameters of the current iteration times are obtained. The training set image features are extracted from the current feature extraction network as a temporary feature database. For the verification set, the same feature is extracted. The image retrieval tests are carried out on the feature database one by one, and the relevant feedback data are counted. Feedback data refers to the statistical results of the misclassification of each image in this round of image retrieval test. The result of misclassification is analyzed and the probability matrix is updated. The updated formula is as follows:

$$\begin{aligned} V_{ij} &= P(W) \times \frac{Num_j}{M}, i \neq j \\ V_{io} &= \frac{P(R)}{N-1-M}, i \neq o, j \neq o \end{aligned} \tag{2}$$

Where N denotes the number of categories of training images, i denotes the category of the current test sample, M denotes the number of categories which is result sequence retrieval, $P(R)$ is the accuracy, $P(W)$ is the error, Num_j represents the number of images which category is j.

Take a test sample belonging to i category as example, suppose there are W images being misclassified among K images from two categories $(M = 2)$, and W_1 images belong to the p category, W_2 images belong to the q category $(W = W_1 + W_2)$. The probability of correct classification is $P(R) = \frac{K-W}{K}$, and the probability of misclassification is $P(W) = \frac{W}{K}$, where p and q account for $\frac{W_1}{W}$ and $\frac{W_2}{W}$, respectively. Consequently, the probability matrix is updated as follows:

$$\begin{aligned} V_{ip} &= P(W) \times \frac{W_1}{W} \\ V_{iq} &= P(W) \times \frac{W_2}{W} \\ V_{io} &= \frac{P(R)}{N-1-2} \end{aligned} \tag{3}$$

The total probability is:

$$\begin{aligned} Total &= V_{ip} + V_{iq} + V_{io} \times (N - 1 - 2) \\ &= P(W) + P(R) \\ &= 1 \end{aligned} \tag{4}$$

Each modification of the probability matrix is a positive feedback of the test results. In this way, MHEF-TripNet focuses more on enlarge the gap between the confusing categories, thus improve the distinction of features, and finally improves the accuracy of image retrieval.

3.2 Mixed Triplet Loss

After extracting image features, the traditional triple method iteratively updates the network parameters by using the loss of distance comparison between feature vectors. This method is suitable for the situation that the number of image categories is constantly changing, or the training data does not contain the category label information, only whether the two images belong to the same category of comparative information. For most of image datasets, the number of image categories is fixed, and all of them have image label information. Therefore, MHEF-TripNet considers to integrate the category loss of images into the training process of the network, and combines the comparative loss to form a hybrid loss training network.

The classification loss is defined as follows:

$$C_{loss} = J(p) + J(n) \tag{5}$$

Where J denotes softmax loss, p and n represent the features of the positive and the negative, respectively. And the mixed loss is defined as follows:

$$M_{loss} = \alpha T_{loss} + \rho C_{loss} \tag{6}$$

Where T_{loss} and C_{loss} denote triplet loss and classification loss, respectively. α, ρ are the corresponding weights. Particularly, in order to balance this two losses, we set α = 2.0, β = 1.0, since the classification loss consists of two parts of softmax loss.

4 Experiments and Analysis

To evaluate the performance of MHEF-TripNet, we design a set of image retrieval experiments on two datasets: UC Merced Land Use and Kdelab Airplane. UC Merced Land Use dataset is land use image dataset meant for research purposes. It is a remote sensing dataset provide by [14] and include 21 classes. Kdelab Airplane dataset is created by our laboratory focus on retrieval airplane, which contains 11 different airplane types. We choose mean average precision (mAP), top 5 precision (P@5), top 10 precision (P@10), top 50 precision (P@50) and 100 precision (P@100) as evaluation criteria. In our retrieval results tables, TripNet denotes the image retrieval network based on triplet loss, M-TripNet denotes the TripNet with mixed triplet loss, HEF-TripNet denotes the TripNet with Hard Example Feedback and MHEF-TripNet represents the combination of the last two. The parameters of the training stage are set as follows: batch size is 64, iteration number is 100, learning rate is 10^{-4}.

4.1 Retrieval Results on UC Merced Land Use

The UC Merced Land Use is one of the most widely used remote sensing image datasets in the field of remote sensing. The dataset contains 2100 images, covering 21 different remote sensing scene categories, each with 100 images. The size of each image is 256 * 256. Figure 3 shows sample images of the dataset. The experiment is divided into two parts. The first part is the comparison with methods based on deep networks use a classification layer. The second part is ablation experiments based on the method proposed in this paper.

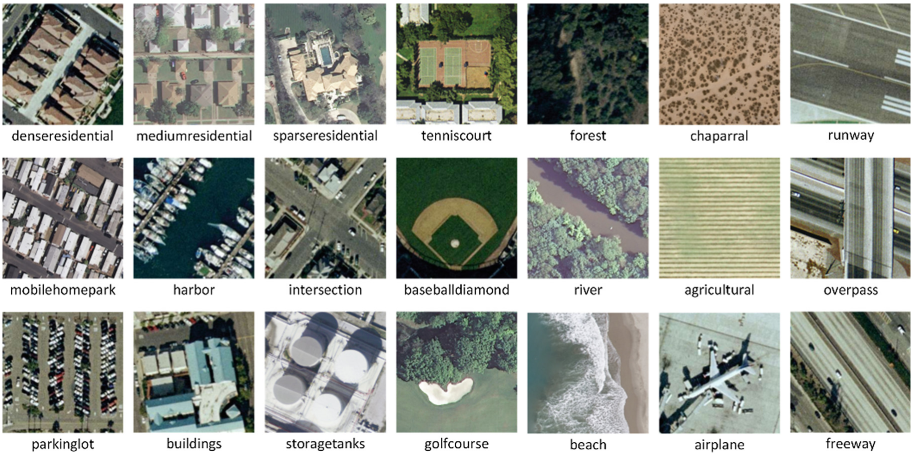

Fig. 3. Sample images of UC Merced Land Use dataset

To further evaluate the power of these methods, we have fine-tuned the networks to the remote sensing domain by using 80% of AID [16] dataset as training set. AID is a remote sensing dataset which is made of aerial image dataset collected from Google Earth imagery. It has a number of 10000 images within 30 classes and about 200 to 400 samples of size 600 * 600 in each class. Afterwards we use 100% of the UC Merced Land Use dataset as the test set. The experimental results are shown in Table 2. VGG16, VGG19, GoogleNet, ResNet-50, ResNet-101, and ResNet-152 are all experimental data from the [15] which is a summary of CBIR in remote sensing images. After comparative analysis, MHEF-TripNet proposed in this paper is the best in the most evaluation criteria, and the mAP is about 1.0% higher than other method. This indicates that MHEF-TripNet extracts discriminative features effectively on the UC Merched Land Use dataset.

Table 2. UC Merced Land Use dataset retrieval results of different approaches

Algorithm	mAP(%)	P@5(%)	P@10(%)	P@50(%)	P@100(%)
VGG16 [15]	52.46	83.91	78.34	61.38	49.78
VGG19 [15]	51.95	82.84	77.60	60.69	49.16
GoogleNet [15]	55.86	85.36	80.96	64.71	52.36
ResNet-50 [15]	56.57	88.26	84.00	65.92	52.69
ResNet-101 [15]	56.63	88.49	83.53	65.69	52.83
ResNet-152 [15]	56.03	88.42	83.08	64.65	52.50
TripNet [1]	57.04	88.95	84.58	66.06	53.09
M-TripNet	59.93	**90.24**	84.61	67.13	54.97
HEF-TripNet	58.31	89.05	84.43	66.73	54.23
MHEF-TripNet	**60.88**	89.42	**84.61**	**67.57**	**55.82**

The second part of the experiment used 80% of the UC Merced Land Use data as the training set and 20% of the UC Merced Land Use data as the test set. The experimental results are shown in Table 3.

Table 3. UC Merced Land Use test retrieval results of ablation experiments

Algorithm	mAP(%)	P@5(%)	P@10(%)	P@50(%)
TripNet [1]	44.70	51.48	51.48	20.09
M-TripNet	48.05	56.76	56.26	22.20
HEF-TripNet	46.60	53.67	52.86	20.67
MHEF-TripNet	**49.31**	**57.52**	**56.31**	**22.21**

The results of three modified algorithms are analyzed as follows:

1. M-TripNet: compared with TripNet, the mAP of M-TripNet is about 3% higher, and the accuracy of the first 5 and 10 are about 5% higher. The result indicates that M-TripNet performs better while training, and the additional label information in M-Triplet help to extract more robust features from remote sensing images.
2. HEF-TripNet: the mAP of HEF-TripNet is about 2% higher than that of TripNet. The P@5 and P@10 are about 2% and 1% higher than TripNet, respectively. It shows that the feedback plays a role in guiding feature extraction network to train hard examples by enlarge the distance of negative samples which are difficult to distinguish, hence improving the distinction of features.
3. MHEF-TripNet: compared to TripNet, the mAP of MHEF-TripNet is about 4.6% higher. P@5 and P@10 are about 6% and 5% higher, respectively. This means that MHEF-TripNet effectively integrates the above two modifications.

4.2 Retrieval Results on UC Merced Land Use

Due to the small number of images in UC Merced Land Use, where only 100 images of each class are available, we cannot measure P@100 in ablation experiments on UC

Merced Land Use dataset. Therefore, ablation experiments on Kdelab Airplane dataset are added. The Kdelab Airplane dataset is created by the Kdelab Laboratory of the University of Science and Technology of China. The dataset contains 2,200 images covering 11 different aircraft types, and 200 images of size 128 * 128 in each class. Figure 4 shows sample images of the dataset. Kdelab Airplane dataset is used to ablation experiments.

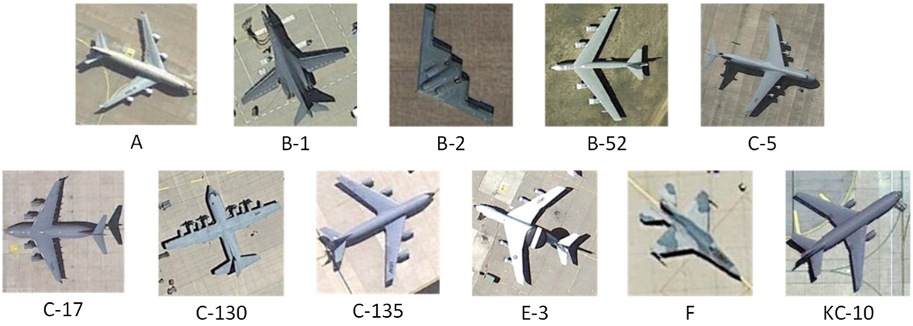

Fig. 4. Sample images of Kdelab Airplane dataset

In this experiment, 80% of the Kdelab Airplane dataset are used as training set and 20% as test set. Since there are 160 training images in each class of the dataset, the overall retrieval performs better than the UC Merced Land Use dataset. The result of retrieval is shown in Table 3 in detail. Comparing the mentioned three algorithms with TripNet, M-TripNet gets 3% higher in mAP, and about 1.5% higher in both P@5 and P@10. As for HEF-TripNet, the mAP is about 1% higher than that of TripNet, and the P@5 and P@10 are about 1.2% and 0.7% higher than TripNet, respectively. Regarding THEF-TripNet, it gets 4.4% higher in mAP, and about 2.2% higher in P@5 as well as in P@10. As can be seen, THEF-TripNet outperforms among other algorithms in all evaluation indicators, which suggests its effectiveness for feature extraction on Kdelab Airplane dataset (Table 4).

Table 4. Kdelab Airplane test retrieval results of ablation experiments

Algorithm	mAP(%)	P@5(%)	P@10(%)	P@50(%)	P@100(%)
TripNet [1]	76.41	79.18	89.00	76.81	74.17
M-TripNet	79.60	80.59	80.48	79.28	78.49
HEF-TripNet	77.44	80.32	79.68	77.81	75.57
MHEF-TripNet	**80.86**	**81.50**	**81.18**	**80.51**	**79.74**

As a result, THEF-TripNet can effectively improve the performance of TripNet in image retrieval. For one thing, the mixed loss not only considers the triplet loss, but also takes the label information of the image into account, which elaborately introduces the label information into the training of the triple. In fact, according to the label information, the cluster center is utilized to make training easier. For another, the feedback-based triple is a process of continuous iteration and adjustment. After each round of training, there is an image retrieval validation. THEF-TripNet evaluates the generalization of the current model, and finds out the hard examples at present, which will be improved pertinently in the next round. Therefore, the model can learn the most discriminative information from the triple, so as to improve the feature availability. Differentiation improves the performance of target retrieval.

5 Conclusion

In this paper, we propose mixed triplet loss with hard example feedback network. The method extracts more discriminative features based on mixed triple loss, focus on the correlation information and category information of images. At the same time, it introduces sample selection probability matrix to select hard triplets according to probability matrix. After each iteration, it adjusts the probability matrix according to the test results of the model, and then improves the effect of difficult sample selection from a global perspective. The experimental results show that this method is superior to the traditional triple method and can effectively improve the accuracy of remote sensing image retrieval.

References

1. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
3. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
5. Smeulders, A.W.M., Worring, M., Santini, S., et al.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 1349–1380 (2000)
6. Rui, Y., Huang, T.S., Ortega, M., et al.: Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol. **8**(5), 644–655 (1998)
7. Liu, Y., Zhang, D., Lu, G., et al.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. **40**(1), 262–282 (2007)

8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
10. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820 (2015)
11. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2892–2900 (2015)
12. Taigman, Y., Yang, M., Ranzato, M.A., et al.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
13. Kumar, N.S., Arun, M., Dangi, M.K.: Remote sensing image retrieval using object-based, semantic classifier techniques. Int. J. Inf. Commun. Technol. **13**(1), 68–82 (2018)
14. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 270–279. ACM (2010)
15. Napoletano, P.: Visual descriptors for content-based retrieval of remote-sensing images. Int. J. Remote Sens. **39**(5), 1343–1376 (2018)
16. Xia, G.S., IEEE, et al.: AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosc. Remote Sens. **55**(7), 3965–3981 (2017)
17. Yan, Y., Wang, X., Yang, X., Bai, X., Liu, W.: Joint classification loss and histogram loss for sketch-based image retrieval. In: ICIG (2017)