An Adaptive Weight Joint Loss Optimization for Dog Face Recognition

Qiwang Wang[1], Jiwei Song[2], Le Chang[3], Qing Tian[4], and Zhaofeng He[1(✉)]

[1] Beijing University of Posts and Telecommunications, Beijing, China
zhaofenghe@bupt.edu.cn
[2] China Electronics Standardization Institute, Beijing, China
[3] SoundAI Technology Co., Ltd, Beijing, China
[4] North China University of Technology, Beijing, China

Abstract. In recent years, the field of human face recognition has developed rapidly, and a large number of deep learning methods have proven their efficiency in human face recognition. However, these methods do not work well in the field of animal face recognition. There are two reasons. One is that the face recognition framework cannot fully extract the features of animal faces, and the other is that there are not enough animal datasets to fully train the model. In this paper, we collect a total of 11889 high-definition pictures containing 174 dog individuals, with an average of more than 60 samples per dog. On this dataset, we trained Swin Transformer as the backbone, and coupled with the triplet loss and cross-entropy loss function, it reaches an accuracy of 88.94%. Compared with the accuracy rate of 86.21% for using the TripletLoss alone and the accuracy rate of 86.60% for using the cross-entropy loss alone, this paper has a big improvement.

Keywords: Dog face recognition · Deep learning · Triplet-Loss · Joint loss function

1 Introduction

Now more and more people have pets and regard pets as their important friends. The increase in the number of pets has brought about a huge demand for pet personal identification technology in the pet insurance and pet medical industries. In animal husbandry, obtaining long-term health data of individual livestock is one of the important measures to know its health status and thus ensure animal health. Today, the main technologies for dog identification are tattoos, collars, and microchip implants. [1] But these methods can cause irreversible damage to animals and are often not reliable. Therefore, it is very important to develop a reliable and convenient method to identify individual animals. There are already many good algorithms in the field of face recognition, such as Facenet [2], DeepFace [3], DeepID [4], DeepID2 [5], and so on. However, there is no very effective method for animal individual recognition. The main reasons include two aspects:

W. Deng et al. (Eds.): CCBR 2022, LNCS 13628, pp. 561–568, 2022.
https://doi.org/10.1007/978-3-031-20233-9_57

one is the lack of sufficient high-quality datasets, and the other is that the framework that works well in face recognition may not be suitable for animal feature extraction. There are two contributions in this paper.

(1) Due to the lack of a high-quality dog identity dataset, we cooperated with the SAICHONGHUI Company to make a high-quality dog individual dataset. After a period of collection, the current dataset includes 174 dogs, with a total of 11,889 high-definition images. It can be used for dog recognition, dog detection etc.
(2) We proposed a method that selected Swin Transformer [6] as the backbone and coupled with the Triplet Loss [2] and Cross-Entropy loss function, it reaches an accuracy of 88.94% on this dataset.

2 Related Work

Face recognition based on deep learning is an emerging artificial intelligence technology. Compared with traditional recognition methods, it has the characteristics of safety, ease of use, and no implantation. The recognition process is mainly divided into the following steps: firstly, the face of the target individual is photographed and videoed to obtain the facial image through the acquisition equipment such as the camera; then the facial feature vector is extracted from the relevant image by using the deep learning model; The extracted facial feature vector is compared with the feature vectors of other pictures or existing categories to determine the identity.

Face recognition technology has been successfully applied in many industries, and many basic theories and practical techniques have been perfected in the process of continuous development. And its related field: animal individual recognition has gradually become a new hotspot and a new direction of research in recent years. At present, there are relatively few research works in the field of animal individual recognition, and most of them focus on cattle recognition [7]. Recently, an article designed a recognition network based on joint loss for goats with high similarity [8], which achieved an accuracy rate of 93.0007% on a closed high-similarity goat dataset. But their work can not transfer well to dog data, because the differences between individual dogs are much greater than the differences between goats. Some small-scale features are very important in the goat dataset, but some small-scale features are relatively less important in the dog data. Mougeot et al. [9] proposed a dataset containing 48 different dog faces, and obtained 88% recognition accuracy on this dataset. But this dataset is too small for recent very deep networks and pre-training weights that are getting better and better as computing power increases. In addition, the resolution of the image is also very low, which is not conducive to the extraction of high semantic features such as nose lines. For animal recognition, a large and good quality dataset is very much needed.

The use of facial recognition technology to determine and identify individual animal identities can not only help breeding enterprises to achieve intelligent

management and control of individual livestock but also help the government to better manage the intelligent annual inspection of the pet industry. Therefore, the research on animal individual identification has very important practical significance and value.

3 Dataset Creation

All the face images of stray dogs used in this paper are from Beijing Pet Association and SAICHOGNHUI Company, which are taken by the staff, and the corresponding dog face images are obtained after sorting and preprocessing (Fig. 1).

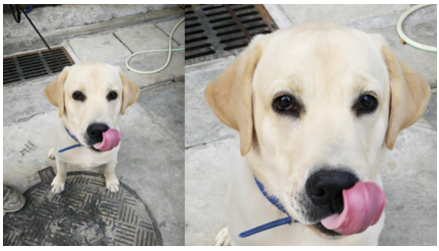

Fig. 1. The left picture is the original picture, the right picture is the cropped picture.

The core difficulty of the individual identification task lies in the scarcity of datasets. Usually, animal datasets are composed of multiple species, and the sources of the same species are scattered and disordered. Many individuals usually have only one sample, and most images are of low resolution, which makes the task of individual identification difficult. For the animal individual recognition task, to make the trained model have better generalization ability, it is necessary to simultaneously train the metric method to increase the distance between different classes and reduce the distance within the same class. At the same time, achieving good recognition results requires very high-quality datasets, which have a large number of different individuals of the same species, and each animal individual needs to have a sufficient number of clear frontal pictures.

In the process of collecting images, we found that it is difficult to obtain high-quality frontal images because it is very hard for animals to obediently look at the camera. For this reason, this paper proposes a shooting method of angle rotation through the field practice of taking pictures. After fixing the animal's head, rotate the angle to the right from the left 45-degree angle to the right 45-degree angle, and take at least 25 high-definition pictures continuously during this period. In the process of cooperating with SAICHONGHUI, a new batch of data is collected every half month. So far, the dataset has included 11,889 photos of more than 174 individual dogs (Fig. 2).

Fig. 2. Examples of our dataset.

4 Proposed Method

Most of the well-known feature vector extraction networks require massive training data to have better feature extraction results. In most cases, our realistic target tasks lack a sufficient number of high-quality datasets for training. Training pre-trained models with large datasets and then fine-tuning model parameters on small scale datasets has become a mainstream approach in both academia and industry. This transfer learning method can alleviate the consequences of insufficient data in the target task domain to a certain extent, that is, the minimum in the global mode cannot be found in the gradient forest, and only the local minimum can be found.

In the selection of the pre-training model, in addition to the size of the dataset, the distance between the source data domain and the target data domain is also an important factor to be considered. For example, if you need to train a model that recognizes dogs, training a pretrained model on a human dataset will generally outperform pre-training on a car dataset with similar dataset sizes. This shows that in the case of transfer learning tasks, it is better to use pre-training datasets with closer domain distances.

So we tried several backbones that perform well on Image-Net recognition tasks, such as ResNet50 [10] EfficientNet [11] Swin [6], etc. Compared with the face dataset, both the intra-class distance and the inter-class distance are larger, especially the intra-class distance will have a greater impact on the recognition results. This is because it is difficult to obtain frontal face images in the process of animal data collection, and there will be various angles or even occlusions, while this situation will be much less in the human face dataset. Therefore, we need to reduce the intra-class distance as much as possible, so we choose Triplet-loss as the loss function, but it will be difficult for the network to fitting, so we

use the combination of the cross-entropy loss function and Triplet-loss as our loss function, and add an adaptive weight to make the network fitting faster and have a higher accuracy.

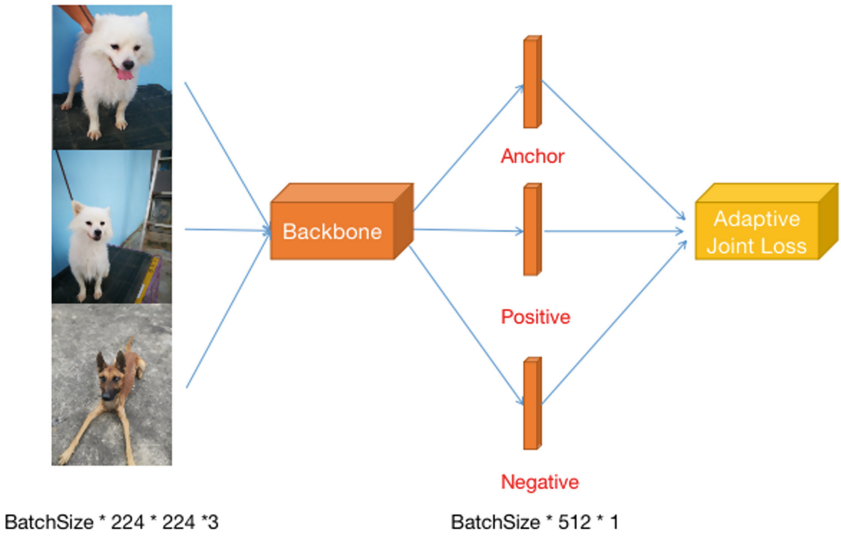

Fig. 3. The pipeline of our network

The overview of the proposed method is shown in Fig. 3. We first randomly select an anchor sample in a batch of data, then find the positive sample with the largest L2 distance of the same class as this anchor, then set a threshold T = 0.5, in this batch of data all normalized L2 distances are less than T is used as a set of negative samples. Then an adaptive loss is applied to the obtained features, because the initial Triplet-Loss is difficult to fit, so we will give Triplet-Loss a small value of weight α in the initial stage of the network, and cross-entropy loss a larger weight of $1 - \alpha$. At this stage, the cross-entropy loss dominates the descending direction of the network. Until the network has some recognition ability, the weights of these two losses are adaptively updated. In each batch, which loss drops faster will gain greater weight in the next batch. Weight update policy is shown in formula 1, $step_{value}$ is the step size for each update of the weight value, $judge_{bool}$ shows which loss from the last batch had a greater impact on the results, that is, loss dropped more.

$$\alpha = \alpha + step_{value} * judge_{bool} \tag{1}$$

$$loss = \alpha * TripletLoss + (1 - \alpha) * CEloss \tag{2}$$

5 Experiments and Results

Here are several experiments to prove the effectiveness of our method. The dataset and some experimental details are introduced first. Then we analyze the results and the comparison with other methods.

5.1 Dataset

Our dataset includes 11819 photos, a total of 174 dogs. On average, each dog has 67 pictures, of which 9447 samples are used for training and 2442 for testing. Both the training set and the test set include all 174 dogs. Before training, we performed data augmentation operations including clipping, random translation, random rotation, Gaussian noise, MotionBlur, Cutout operations. This improves the accuracy of the test set by 0.2%.

5.2 Implementation Details

We initially thought that since the original image contained the surrounding background, it would have some bad influence on the recognition results. So we performed object detection on the original data, detected the dog body according to the object detection framework FasterRCNN [12] pre-trained on ImageNet, and then segmented a new dog body dataset. We also manually marked the dog face data, then train a dog face detector, and performed dog face detection on the dataset, segmented a new dog face dataset and train it. The results show that the effects of these three datasets are similar. It may be because the network has been trained to have the ability to filter the ambient background noise.

So we use the original data to feed the backbone after data augmentation. Due to the use of Triplet-Loss, each batch of BatchSize is one third of the original size. This is because each time an anchor image is randomly selected, the corresponding positive image and negative image must be selected according

Table 1. Recognition performance of different method

Ablation study		
Model	Loss function	Recognition acc (%)
Resnet18 [10]	CE	85.94
Resnet50 [10]	CE	86.04
Swin Transformer [6]	CE	86.60
EfficientNet [11]	CE	86.72
Resnet18 [10]	CE+TripletLoss	86.51
Resnet50 [10]	CE+TripletLoss	86.49
Swin transformer [6]	CE+TripletLoss	**88.94**
EfficientNet [11]	CE+TripletLoss	87.88

to the anchor image. We take a uniform batch size of 64 * 3 with a learning rate of 3e−4. The initial joint loss weight parameter α is 0.01. When cross-entropy loss <9, the initial weight parameter α is 0.5, and the subsequent α is updated by Formula 1.

5.3 Ablation Study

We selected four models that performed well on the ImageNet recognition task as backbones and initialized them with pre-trained weights. We tested cross-entropy loss alone and the combination of cross-entropy loss and TripletLoss, because the network using TripletLoss alone will be difficult to fitting, so it is not listed in the Table 1.

The experimental results show that using the joint loss is about 1% higher than using CE as the loss, especially when the Swin Transformer is selected as the backbone, the improvement is most obvious, with an increase of 2.34%. This may be from the global attention mechanism of the Swin Transformer. Compared to other networks based on convolutional structures, Triplet Loss helps the network learn more inter-class differences in global features.

5.4 Comparison with Other Methods

We use the Swin Transformer as the backbone, with the adaptive joint loss function and get a pretty good results, getting accuracy of 88.94% on our dataset, which is a big advantage over other networks. Other ResNet18, Swin Transformer, and EfficientNet that only use CE as loss have slightly lower accuracy. For the two methods using Swin Transformer as the feature extraction framework, using the adaptive joint loss function can improve the accuracy by about 2%, which indicates that the joint loss function can indeed help the network learn richer features. Also, from the results, it can also help the network adapt to fewer training epochs. As shown in Fig. 4, in the case of the same epochs, our results are better than others.

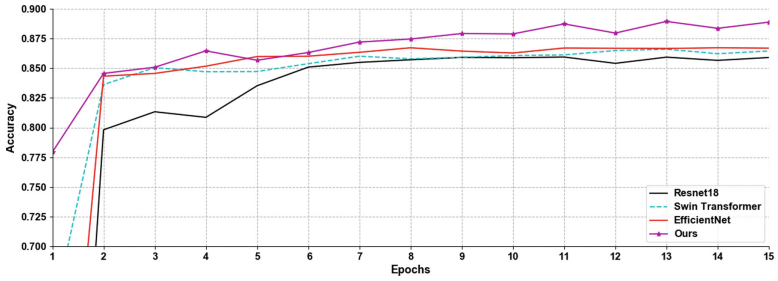

Fig. 4. Comparison with other methods

6 Conclusion

After recent development, face recognition technology has achieved very high accuracy and stability thanks to the rich face datasets and the work input of a large number of researchers. But in terms of animal identification, there has not been much breakthrough progress. Dog face recognition is much more complicated than human face recognition due to the lack of available data and the wide range of texture variations in dog face images. In this paper, we design a new method to solve the dog recognition problem by establishing a new dog dataset containing multiple dogs, each with a large number of high-definition samples, and achieve satisfactory accuracy. However, there is still room for further improvement in the recognition accuracy. In the future, it may be possible to study the fusion of dog nose texture information to further improve the recognition accuracy.

References

1. Blancou, J.: A history of the traceability of animals and animal products. Revue scientifique et technique (International Office of Epizootics) (2001)
2. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. CoRR abs/1503.03832 (2015)
3. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification (2014)
4. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
5. Sun, Y., Chen, Y., Wang, X., et al.: Deep learning face representation by joint identification-verification. Adv. Neural Inf. Process. Syst. **27** (2014)
6. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
7. Yao, L., Hu, Z., Liu, C., et al.: Cow face detection and recognition based on automatic feature extraction algorithm. In: Proceedings of the ACM Turing Celebration Conference-China, pp. 1–5 (2019)
8. Shang, C., Wang, M.L., Ning, J.F., Li, Q.H., Jiang, Y., Wang, X.L.: Joint loss optimization based high similarity identification for Milch goats. J. Image Graph. **27**(04), 1137–1147 (2022)
9. Mougeot, G., Li, D., Jia, S.: A deep learning approach for dog face verification and recognition. In: Nayak, A.C., Sharma, A. (eds.) PRICAI 2019. LNCS (LNAI), vol. 11672, pp. 418–430. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29894-4_34
10. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
12. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28** (2015)