Multi-lingual Speaker Recognition based on Asymmetric Convolution and Central Loss Function

Xiaofang Zhang¹, Askar HAMDULLA² and Mijit ABLIMIT²,*

¹School of Information Science and Engineering Xinjiang University, Urumqi, China
²School of Information Science and Engineering Xinjiang University, Urumqi, China
*Corresponding author e-mail: askar@xju.edu.cn

Abstract. In order to solve the problem of poor speaker recognition performance on multi-language corpus on convolutional neural networks and large amount of calculations on multiple parameters, this paper draws on asymmetric convolution and center loss (Center Loss) functions to improve and optimize the ResNet model. And perform speaker recognition tasks in the Chinese-Uyghur corpus. The results show that, compared with the original model, the improved model has higher accuracy, fewer parameters, and reduced computational complexity.

1. Introduction

Two typical speaker recognition methods that are still widely used are: 1. X-vector/PLDA speaker recognition system; 2. Neural network-based end-to-end speaker recognition system. Among them, the convolutional neural network (CNN) is a common neural network algorithm. The weight sharing of CNN and the partial connection of the convolutional layer can reduce the parameters to slow down the gradient disappearance and model degradation. This paper proposes a method based on the improved convolutional network results. The speaker recognition system uses the idea of asymmetric convolution in Inception V3 to improve the convolution kernels in the dense blocks and residual blocks of ResNet respectively, and adds the center loss (Center loss) to the fully connected layer. Loss) function is used for speaker recognition task, which reduces the parameters and calculation amount of the model, and introduces the central loss function to pull the features in the same category to the center point as much as possible, so that the distance between the features becomes larger, and further improve the accuracy of recognition.

2. Model Introduction

2.1. Asymmetric Convolution

The idea of asymmetric convolution is to replace the n×n convolution with n×1 convolution and 1×n convolution[1]. The original n×n multiplication has become a 2×n multiplication, the bigger n is, the less computation you have to do. By applying the idea of asymmetry, the 3×3 convolution kernel in the RESNET model is replaced by 3×1 and 1×3 convolution kernel, so that the architecture of the RESNET model remains unchanged and the amount of calculation is reduced [2][3].
As shown in Figure 1, a 3×3 convolution is replaced with a 3×1 convolution and then a 1×3 convolution in series. Each 1×1 area of feature map B feels each of feature map A 3×3 area, and each 1×3 area of feature map C feels each 3×3 area of feature map A, and each 1×1 interval of feature map D feels each 1×3 area of feature map C. Each 1×1 area equivalent to feature map D feels every 3×3 area of feature map A.

2.2. Center Loss Function

The classification layer uses Softmax only requires that the features can be correctly classified by the classification layer, and there is no requirement for the features. When some features are difficult to distinguish or are at the boundary of classification, it will affect the accuracy of recognition. Therefore, the center loss function (Center Loss) is introduced. It pulls the features in the same category to the center point as much as possible, making the distance between the features larger [4][5]. The total loss function of the two loss function training of Softmax and Center Loss formula can be expressed by formula 1:

$$L_{sc} = \text{loss}_{\text{softmax}} + \lambda \text{loss}_{\text{center}}$$

$$= -\sum_{i=1}^{m} \log \frac{e^{W^T x_i + b_i}}{\sum_{j=1}^{n} e^{W^T x_j + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \| x_i - c_{yi} \|_2^2$$

(1)

The specific parameters of the improved ResNet network model are shown in Table 1.
### Table 1. Improve the specific parameters of the ResNet network model

| Number | Network layer | Structure | Number | Network layer | Structure |
|--------|---------------|-----------|--------|---------------|-----------|
| 1      | Conv          | 7×7       | 7      | ResBlock3     | \[1×l_{conv}\] × 24 |
|        |               |           |        |               | \[1×3_{conv}\]       |
|        |               |           |        |               | \[3×1_{conv}\]       |
|        |               |           |        |               | \[1×1_{conv}\]       |
| 2      | Pool          | 3×3 max pool | 8      | 3             | BN+Relu |
|        |               |           |        | 9             | ResBlock4 |
|        |               |           |        |               | \[1×3_{conv}\] × 16 |
|        |               |           |        |               | \[3×1_{conv}\]       |
|        |               |           |        |               | \[1×1_{conv}\]       |
| 3      | ResBlock1     | \[1×1_{conv}\] × 6 | 10     | pool          | 7×7 max pool |
|        |               | \[1×3_{conv}\]       |        |               |           |
|        |               | \[3×1_{conv}\]       |        |               |           |
|        |               | \[1×1_{conv}\]       |        |               |           |
| 4      | 1             | BN+Relu |
|        |               |           | 11     | Classification layer | Softmax+Center Loss |
| 5      | ResBlock2     | \[1×1_{conv}\] × 12 |        |               |           |
|        |               | \[1×3_{conv}\]       |        |               |           |
|        |               | \[3×1_{conv}\]       |        |               |           |
|        |               | \[1×1_{conv}\]       |        |               |           |

### 3. Experiment and Analysis

#### 3.1. Corpus and Experimental Design

The Free ST Chinese Mandarin Corpus Chinese corpus and the THUYG-20 Uyghur corpus are mixed to generate a Chinese-Uyghur corpus. The total number of speakers is 1203, the total sentence is 20,000 sentences, the male to female ratio is 1:1, and the length of the audio is not fixed, mostly distributed in 3-4 seconds, and there are more audios above 10s.

The baseline experiment is to carry out the X-vector speaker recognition system on the Kaldi platform[6]. The learning rate of the model is 0.001, batch_size is 32, and the optimization method is Adam's algorithm. In order to prevent overfitting, this experiment uses Dropout with a ratio of 0.5, and Activity Regularization selects L2 regularization.

#### 3.2. Analysis of Results

### Table 2. Comparison of experimental results on ResNet models before and after improvements

| Corpus               | Chinese-Uyghur Corpus | Parameters |
|----------------------|------------------------|------------|
| X-vector             | 0.8952                 | -          |
| ResNet Acc           | 0.9164                 | 23,677,370 |
| Improve ResNet Acc   | 0.9313                 | 9,367,835  |
| Improve Acc          | **0.0149**             | -          |

From the experimental results Table 2, it can be seen that the accuracy of the improved ResNet model is about 1% higher than that of the original network model, and the parameters of the improved model are significantly reduced.
ResNet model are reduced by about six times compared with those before the improvement, which greatly reduces the computing power and shortens the training time of the model. Time, shorten the training time of the model. It shows that the improved ResNet model has certain advantages in recognition, and the fit of the model has been improved, which shows that the improvement and optimization of the model are effective.

The loss value of the ResNet network on the two corpora changes before and after the improvement. As shown in Figure 2, the loss of the improved ResNet network on any corpus decreases and converges faster, which shows that the improved optimization scheme of the model is effective.

![Figure 2. Model loss on the Chinese-Uyghur corpus](image)

4. Conclusion
The improved ResNet network model converges faster, reduces the number of parameters and computation, improves the recognition accuracy and has better performance, which shows that the improved and optimized method is effective. It solves the problem of poor speaker recognition performance and large amount of multi-parameter calculation in convolutional neural network multilingual corpus.

References
[1] Ding X, Y Guo, Ding G, et al. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. ICCV[J]. 2019.
[2] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. IEEE, 2016:2818-2826.
[3] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.
[4] Qi C, Fei S. Contrastive-center loss for deep neural networks[C]// 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017..
[5] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[J]. Springer, Cham, 2016.
[6] Kanagasundaram A, Sridharan S, Ganapathy S, et al. A Study on Pairwise LDA for X-vector based Speaker Recognition[J]. Electronics Letters, 2019.