IBG_AI Language Identification System for AP20-OLR

Feifei Zhou,Chuan Ke, Mingqing Zhu, Yi Peng, Zhihao Liu

International Business Group, Tencent Inc., Shenzhen, China
{feifeizhou,jayke,migosszhu,yipeng,zhihaoliu}@tencent.com

Abstract

In this report, we describe the submissions of International Business Group AI (IBG_AI) team to AP20-OLR. Our submitted systems for task 1 and task2 are both a fusion of five models based on ResNet with Squeeze-Excitation. One model makes use of MFCC features, while the other four models use FBank features. Language identification is simply regarded as classify problem, and trained by Softmax cross entropy in our systems. As a result, our best systems for task1 and task2 achieved 0.0079 C_{avg} and 0.025 C_{avg} on our own development sets respectively.

Index Terms: language identification, deep neural network

1. Introduction

Language identification refers to identifying language categories from utterances. The AP20-OLR challenge[1] includes three tasks: task 1 focus on cross-channel language identification (LID), task 2 is an open-set identification task, and task 3 pays more attention on noise.

We submitted the final results of the task 1 and task 2 with required test conditions in this challenge. We regard the task as a classify problem, and the Fbank features or MFCC features of the audio are extract. The rest of this document is organized as follows: in Section 2, we describe the details of each part of our system, in Section 3, several experiments are conducted to get the result, and the final conclusion are presented in Section 4.

2. System components description

In this section, we introduce all the components used in our systems.

2.1. Front-End

2.1.1. Training data, Augmentations

According to the evaluation plan of AP20-OLR, the following datasets are allowed to construct the system: AP16-OL7, AP17-OL3, AP17-OLR-test, AP18-OLR-test, AP19-OLR-dev, AP19-OLR-test, AP20-OLR-dialect, THCHS30. For task 1, there are 6 languages, i.e., Cantonese, Indonesian, Japanese, Korean, Russian, Vietnamese, so we picked out the audios which are in the 6 languages as our dataset. For task 2, there are 3 languages, i.e., Minnan, Shanghai, Sichuan, besides, there also exists an unknown language. We select the three language data from the allowed datasets, and other data are regarded as unknown language. So for task 2, it is a 4-class model.

The selected data are split into training set and development set in a ratio of 9: 1. Although there is a gap between the development data and evaluation data, it is still helpful for parameters tuning and model selection.

As argument data is forbidden in this challenge, so in order to improve the generalization of the model, we crop the original audio into pieces with different lengths.

2.1.2. Features and VAD

Our models make use of FBank features and MFCC features. The 90-dimensional FBank features are extracted in the way similar to BUT system[2]. It is extracted from audios which is down-sampled to 16kHz, its frequency is limited to 20-7600Hz, and the frame length is 25ms with 10ms shift. As audios in different dataset have various volume levels, the FBank feature is normalized by subtracting the mean and divided by the standard deviation with a sliding window of 3 seconds. As for MFCC features, it has 60 dimensions. And we didn't use any VAD for our system.

2.2. Model

2.2.1. Model structure

The backbone network of our system is the well-known ResNet topology[3]. Similar to BUT system, we halved the number of channels of each ResNet block, thus can get a deeper network with the similar number of model parameters. The details of the ResNet101 topology is shown in Table 1.

In addition, attentive statistics pooling layer[4] is utilized to aggregate frame-level representation on utterance level, benefit from the attentional mechanism, the model is able to assign bigger weights to more important frames when calculate statistics, which makes the model more robust to the external environment interference and silence frames.

In our models, the shape of the last ResNet block output is (d_1, T_1, c), c is the number of channels, d_1 and T_1 represent the number of feature dimensions and frames after they pass through ResNet blocks respectively. In order to reduce the attention network parameters, the output is averaged along feature dimension, which leads to a matrix H of shape (T_1, c), and then additive attention with two fully-connected layers is utilized to get the weights of every frames e_t:

$$e_t = \mathrm{softmax}(f(f(h_t))) \tag{1}$$

where $f(.)$ is fully connected layer with nonlinear activation function, h_t is each row of H. Finally, the weighted mean and standard deviation of each dimension is concatenate together as the input of the rest network.

Besides, as the model benefits from a wider temporal context, it could be beneficial to rescale the frame-level features given global properties of the recording. For this purpose, 2-dimensional Squeeze-Excitation (SE) blocks is introduced to the get the weights of each channel[5] in the last two ResNet block as shown in Table 1. The first component of an SE-block is the freeze operation which calculates the mean of each channel:

$$\boldsymbol{z} = \frac{1}{T_0 d_0} \sum_{t=1}^{T_0} \sum_{i=1}^{d_0} \boldsymbol{u}_{t,i} \tag{2}$$

where $\boldsymbol{u}_{t,i}$ is the vector consists of every channel data at time t and feature I, d_0 and T_0 represent the number of fea-

Table 1: *The proposed ResNet101 architecture. C in the last row is the number of language categories. S in Structure column indicates the stride of convolution. The first dimension of the input shows number of filter-banks and the second dimension indicates the number of frames*

Layer name	**Structure**	**Output**
Input	-	$(90, L, 1)$
Conv2D-1	3×3,S=1	$(90, L, 32)$
ResNetBlock1	$\begin{bmatrix} 1 \times 1, 32 \\ 3 \times 3, 32 \\ 1 \times 1, 32 \end{bmatrix} \times 3$, S=1	$(90, \frac{L}{2}, 128)$
ResNetBlock2	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 32 \end{bmatrix} \times 4$, S=2	$(45, \frac{L}{4}, 256)$
ResNetBlock3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 32 \\ \text{SE} - \text{Block} \end{bmatrix} \times 23$, S=2	$(23, \frac{L}{6}, 512)$
ResNetBlock4	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 32 \\ \text{SE} - \text{Block} \end{bmatrix} \times 3$, S=2	$(12, \frac{L}{8}, 1024)$
Att-StatPooling	-	$(24, 1024)$
Flatten	-	24576
Dense1-Relu-BN	-	64
Dense2	-	C

ture dimensions and frames after passing through the convolution network in ResNet block respectively. Then two-layer fully connected layer is used to get the weights of each channel:

$$\boldsymbol{s} = \sigma(f(f(\boldsymbol{z}))) \tag{3}$$

where σ denotes the sigmoid function, and the resulting vector $\boldsymbol{s}$ contains weights s_c between 0 and 1, which are applied to the original input by channel-wise multiplication.

2.3. Fusion

For score level, fusion was performed by computing the average of the scores of the individual systems.

3. Results and Discussion

At every training step, for each language, 16 segments which are randomly cropped between 200 frames and 250 frames are selected, if the length of a audio is smaller than 200 frames, we repeat it util reach the minimum length, i.e. 200 frames. Softmax with cross entropy is used to train the model. We select stochastic gradient descent (SGD) as the optimizer and the weight decay is set to 0.0002 in Pytorch. The initial learning rate to train the raw model is set to 0.1. During training, learning rate halved every 1400 steps and the model is trained for 14000 steps.

The results of the models for task 1 are displayed in Table 2, the fusion which makes use of Model 1 to Model 5 is our best submission, and achieve 0.0079 C_{avg} on the development set.

Table 2: *Results of our systems for AP20-OLR task 1 challenge. The 60-dims prefix indicates 60-dimensional MFCC features, SE indicates Squeeze-Excitation Block*

ID	Embd NN	C_avg on Development Set
1	60dims-ResNet101-SE	0.0179
2	ResNet18-SE	**0.0113**
3	ResNet34-SE	0.0152
4	ResNet50-SE	0.0122
5	ResNet101-SE	0.0125
Fusion 1-5		0.0079

Table 3: *Results of our systems for AP20-OLR task 2 challenge. The 60-dims prefix indicates 60-dimensional MFCC features, SE indicates Squeeze-Excitation Block*

ID	Embd NN	C_avg on Development Set
1	60dims-ResNet101-SE	0.098
2	ResNet18-SE	**0.049**
3	ResNet34-SE	0.054
4	ResNet50-SE	0.06
5	ResNet101-SE	0.056
Fusion 1-5		0.025

The results of the models for task 2 are displayed in Table 3, and the fusion model get 0.025 C_{avg} on development set.

4. Conclusions

In this work, we presented our language identification system for AP20-OLR task1 and task 2. And our fusion submission achieved 0.0079 C_{avg} and 0.025 C_{avg} for task 1 and task 2 on our own development set respectively.

5. References

[1] Z. Li, M. Zhao, Q. Hong, L. Li, and C. Yang, "Ap20-olr challenge: Three tasks and their baselines," 2020.

[2] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Proc. Interspeech 2018*, pp. 2252–2256, 2018.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.