Focal Loss for End-to-end Short Utterances Chinese Dialect Identification

Qiuxian Zhang*, Jiangyan Yi†, Jianhua Tao†‡, Mingliang Gu* and Yong Ma*§
* School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou, China
† National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
‡ School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
§ Jiangsu Normal University Kewen College, Xuzhou, China
E-mail: zhang_qiuxian@126.com, jiangyan.yi@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, mlgu@jsnu.edu.cn, may@jsnu.edu.cn

***Abstract*—Short utterances dialect identification is a challenging task because of the substantial similarity between dialects. The previous cross-entropy loss function does not consider the category and probability of prediction error, which result in insensitivity to easily misclassified and unbalanced samples. To solve this problem, we propose to use an improved cross-entropy loss function, namely focal loss, introducing category weights and tunable focusing parameter to improve the classification accuracy. Experiments are carried out on AI Dialect Contest database. The results demonstrate that our proposed end-to-end model trained with focal loss achieves better performance than the model trained with cross-entropy loss function.**

I. INTRODUCTION

Dialect identification (DID) is a special case of language identification (LID) [1], but due to the similarity between dialects, DID is considered to be a more challenging problem than general LID. In particular, duration mismatch between training and test utterances is a long-standing problem in LID [2, 3]. Commonly, long utterances are available for model training but test utterances may be very short during the language recognition stage, thus reducing the accuracy of recognition. Similarly, improving the performance of DID on short utterances is also one of the important tasks for practical applications.

In recent decades, a number of techniques have been applied successfully on language identification or dialect identification [4]. For LID/DID, i-vector is regarded as the state-of-the-art latent feature extraction method, which can obtain fixed-dimensional representations of utterances by the total variability i-vector modeling. However, the performance is degradation when dealing with short utterances, the main reason is that there is a large distribution variation in the i-vector representation of short utterances [5].

Recently, end-to-end approaches have been explored with deep neural networks (DNN), recurrent neural networks (RNN), convolutional neural networks (CNN) for LID [6, 7, 8, 9, 10], which can not only cross the framework level to the utterance level LID identity by extracting the discriminative feature representation of the species speech, but also avoid the need for the back-end discriminant algorithm [11]. For short utterances LID tasks, Gonzalez et al. [12] and Geng et al. [7] proposed building long short term memory-recurrent neural networks (LSTM-RNN) to identify languages. Especially, Geng et al. [7] used a unified attention based-LSTM to establish an utterance level end-to-end short duration LID system. However, unidirectional LSTMs cannot extract context information concerning future frames. Fernando et al. [13] showed that bidirectional long short term memory network (BiLSTM) performs well for short durations (3 seconds) LID tasks by modelling temporal dependencies between past and future frame based features in short utterances. Shen et al. [14, 15] used deep convolutional neural networks (DCNN) and interactive learning of teachers-student model for short test durations (segments up to 3 seconds of speech). These approaches have been shown to be effective for short utterances language identification.

For multi-classification tasks, cross-entropy loss function (CE) [16] CE simply calculates the distance between the actual category label distribution and the predicted category label distribution, but does not take into account the problems of the "more easily errored" and unbalanced samples. In 2017, Kaiming He et al. [17] proposed the focal loss (FL) for dense object detection. Experiments show that the focal loss can effectively address the problem and improve the performance of the model. Zhao et al. [18] used the focal loss as one of the three strategies for optimizing in large-scale Mandarin Chinese speech recognition tasks during the training process, which is a dynamically scaled cross- entropy loss and down-weights the loss assigned to well-classified examples. In addition, focal loss has been applied in natural language processing [19].

Inspired by the above methods [17, 18, 19], we adopt the focal loss for end-to-end short utterances Chinese dialect identification. In this paper, we focus on two goals to further improve the performance of DID. The first one is to reduce the weight of easily classified samples by the focal loss, so that the model is more focused on difficultly classified samples during training. The second one is to consider the imbalance of dialect dataset. Experimental results on AI Dialect Contest dataset show that our proposed method based convolutional neural networks-bidirectional gated recurrent unit (CNN-BiGRU) is effective in improving the accuracy of short utterances.

The rest of this paper is organized as follows. We review of cross-entropy loss function in Section II and introduce focal loss in Section III. Section IV is to describe the proposed method that focal loss using CNN-BiGRU for short utterances DID. Section V presents our experiments and results. Section VI discusses the experimental results in detail. Finally, conclusions are summarized in Section VII.

II. CROSS-ENTROPY LOSS FUNCTION

For end-to-end language or dialect identification tasks, cross-entropy loss function (CE) is usually used and minimize it for optimization. CE is to calculate the distance between the real category label distribution and the predicted category label distribution, which is defined as shown in equation (1), the representation of the category label distribution of the real category label distribution in the model prediction is shown in equation (2):

$$J(\theta) = -\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\log\left(\hat{y}_{ic},\theta\right) \tag{1}$$

$$L_{CE} = -\sum_{c=1}^{M} y_c\log\left(\hat{y}_c,\theta\right) \tag{2}$$

where N is the number of speech samples, M is the number of classes in the multi-classification, $y_c \in \{0,1\}$ is the pre-marked true value, $\hat{y}_c$ is the predicted probability value of the observed sample belonging to the class c, and θ is the representation of all the parameters in the model.

III. FOCAL LOSS

In the work, due to the confusing and unbalanced dialects, we consider a variant of the cross-entropy loss, namely the focal loss (FL). It is first proposed in dense object detection. The essence of focal loss is to use a suitable function to measure the contribution of easily/difficultly classified samples to the total loss function.

For the class imbalance problem, a weighting factor α is introduced to reduce the impact of the category with more data on cross-entropy loss function:

$$L_{CE} = -\alpha\sum_{c=1}^{M} y_c\log\left(\hat{y}_c,\theta\right) \tag{3}$$

While α balances different categories samples, it does not differentiate between easily/difficultly classified samples, such as confusing dialects. Therefore, β is introduced to down-weight easily classified samples and focus training on difficultly classified samples. In each iterative process, it is possible to adaptively identify whether the sample is easily or difficultly classified. It allows the model to learn as much as possible the difficultly classified samples, β is defined as follows:

$$\beta = (1-p_c)^{\gamma} \tag{4}$$

$$p_c = y_c p\left(\hat{y}_c,\theta\right) + (1-y_c)\left(1-p\left(\hat{y}_c,\theta\right)\right) \tag{5}$$

where β is the modulating factor with tunable focusing parameter $\gamma \geqslant 0$, the focal loss is defined as follows:

$$L_{FL} = -(1-p_c)\log(p_c) \tag{6}$$

When the samples are correctly classified, p_c is tend to 1 and β is tend to 0, the influence of the correctly classified samples on the gradient of the loss function is close to 0. Conversely, when the samples are misclassified, it approximates to the original loss function value.

Here we use an α-balanced variant of the focal loss:

$$L_{FL} = -\alpha(1-p_c)^{\gamma}\log(p_c) \tag{7}$$

where $\alpha\epsilon[0,\ 1]$, $\gamma\epsilon[0,\ 5]$.

According to our investigation, this is the first time that FL is introduced into the DID task. Our experimental results demonstrate that the focal loss can improve the performance of DID.

IV. FOCAL LOSS FOR SHORT UTTERANCES DID

A. System overview

In this paper, we combine the focal loss and batch normalization based our previous work [20] that proposed end-to-end method using convolutional neural networks-bidirectional gated recurrent unit (CNN-BiGRU). Fig. 1 shows our overall architecture.

Firstly, the raw dialect audios are preprocessed and acoustic features are extracted (such as 40-dimensional filter bank features, Fbank). Secondly, CNN-BiGRU model plays a role as local pattern extractor for the inputs. It effectively extracts the deep information features representation of short utterances, and uses the multi-layer nonlinear mapping of the network model to obtain a more differentiated high-level expression of short utterances. In addition, the batch normalization layer is used to CNN and BiGRU respectively to reduce the weight offset and accelerate the convergence of the network during training. Then, connecting the fully connected layer after CNN-BiGRU and making the global average of the frame level output. Finally, the output vector is mapped to the (0, 1) interval by the softmax function, and the probability between the dialect categories is calculated to predict the class label of the Chinese dialect. We apply the focal loss (FL) instead of the traditional cross entropy loss function (CE) to optimize the model parameters during training.

B. CNN-BiGRU Model

For this DID task, we extract 40-dimensional Fbank features as inputs of the model, as shown in Fig. 2. Generally, the network is divided into two parts.

The first part is CNN, it includes three convolution layer blocks. Each block includes one convolution layer, one batch normalization layer and one ReLU layer. The convolution layer can overcome the diversity of speech signals by using the invariance of convolution, and extract local features by

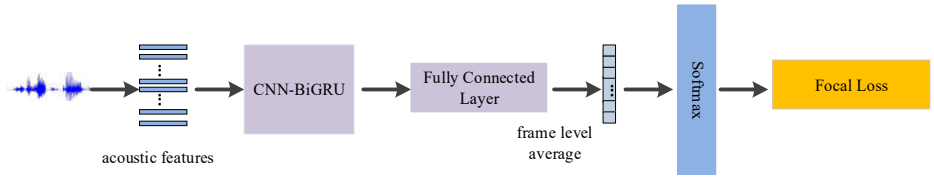

Fig. 1. Architecture of our proposed focal loss-based CNN-BiGRU for short utterances DID

TABLE I
THE TRAINING DATA USED IN EXPERIMENTS

Dialects	#Speakers	#Utterances	#Hours
changsha	30	6000	8.37
hebei	30	6000	7.04
hefei	30	6000	8.40
kejia	30	6000	6.91
minnan	30	6000	7.27
nanchang	30	6000	8.38
ningxia	30	6000	6.92
shan3xi	30	6000	7.74
shanghai	30	6000	7.19
sichuan	30	6000	7.21

TABLE II
SHORT UTTERANCES TEST DATA ($\leqslant 3s$) USED IN EXPERIMENTS

Dialects	#Speakers	#Utterances	#Hours
changsha	5	250	0.23
hebei	5	250	0.24
hefei	5	250	0.25
kejia	5	250	0.22
minnan	5	250	0.19
nanchang	5	250	0.26
ningxia	5	250	0.17
shan3xi	5	250	0.34
shanghai	5	250	0.21
sichuan	5	250	0.21

using the invariance of convolution, and extract local features by sliding convolution kernel. The batch normalization layer normalizes the output of the convolution layer to effectively control overfitting.

The second part is BiGRU, it includes the forward GRU and the reverse GRU. Considering insufficient information and time series correlation of short utterances, BiGRU extracts the global context feature information vector. That is, the forward implicit state and the reverse implicit state are combined to form a hidden state of the speech.

V. EXPERIMENTS

A. Dataset

Experimental dataset is obtained through AI Dialect Contest [21] organized by IFLYTEK. It includes speech from 10 different dialects, namely changsha, hebei, hefei, kejia, minnan, nanchang, ningxia, shan3xi, shanghai and sichuan.

The data is stored in a PCM format with a sampling rate of 16000 Hz and 16 bits of quantization. The recording environment includes a quiet environment and a noisy environment. There are 6000 utterances in each dialect on the training data, including 30 speakers. In all the experiments, we select short utterances (segments up to 3 seconds of speech) of the test data. The training data and short utterances test data are shown in Table I and Table II in detail.

B. Baseline models

I-vector: I-vector-based method was examined for comparison. The script of Kaldi toolkit [22] was used for the i-vector system preparation. We extracted 40-dimensional Fbank features from raw audio, and a frame-level energy-based voice activity detection (VAD) selects features corresponding to speech frames. Then, the GMM-UBM with 512 mixtures was trained, along with a 400 dimensional i-vector extractor. After extracting i-vectors, a whitening transformation and length normalization are applied. Linear Discriminant Analysis (LDA) was applied after extracting i-vectors with language labels, LDA reduced the 400 dimensional i-vector to 9 dimensional. Finally, a support vector machine (SVM) classifier is trained on these i-vector to use for dialect identification.

GRU/BiGRU-CE: For better result comparison, we use end-to-end methods to build system with GRU for DID task. Audio is converted to 40-dimensional Fbank features with a frame length of 25ms and a frame shift of 10ms. The model includes two layers GRU with 256 units in each layer followed by an output softmax layer as a baseline end to end system. The dropout layer is added between the GRU levels and the random deactivation value was 0.3. The batch size is 64. We use stochastic gradient descent (SGD) with learning rate 0.1 and momentum 0.9. CE is adopted for model optimization. In addition, we also used BiGRU with 128 units in each layer instead of GRU.

CNN-BiGRU-CE: As mentioned in Fig. 1, in the CNN, there are 100, 128, 128 convolution kernels with a size of 7 and the stride of 1. We adopt the BN layers with momentum 0.9 and decay_rate 1e-05. In the BiGRU, there are two layer BiGRU with 128 units and the BN layer with the same parameter setting as above. Table III shows performance com-

TABLE III
THE PERFORMANCE COMPARISON OF DIFFERENT MODELS BASED ON CE IN SHORT UTTERANCES

Model	Acc (%)
i-vector	70.20
GRU - CE	72.17
BiGRU - CE	75.46
CNN-BiGRU - CE	**79.36**

TABLE IV
VARYING α AND γ FOR FL BASED ON CNN-BIGRU

α	γ	Acc (%)
0.25	1	78.87
0.50	**1**	**79.39**
0.75	1	78.53
1	1	78.21
0.25	2	76.10
0.50	**2**	**80.57**
0.75	2	77.20
1	2	77.38
0.50	0	76.42
0.50	5	76.66

TABLE V
RESULTS OF DIFFERENT LOSS FUNCTIONS BASED ON CNN-BIGRU

Model	Acc (%)
CNN-BiGRU - CE	79.36
CNN-BiGRU - FL	**80.57**

parison of different models based on CE in short utterances.

By comparing the experimental results of different baseline from Table III, we choose the best performance system with CNN-BiGRU as the final baseline.

C. *The proposed method*

The objective problem in DID task as mentioned above, we further investigate the role of the focal loss (FL). The proposed method is implemented using focal loss with CNN-BiGRU model, which is introduced into the training process. In this method, it is very important to adjust weighting factor α and the focusing parameter γ. For parameter settings (α and γ), we select several sets of experience values for testing by referring to the results of other tasks [17, 18, 19]. The results of FL about varying α and γ are listed in Table IV, we find that the best performance of all experiments is when α is set to 0.5 and γ is set to 2. Under this condition, Table V shows that the experimental result of CE and FL based on CNN-BiGRU for short utterances.

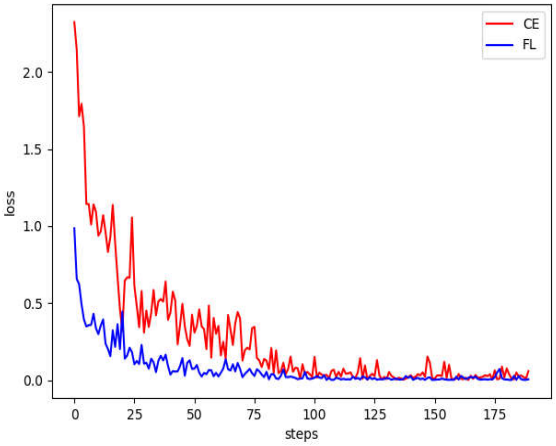

Fig. 2. Comparing loss for different loss function based on CNN-BiGRU in training (FL with $\alpha = 0.5, \gamma = 2$).

VI. DISCUSSION

It can be seen from Table III that the end-to-end systems outperform the classic i-vector system significantly for short utterances Chinese dialect identification. Especially, the results of different models with CE show that CNN-BiGRU model performs the best, which obtain 13.05% and 5.17% relative improvements than i-vector and BiGRU model respectively. It means that this model is effective in feature representation of short utterances.

For FL method, due to the great similarity between dialects that results in the existence of samples that are not easily distinguishable, Eq. (7) is used for optimization. Focal loss is equivalent to the traditional cross-entropy loss when $\gamma = 0$, and as γ is increased the effect of the modulating factor β is likewise increased. γ smoothly adjusts the proportion at which the easily classified samples are down-weighted. From Table IV, we find $\gamma = 2$ and $\alpha = 0.5$ to work best in all experiments that the recognition accuracy is 80.57% in Table IV.

From the presentation of Table V, our proposed model with FL is improved 1.21% than with CE. Moreover, Fig. 2 shows that the performance of FL is significantly better than CE by comparing loss for different loss function in training. That is, compared with CE based on CNN-BiGRU, FL has a small initial loss, which also accelerates the convergence of the network and makes its performance more stable. Therefore, our proposed method can further improve the performance of short utterances Chinese dialect identification.

VII. CONCLUSIONS

In this paper, a detailed analysis of the use of focal loss based on CNN-BiGRU for short utterances Chinese dialect identification has been presented. Specifically, it is shown that focal loss can address the problems of unbalanced and "more easily errored" samples to a certain extent by introducing weighting factor and the focusing parameter. Meanwhile, the proposed model with focal loss effectively extracts the deep

nonlinear features of short utterances and accelerates network convergence. Experimental results on 10 Chinese dialects demonstrates that our proposed model trained with focal loss further improve recognition accuracy than with cross-entropy loss, which achieves 80.57%. In future work, we plan to explore internal dataset augmentation methods on short utterances Chinese dialect identification task.

ACKNOWLEDGMENT

This work is supported by the National Key Research Development Plant of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), Inria-CAS Joint Research Project (No.173211KYSB20170061), and the Natural Science Foundation of Colleges and Universities in Jiangsu Province (No.17KJB510018).

REFERENCES

[1] Zhang C, Zhang Q, Hansen J H L, "Semi-supervised Learning with Generative Adversarial Networks for Arabic Dialect Identification," in ICASSP, 2019, pp. 5986-5990.
[2] R. Travadi, M. Van Segbroeck, and S. Narayanan, "Modifiedprior i-vector estimation for language identification of short duration utterances," in Interspeech, 2014, pp. 3037-3041.
[3] M.-G. Wang, Y. Song, B. Jiang, L.-R. Dai, and I. McLoughlin, "Exemplar based language recognition method for short-duration speech segments," in ICASSP, 2013, pp. 7354-7358.
[4] N. Dehak, P. Torres-Carrasquillo, D. Reynolds and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in Interspeech, 2011.
[5] Zhang, Qian, and John HL Hansen, "Language/dialect recognition based on unsupervised deep learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, pp. 873-882.
[6] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez and P. Moreno, "Automatic language identification using deep neural networks," in Proc. of ICASSP, 2014.
[7] Wang Geng, Wenfu Wang, Yuanyuan Zhao, Xinyuan Cai, and Bo Xu, "End-to-end language identification using attentionbased recurrent neural networks.," in Interspeech, 2016, pp. 2944–2948.
[8] Ma Jin, Yan Song, Ian McLoughlin, Wu Guo, and Li Rong Dai, "End-to-end language identification using high-order utterance representation with bilinear pooling," in Interspeech, 2017, pp. 2571–2575.
[9] A. Lozano-Diez, R. Zazo Candil, J. G. Dominguez, D. T. Toledano and J. G. Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in Proc. of Interspeech, 2015.
[10] Shon, Suwon, Ahmed Ali, and James Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," arXiv preprint arXiv:1803.04567, 2018.
[11] Zhang Q, Hansen J H L, "Dialect Recognition Based on Unsupervised Bottleneck Features," in Interspeech, 2017, pp. 2576-2580.
[12] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in Interspeech, 2014, pp. 2155–2159.
[13] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional Modelling for Short Duration Language Identification," in Interspeech, 2017, pp. 2809–2813.
[14] Shen P, Lu X, Li S, et al. "Feature Representation of Short Utterances based on Knowledge Distillation for Spoken Language Identification," in Interspeech, 2018, pp. 1813-1817.
[15] Shen P, Lu X, Li S, et al. "Interactive Learning of Teacher-student Model for Short Utterance Spoken Language Identification," in ICASSP, 2019, pp. 5981-5985.
[16] Gelly G, Gauvain J L, "Spoken Language Identification Using LSTM-Based Angular Proximity," in Interspeech, 2017, pp. 2566-2570.
[17] Lin T Y, Goyal P, Girshick R, He K, et al. "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision., 2017, pp. 2980-2988.
[18] Li J, Wang X, Li Y, "The Speechtransformer for Large-scale Mandarin Chinese Speech Recognition," in ICASSP, 2019, pp. 7095-7099.
[19] Li Y, Guo H, Zhang Q, et al. Imbalanced text sentiment classification using universal and domain-specific knowledge[J]. Knowledge-Based Systems, 2018, 160: 1-15.4.
[20] Qiuxian Zhang, Yong Ma, Mingliang Gu, Yun Jin, Zhaodi Qi, Xinxin Ma, Qing Zhou, "End-to-End Chinese Dialects Identification in Short Utterances using CNN-BiGRU," in Proc. 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, 2019.
[21] http://challenge.xfyun.cn/2018/aicompetition/tech
[22] Povey D, Ghoshal A, Boulianne G, et al. "The Kaldi speech recognition toolkit," in IEEE Signal Processing Society, 2011