A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition

Ying Zhong\textsuperscript{1,2}, Ying Hu\textsuperscript{1,2}, Hao Huang\textsuperscript{1,3}, Washour Silamu\textsuperscript{1,3}

\textsuperscript{1}School of Information Science and Engineering, Xinjiang University, Urumqi, China
\textsuperscript{2}Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region
\textsuperscript{3}Key Laboratory of Multilingual Information Technology in Xinjiang Uygur Autonomous Region

zhongyangdl3@gmail.com, huying.75@sina.com, hwanghao@gmail.com

Abstract

One of the major challenges in Speech Emotion Recognition (SER) is to build a lightweight model with limited training data. In this paper, we propose a lightweight architecture with only fewer parameters which is based on separable convolution and inverted residuals. Speech samples are often annotated by multiple raters. While some sentences with clear emotional content are consistently annotated (easy samples), sentences with ambiguous emotional content present important disagreement between individual evaluations (hard samples). We assumed that samples hard for humans are also hard for computers. We address the problem by using focal loss, which focuses on learning hard samples and down-weight easy samples. By combining attention mechanism, our proposed network can enhance the importing of emotion-salient information. Our proposed model achieves 71.72\% and 90.1\% of unweighted accuracy (UA) on the well-known corpora IEMOCAP and Emo-DB respectively. Comparing with the current model having fewest parameters as we know, its model size is almost 5 times of our proposed model.

Index Terms: Speech emotion recognition, lightweight, inverted residuals, focal loss

1. Introduction

Emotion plays an important role in daily human interactions, it helps us to contact with each other by expressing our feelings and providing feedback. Recognizing emotion from speech correctly can help intelligent spoken interaction system to understand the potential user’s intention, and further improve the user’s experience. SER is an important technology to understand human feelings. There has been a growing number of researches and applications in recent years.

Recently, deep learning has attracted increasing attention due to its outstanding performances for many tasks, more and more methods utilizing neural networks to extract valid features from raw data have emerged in the field of SER. Most of them focus on training strategy or modeling networks. However, even though CNN has been innovated to achieve better recognition performance, it needs a large training parameters. Thus, the SER task is not suitable for models with large number of parameters. Consequently, the training data for SER is extremely limited. Thus, the SER task is not suitable for models with large number of parameters. Therefore, the training data for SER is extremely limited. Thus, the SER task is not suitable for models with large number of parameters.

2. Related Work

Depthwise separable convolution [13], factorizing a standard convolution into a depthwise convolution followed by a pointwise convolution (i.e., $1 \times 1$ convolution), drastically reduces computational complexity. Specifically, the depthwise convolution performs a spatial convolution independently for each input channels, while the pointwise convolution is employed to combine the outputs from the depthwise convolution. Results reported on Xception, which is based on depthwise separable convolution, showed that the absence of any non-linearity leads to both faster convergence and better final performance. MobileNets [15] is based on a streamlined architecture that uses depthwise separable convolutions to build lightweight deep neural network. Subsequently, the MobileNetV2 [14] presented inverted residuals and linear bottlenecks. It is also based on the depthwise separable convolution. The authors found that it’s important to remove non-linearities in the narrow layers in order to maintain representational power. Inspired by Xception and MobileNetV2, we build a model based on separable convo-
The proposed lightweight model. The entry flow maps the log-Mels of an utterance to a high-dimensional representation and the middle flow extracts richer information. The exit flow outputs predicted class label.

3. The Proposed Method

In this section, we describe the proposed lightweight model as shown in Figure 1. The input features first go through the entry flow which using a 2-D convolution layer with stride of 2, then through the middle flow which is used to automatically extract discriminative feature representations. Finally, these feature representations further produce higher level features for SER through the exit flow.

3.1. Inverted Residual

Figure 1 describes a complete architecture of proposed model. The entry flow extracts shallow information from the features with variable length. We evaluated proposed model on the IEMOCAP and Emo-DB corpus.

3.2. Attention Layer

In the exit flow, attention mechanism is applied after a Bi-RNN which compresses variable length sequences produced by middle flow to a fixed-length vector. Each direction of Bi-RNN contains 128 Gated Recurrent Units (GRUs) [4]. Then we can obtain a sequence of 256-dimensional high-level features by con-
catenating the outputs of two directions. The attention layer is employed to focus on emotion relevant parts and produce discriminative utterance-level representations for SER [9]. As shown in (1), the weight $\alpha_i$ is first computed by a softmax function, where $h_i$ is the Bi-RNN output, then the utterance-level representations $c$ are calculated by performing a weighted sum on $h_i$ according to the weights, as shown in (2).

$$\alpha_i = \frac{\exp(W \cdot h_i)}{\sum_{t=1}^{T} \exp(W \cdot h_t)}$$  \hspace{1cm} (1)

$$c = \sum_{t=1}^{T} \alpha_t h_t$$  \hspace{1cm} (2)

Finally, the utterance-level representations are passed into a fully connection layer with 64 output units, then followed by $PRelu$ [4] activation function and use one softmax layer to calculate the probability of per emotion.

### 3.3. Focal Loss

The training of a deep network is based on updating the network parameters to minimize a loss function that expresses the divergence between the predictions and the ground truth labels [19]. For SER, each sentence is often annotated by multiple raters, which are aggregated with methods such as majority vote rules. The inconsistency of evaluations of emotional content may lead to that emotion recognition becomes more difficult. In addition, the imbalance of categories in training data also makes SER more difficult. A common method for addressing class imbalance is to introduce weighting factors [4, 17]. We assigned weights to cross entropy (CE) loss, shown in Eq.(3), the weight $w_i$ is in inverse proportion to the sample number of the class in training set, $\hat{y}_i$ is the $i$-th element of network predictions. The weighted CE loss as follows:

$$CE_w = -\sum_{i=1}^{m} w_i y_i \log(\hat{y}_i)$$  \hspace{1cm} (3)

Focal loss was proposed in [17] to address class imbalance and hard examples by focusing on learning hard examples and down-weight easy examples. As shown in Eq.(4), adding a factor $(1 - \hat{y}_i)^\lambda$ to the weighted cross entropy, where $\lambda$ is hyper-parameter adjusting the rate at which easy examples are down-weighted. Setting $\lambda > 0$ reduces the relatives loss for well-classified examples, putting more focus on hard and misclassified examples. When $\lambda = 0$, the model is trained using only weighted cross entropy loss. As $\hat{y}_i$ closes to 1, the factor goes to 0 and the loss for well-classified examples is down-weighted.

$$FocalLoss = -\sum_{i=1}^{m} w_i (1 - \hat{y}_i)^\lambda y_i \log(\hat{y}_i)$$  \hspace{1cm} (4)

### Table 1: Exemplary complete annotations of utterance of Ses01F_impro03_F025.

<table>
<thead>
<tr>
<th>Annotators</th>
<th>Annotations</th>
<th>Label</th>
<th>Vote proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-E1</td>
<td>Happiness</td>
<td>hap</td>
<td>4/6</td>
</tr>
<tr>
<td>C-E2</td>
<td>Happiness</td>
<td>hap</td>
<td>4/6</td>
</tr>
<tr>
<td>C-E4</td>
<td>Happiness; Excited</td>
<td>hap</td>
<td>4/6</td>
</tr>
<tr>
<td>C-F1</td>
<td>Happiness; Excited</td>
<td>hap</td>
<td>4/6</td>
</tr>
</tbody>
</table>

![Figure 3: (a) The distribution of vote proportion of label and (b) samples distribution in IEMOCAP database.](image-url)

### 4. Experiments and Results

#### 4.1. Datasets

The IEMOCAP [20] database containing 12 hours English conversations is employed for performance assessment. They are segmented and categorized into utterances with 9 emotion classes. We conducted the classification task only on the same 5 emotion classes as [4, 10, 21]. Same to the reported procedure, utterances in exciting class are combined to the happy class in evaluation, to form a four-class database labeled with {happy, angry, sad, neutral}, each class contains {1636, 1103, 1084, 1708} utterances respectively. Each utterance is labeled by three or four annotators and the classification label is the majority label among the annotations [22]. As shown in Table 1, there are six annotations, four of which are happiness, thus the vote proportion of label hap is 4/6.

The distribution of vote proportion of labels is shown in Fig.3(a). The distribution of vote proportion of labels is shown in Fig.3(a). As shown in Table 1, there are six annotations, four of which are happiness, thus the vote proportion of label hap is 4/6. The distribution of vote proportion of labels is shown in Fig.3(a). We define the samples with vote proportion of 1 as easy samples (A), between 1 and 0.65 as medium difficulty samples (B), and less than 0.65 as the difficulty samples as shown in Fig.3(b). Emo-DB [23] consists of 535 utterances that displayed by ten professional actors, covering seven emotions {angry, bored, disgust, fear, happy, sadness, neutral}. The number of each class is {127, 81, 46, 69, 71, 62, 79} and all seven emotions are used for our tasks. The sample rate of IEMOCAP database is 16kHz. The Emo-DB database sampled at 44.1kHz, and later downsampled to 16kHz. Because each sample in Emo-DB has only one annotation, so we just apply focal loss on IEMOCAP dataset. The code is available at [1].

#### Table 2: The validation of the effect of attention layer.

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th>Emo-DB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UA(%)</td>
<td>WA(%)</td>
</tr>
<tr>
<td>No Attention</td>
<td>68.21</td>
<td>66.73</td>
</tr>
<tr>
<td>Attention</td>
<td>70.51</td>
<td>69.63</td>
</tr>
</tbody>
</table>

#### Table 3: Average results (%) of three kind of losses.

<table>
<thead>
<tr>
<th>Losses</th>
<th>UA</th>
<th>WA</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft_loss</td>
<td>69.95</td>
<td>69.05</td>
<td>69.29</td>
</tr>
<tr>
<td>Lq_loss</td>
<td>68.74</td>
<td>68.51</td>
<td>68.75</td>
</tr>
<tr>
<td>Focal_loss</td>
<td>71.72</td>
<td>70.37</td>
<td>70.37</td>
</tr>
</tbody>
</table>

4.2. Experimental Settings

We used 128-dimensional log scale Mel-spectrogram (log-Mel) as input features [4, 12, 24]. The spectrogram is extracted using 1024-point short-time Fourier transformation (STFT) with 25% overlap. Neumann et al. found that 7s long utterance contains enough emotional information [12]. So if the utterance is longer than 7s, only the middle part with the length of 7s was calculated.

We employed TensorFlow to implement the proposed method. Adam is used as optimizer. Learning rate was set with 0.0003 and batch size 64. Both of datasets were divided into 10 subsets randomly keeping the emotion distribution, 8 subsets were used for training, one for validation and one for testing (fixed test set). The experimental results are the average of 9 times cross validation. We use three metrics, Unweighted Accuracy (UA), Weighted Accuracy (WA) and F1-score, to evaluate proposed method.

4.3. Results and Discussions

In our first set of experiments, we evaluate the effect of attention layer in Exit Flow (Fig.1) using IEMOCAP and Emo-DB datasets. The system adapted weighted CE loss (Eq.3) for training. Table 2 shows the effect of attention mechanism. No Attention in Table 2 indicates that the outputs of Bi-RNN are fed into fully connection layer directly but not pass the attention layer.

Then, we conducted experiments on IEMOCAP adopting focal loss (Eq.4) and comparing with two kinds of losses[19]. For easy samples (A), \( \lambda = 0 \) (the loss of easy samples was calculated by Eq.3). \( \lambda_B \) and \( \lambda_C \) denote the values of \( \lambda \) for medium difficulty samples (B) and difficulty samples (C). Table 3 shows that when \( \lambda_B = 1, \lambda_C = 2 \), UA, WA and F1-score can achieve the best performance. \( \lambda_B = 0, \lambda_C = 0 \) means that the model was trained by using weighted CE loss. When \( \lambda_B \) was increased to 1.5 and \( \lambda_C \) remained 2, the performance decreases slightly. The medium difficulty samples accounts for nearly 64% of training data. So increasing \( \lambda_B \) slight means that it will reduce the contribution of difficulty samples relatively. The followed experiments all adopted the settings of \( \lambda_B = 1 \) and \( \lambda_C = 2 \).

Followed, we explored the performance of the system with various number of inverted residual blocks. As shown in table 4, the model achieves the best performance on IEMOCAP when the number of blocks is 5, and on Emo-DB when the number of blocks is 4. Using focal loss, the performance of model increased by 1.7%, 1.08% and 1.3% respectively for UA, WA and F1-score compared with weighted loss (with blocks are 5).

It further proves that focal loss can facilitate the generalization of network. Although the network become deeper, the amount of parameters doesn’t increase obviously.

Finally, we compared our system with several baseline on IEMOCAP and Emo-DB datasets. As shown in table 5, the proposed model has achieved significant improvement comparing to state-of-the-art models with their reported results especially on Emo-DB dataset. 3D-ACRNN has the largest number of model parameters which are almost 32 times of DRN and 73 times of the BCRNN. Among the compared models, BCRNN has the least number of model parameters, however, its model size is almost 5 times of our proposed model. Meanwhile, our proposed model achieves better performance with only fewer parameters.

5. Conclusions

To facilitate the SER application to real-time system, we proposed a lightweight model based on separable convolution network and inverted residuals. By employing attention layer, the model can focus on the parts of emotion relevant. The model also use focal loss to address the problem of class imbalance and difficult samples, and to help the network focus on learning hard examples. IEMOCAP and Emo-DB databases are used to evaluate the performance of the model in terms of UA, WA and F1-score. Results indicate that our proposed model can yield better results compared with state-of-the-art models with fewer parameters.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (U1903213, 61761041, 61663044), Natural Science Foundation of the Xinjiang (2016D01C061) and University Scientific Research Project of Xinjiang (XJEDU2017T002).
7. References


