META-LEARNING FOR SPEECH EMOTION RECOGNITION
CONSIDERING AMBIGUITY OF EMOTION LABELS

Takuya Fujioka¹, Takeshi Homma¹, Kenji Nagamatsu¹

¹Research & Development Group, Hitachi, Ltd., Tokyo, Japan.
{takuya.fujioka.qh, takeshi.homma.ps, kenji.nagamatsu.dm}@hitachi.com

Abstract

Emotion labels in emotion recognition corpora are highly noisy and ambiguous, due to the annotators’ subjective perception of emotions. Such ambiguity may introduce errors in automatic classification and affect the overall performance. We therefore propose a dynamic label correction and sample contribution weight estimation model. Our model is based on a standard BLSTM model with attention with two extra parameters. The first learns a new corrected label distribution and aims to fix the inaccurate labels in the dataset. The other estimates the contribution of each sample to the training process and aims to ignore the ambiguous and noisy samples while giving higher weights to the clear ones. We train our model through an alternating optimization method, where in the first epoch we update the neural network parameters, and in the second we keep them fixed to update the label correction and sample importance parameters. When training and evaluating our model on the IEMOCAP dataset, we obtained a weighted accuracy (WA) of 65.9% and 61.4%, respectively. This yielded an absolute improvement of 2.3% and 1.9%, respectively, compared to a BLSTM with attention baseline, trained on the corpus gold labels.

Index Terms: speech emotion recognition, meta-learning

1. Introduction

Automatic recognition of emotion is important to enable more natural and engaging communication between humans and machines. In this work we concentrate on emotion recognition from speech, which is the task of estimating the emotional content of a spoken utterance.

In the past, emotion recognition was performed by extracting a set of low-level features from each frame of an audio sample. These features were then aggregated through various statistical aggregation functions (mean, standard deviation, minimum, maximum, etc.) to a global utterance-level vector representation [1], to be finally fed through a shallow classifier such as support vector machines (SVMs) [2, 3]. However, in recent years, the accuracy of speech emotion recognition has dramatically improved with the introduction of deep neural networks (DNNs). Initial DNN-based models [4] were still based on the same utterance-level feature extraction. However, in subsequent approaches, speech features extracted from each frame were used as inputs of more complex neural network architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and the accuracy was further improved [5, 6, 7]. Recent years have seen the application of novel methods developed from other AI fields, such as self-attention models [8], connectionist temporal classification (CTC) [9] and dilated residual network (DRN) [10]. Even higher performance was achieved by using multi-modal information, such as audio and image together with speech [11].

While most of the studies concentrated on the development of more accurate classification models, there were other aspects of emotion classification regarding the data itself that could help improve the performance but were mostly ignored. In many datasets, the emotion labels are annotated based on human annotators’ perceptions to emotion. Emotion perception is highly subjective [12]; therefore, the labels often contain some noise due to humans’ decision ambiguities. For instance, an annotator may assign the label neutral not when the sample is actually neutral, but when he/she is unsure about the most appropriate emotion class. Likewise, he may mistakenly recognize some loud enthusiastic speech as angry, but actually it is happy. Training a model on such noisy labels is likely the cause of some performance degradation because the model may become confused and may not clearly distinguish one emotion from another.

Another important issue is that, in many emotion recognition datasets, the numbers of utterances for each emotional category are imbalanced. Generally, in a classification task using these category imbalanced datasets, the accuracy of the small class is decreased [13, 14], which in turn affects the overall accuracy. To overcome these problems, some methods have been proposed to use soft target approaches to correct the annotation ambiguities [15] or to augment the dataset with synthetic data to reduce the effect of the data imbalance [16]. However, the former method only performs a static label contribution estimation based on the original annotation data, while the latter method is complex and the generated data might still be affected by the original labeling noise. In other domains, such as image recognition, similar issues were tackled by performing a label update, not a priori but during training, by gradually tuning the estimation [17].

Inspired by the achievements in image recognition [17], we propose a method to automatically tune the contribution of each data sample during training. We do this by alternately updating the parameters of a DNN emotion classification model, and then use the neural network prediction to correct the relative contribution and the target labels of each sample, to reduce the overall loss. The main purpose is to correct or ignore altogether the ambiguously labeled utterances, while giving higher importance to the clear and unambiguous ones. The results obtained in the interactive emotional dyadic motion capture (IEMOCAP) dataset [18] show that our proposed method is effective in removing the annotation noise. It achieves an improvement of 2.5% for weighted accuracy, and of 2.7% for unweighted accuracy compared to a state-of-the-art BLSTM model trained on the original labels only [7].

2. Methodology

Given an input audio sample $x_n = [x_{n,1}, x_{n,2}, \cdots, x_{n,T}]$, where $n$ is the utterance index, $x_{n,t}$ is the frame-based feature vector, and $T$ is the total number of frames, we want
to estimate the probabilities of each emotion category \( y_n = [y_{n,1}, y_{n,2}, \cdots, y_{n,C}] \), where \( C \) is the number of discrete emotion classes. We use a BLSTM model with attention [7] to perform the classification.

To improve the classification performance and reduce the ambiguities of the human-annotated labels during training, for each training speech sample we also learn two parameters: \( l_n = [l_{n,1}, l_{n,2}, \cdots, l_{n,C}] \), and \( w_n \). \( l_n \) represents a new estimate of each sample emotion class, aiming to correct the ambiguities and inaccuracies during the training process, while through \( w_n \) we learn the contribution weight for each training utterance.

### 2.1. Emotion classification model

Fig. 1 shows the structure of our main BLSTM emotion recognition model, which closely follows the state-of-the-art by [7]. At first, we input the feature sequence \( x_n \) through a bi-directional LSTM (BLSTM), which yields \( h_n = [h_{n,1}, h_{n,2}, \cdots, h_{n,T}] \) as the output. We then weight the contribution of each frame through an attention layer, where its weights \( \alpha_{n,t} \) are calculated as follows:

\[
\alpha_{n,t} = \frac{\exp(h_{n,t}^u)}{\sum_{\tau=1}^{T} \exp(h_{n,\tau}^u)}.
\]

In the equation above, \( u = [u_1, u_2, \cdots, u_C] \) are the learned attention parameters. The obtained attention weights \( \alpha_{n,t} \) are used to calculate the weighted average over time of the BLSTM output vectors, to get a fixed-length utterance-level vector representation \( h'_n \). We get the output emotion probabilities \( y_n \) by applying a softmax layer to \( h' \):

\[
h'_n = \sum_{t=1}^{T} \alpha_{n,t} h_{n,t},
\]

\[
y_{n,c} = \frac{\exp(h'_{n,c})}{\sum_{c=1}^{C} \exp(h'_{n,c})},
\]

### 2.2. Update of target labels and contribution weights

Normally, an emotion classification system such as the one explained in the previous section, is trained from the gold standard labels, which closely follows the state-of-the-art by [7]. At first, we input the feature sequence \( x_n \) through a bi-directional LSTM (BLSTM), which yields \( h_n = [h_{n,1}, h_{n,2}, \cdots, h_{n,T}] \) as the output. We then weight the contribution of each frame through an attention layer, where its weights \( \alpha_{n,t} \) are calculated as follows:

\[
\alpha_{n,t} = \frac{\exp(h_{n,t}^u)}{\sum_{\tau=1}^{T} \exp(h_{n,\tau}^u)}.
\]

In the equation above, \( u = [u_1, u_2, \cdots, u_C] \) are the learned attention parameters. The obtained attention weights \( \alpha_{n,t} \) are used to calculate the weighted average over time of the BLSTM output vectors, to get a fixed-length utterance-level vector representation \( h'_n \). We get the output emotion probabilities \( y_n \) by applying a softmax layer to \( h' \):

\[
h'_n = \sum_{t=1}^{T} \alpha_{n,t} h_{n,t},
\]

\[
y_{n,c} = \frac{\exp(h'_{n,c})}{\sum_{c=1}^{C} \exp(h'_{n,c})},
\]

### 3. Experiments

To evaluate the performance of the proposed learning method, we use the IEMOCAP dataset [18], one of the most commonly used benchmark datasets in emotion recognition tasks. The
corpus is organized into 5 sessions, in each of which two actors performed a conversation. The total number of speakers in the corpus is 10. We only considered the samples belonging to the four emotion categories of happiness, sadness, neutral and anger, to keep the analysis consistent with previous works [6, 7, 8, 9, 10, 11, 15, 16]. The number of utterances in each emotional class of each speaker is shown in Table 1. We performed a leave-one-speaker-out 10-fold cross-validation using a leave-one-out strategy [10]. We applied early-stopping criteria in all conditions to minimize the loss of the validation set [15].

3.2. Experimental setup

We extracted 32-dimensional acoustic features from the raw audio samples using the openSMILE toolkit [19]: 12-dimensional mel-frequency cepstral coefficients (MFCCs), loudness, fundamental frequency (F0), voicing probability, zero-crossing rate, and their first order derivatives. The frame length and frame shift were set to 25 ms and 10 ms, respectively. All features were normalized by mean and standard deviation calculated over all of the utterance features in the training set.

The emotion classification model was composed of a fully-connected layer with rectified linear unit (ReLU), a BLSTM layer and a fully-connected layer. The numbers of hidden units were 512, 128 and 4, respectively. We applied dropout to all layer and a fully-connected layer. The numbers of hidden units were normalized by mean and standard deviation calculated over all of the utterance features in the training set.

We evaluated our model using two common evaluation measures in the previous works: weighted accuracy (WA) and unweighted accuracy (UA). We also calculated per-class precision, recall and F1-score, to get a performance estimate over each individual emotion class.

We compare our model (BLSTM + ATT + L + w) against the following baselines:

- **BLSTM + ATT**: our reimplementation of the attention based BLSTM model as proposed in [7].
- **BLSTM + ATT + Oversampling/Undersampling**: same as the above, but applying oversampling or undersampling instead of w to address the class imbalance problem. In this baseline, utterances are randomly resampled in each epoch.
- **BLSTM + ATT + L**: the full model and training algorithm, but only updating L, while keeping w fixed.
- **BLSTM + ATT + w**: the full model and training algorithm, but only updating w, while keeping L equal to the gold standard labels.
- **BLSTM + ATT + L + w pretrained**: similar to the full model, but the neural network was first pretrained using the gold standard labels.
- **Soft-target**: the soft label method proposed in [15].
- **Cycle-GAN**: the data augmentation method proposed in [16].

3.3. Results

The final results are shown in Table 2. For the Soft-target and Cycle-GAN baselines, we show the reported results from the original papers, since in the former case we were unable to replicate the same results, and the latter used a different methodology. Our proposed model performed the best in terms of WA, and the second best in terms of UA, achieving 65.9% and 61.4%, respectively. This yields an absolute improvement of +2.3% and +1.9%, respectively, over the BLSTM + ATT baseline. It is also worth noticing that the originally reported WA and UA in [7] are 63.5% and 58.8%, respectively. These values are not significantly different from the ones obtained by our reimplementation. The soft-target baseline achieved a 2.3% higher UA than our proposed method. However, they only used one-fold cross validation instead of ten-fold, and they did not report the performance for the individual emotion classes [15]; therefore, the results are not fully comparable.

In terms of per-class performance, our model achieves an F1-score of 35.7%, 67.1%, 63.3% and 75.9% for the happiness, sadness, neutral and anger classes, respectively, with absolute improvements of +0.5%, +2.1%, +2.4% and +1.9%, respectively. The lower improvement in the F1-score of happiness is compensated by a significant improvement in precision of +14.6%.

3.4. Discussion

By looking at the results in Table 2, it clearly emerges how our proposed model achieves a much better performance than just applying some simple imbalance corrections such as data undersampling or oversampling. In terms of performance, the introduction of per-sample importance weighting w had a slightly higher effect than emotion correction parameter L presumably because it is less sensitive to errors. w had the main effect of improving the precision in happiness, and of improving the precision in anger, while the main contribution of L was to improve the recall for neutral samples. Pretraining the model with the original labels did not seem to work better than starting by immediately updating L and w, presumably due to a greater
evaluation, and the characteristics of the L transition are similar to those of the imbalance dataset evaluation, especially regarding happiness, anger and neutral. This result supports the motivation of label correction by L and handling imbalance by w. Besides, BLSTM + ATT + L + w scored 61.5% and 59.4% for WA and UA, respectively, while BLSTM + ATT scored 60.1% and 56.9% for WA and UA, respectively. The contribution of updating L and w to improve WA and UA is also confirmed in this evaluation.

4. Conclusions

We have proposed a novel meta-learning approach, built on top of a traditional BLSTM with attention classifier, to address the issue of labeling inaccuracy and ambiguity in speech emotion recognition. Our proposed method is effective for dynamically updating each sample label during training, and learning an estimate of each sample contribution to reduce the relative weight of ambiguous utterances. We obtained an overall performance of 65.9% and 61.4%, for weighted and unweighted accuracy, respectively, on the IEMOCAP dataset, giving an absolute improvement of 2.5% and 2.7%, respectively over the same BLSTM model trained on the original gold labels. We also showed how our proposed framework clearly managed to reduce the importance of the most ambiguous label (happiness), and fix the initial label annotation to the most appropriate classes for each sample, thus improving the classification performance.

5. Acknowledgements

The authors would like to thank Dr. Dario Bertero for his valuable suggestions to improve the manuscript.
6. References


