Title: Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition (ICASSP 2018)

Author: Po-Wei Hsiao and Chia-Ping Chen

Summary: This paper is focused on speech emotion classification. Based on former LSTM-RNN models, the authors present a new LSTM with attention mechanism. The proposed framework first extracts acoustic features of speech signal, then feeds them into an attention-based LSTM model. The attention layer is inserted between LSTM layer and the output layer. Experiments on baseline models HMM and RNN prove the effectiveness of the proposed method.

Good Things about This Paper

1. The proposed model takes the sparsity of speech emotion into consideration by using attention mechanism, which has the ability to locate and focus on the salient and reliable parts of the signal that contains useful emotion representations.
2. The authors also pay attention to the unbalanced data issue and apply class weight to relieve this problem, as most other researcher did not notice.
3. Emotions are expressed differently for each individual. The authors apply speaker normalization for this problem.

Major Comments

There is a growing trend to feed row data instead of handcrafted features into the neural networks, as the NNs are good at figuring out the linear and non-linear combinations of the input vectors themselves. So, the authors may want to use row speech input instead of or more promisingly, apply several CNN layers to extract multi-level features automatically [1].

Comparing table 6 with table 8, the results show that the performance varies among the different classes of the output. Firstly, the accuracy of class anger is greatly improved by 48%, which is astonishing. Then, emphatic is almost the same, and the rest class stays hard to classify. However, we can also see an accuracy drop in neutral and positive classes. The authors did not explain or discuss this result, but it is important because it is relevant to evaluate the effect of introducing the attention mechanism. To explain this, we can see deeply into the attention weight distribution of each class. The anger class is negative skewed, and the natural and positive classes are more like a Gaussian distribution. Also, anger distributions of train and test set are different. Maybe the authors can look deeper into acoustics characteristics or deeper compare train and test set to find the reason.

Besides, because the rest class in a catch-all class contains all data that not belong to any of the other classes, the authors think their model cannot represent such a highly twisted space. So in the conclusion part, the authors said one of the future work they will do is introducing a deeper network to solve the problem. However, as the rest class seems to contain no useful information, it needs discussion if we need to bigger our model and use more computational sources for this.

[1] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5089-5093.

Minor Comments

1. Section 3.1 Data: speech chucks are not described in detail.
2. Fig. 3: no units of the axis.
2. Fig. 4: no units of the axis; abbreviation ‘t-SNE’ not described before and after.