Title: Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition
Author: Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina

Summary: This paper focused on emotion recognition from music tracks. The proposed model first extracts its baseline feature set with openSMILE toolbox, then calculates the log Mel-band energies as raw audio features. After that, a stacked convolutional and recurrent neural network (CRNN) is described and it is compared with another CRNN without branching (CRNN_NB) then. Finally, the proposed model is tested on different sequence length from 10 to 60 and also on different feature sets respectively. Compared with the Li (2016) system, experiments results outperform the baseline and the average training time is reduced.

Good Things about This Paper

1. Based on the belief that neural networks have the ability to learn first and second derivative information from raw features on its own, the proposed method uses log Mel-band energies instead of extracted features, which highers accuracy and lowers training consumption.
2. The model has fewer parameters, as the number of features is reduced.
3. This method trains two models to predict arousal and valence separately, which is good because these two arises are thought to be orthogonal.

Major Comments

Firstly, although it is tested in the evaluation stage, different sequence length should be viewed as an important influence of model architecture. As temporal context and hierarchical structures on emotion in music are considered as the major motivation of model design in some papers [1], it is a shame to lose the information of emotion variation in different levels of musical structure. The varying sequence length analysis and its following analyzing procedures could be integrated into the proposed model.

In the training process, the proposed model uses convolutional neural network for automatic feature extraction, and then applies recurrent networks for prediction. As the RNN has some mechanism for memorizing context, it is possibly better to employ method with longer memories as it has been proved more efficiency in some former literatures. LSTM could be a suitable choice, BLSTM is even better with both past and future memories. Furthermore, attention mechanism could be considered as it potentially provides longer memories than the methods above.

Another problem is about the testing procedure. As it is described in the 5th paragraph of introduction, the best reported result is a LSTM with varying-length sequence input, which achieved a RMSE of 0.203, 0.303, 0.253. However, the baseline chosen by the authors is Li (2016), whose RMSE is 0.285, 0.225, 0.255, which is not the best performance existed. Why not compare to the best one? Is there anything common between the proposed and Li's model than drives the authors to choose it as baseline? Needs more explanation.

In addition, this paper possibly did not repeat the baseline model as the model details are not given in the original paper. And the reasons for choosing the parameters in figure 1,2 is are not mentioned.

[1] Li, X., Tian, J., Xu, M., Ning, Y., & Cai, L. (n.d.). DBLSTM-BASED MULTI-SCALE FUSION FOR DYNAMIC EMOTION PREDICTION IN MUSIC. 2016 IEEE International Conference on Multimedia and Expo (ICME), 1–6.

Minor Comments

1. Fig. 1, 2: The use of multiple annotation is not consistent (* and x), or needs further explanation.
2. Fig. 1, 2: The numbers on the figures e.g. 60x16 need more clarification.