Title: DBLSTM-Based Multi-scale Fusion for Dynamic Emotion Prediction in Music
Author: Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, Lianhong Cai

Summary: This paper is focused on dynamic music emotion prediction for music retrieval and recommendation. The proposed procedure first extracts low-level features with openSMILE toolbox, then put them in a bidirectional LSTM-RNN model with different sequence length. After that, a hybrid fusion and post-processing model is applied to integrate the output of the former BLSTM model. Finally, emotion classification result is generated. Experiments among different methods prove the effectiveness of the proposed method.

Good Things about This Paper

Music has both internal hierarchical structures and textual continuity, which former researches paid little attention to. The proposed method utilizes the hierarchical information by using different sequence lengths when training the RNN. As the sequence length indicates different levels of music structure, this procedure helps to capture the emotion variation in different levels. And it also explores the temporal context with long memory models like LSTM, both frames in the past and in the future.

Major Comments

The most significant drawback of this paper is the lack of explanations and theoretical discussions of the model performance. That is to say, it seems that the authors tried several similar structures and pick up one model with the best testing result, with no explanations of their choice and no further discussions. For example, the authors put forward several combinations of fusion and post-processing in section 3.3.3 and figure 2. Then a testing result is given later indicating one of these combinations is more effective than the others. But the reasons why the combinations are designed like that or why one specific architecture outperforms the others are not discussed. Besides, whether the best combination works as good on other datasets is not mentioned either. The same problem comes repeatedly in sections 5.1 and 5.3 when the authors find out the best sequence length is 10 and the best fusion method is neural networks. Why is that? Does the 10-length performers the same in different music genres? This issue generally hurts the significance of this paper.

Besides, there are other minuses in the problem definition part. The title of this article includes the word “dynamic”, but what “dynamic prediction” is not identified throughout the paper. What is called dynamic? Does it mean the sequence length in the BLSTM training changes? Or the prediction is on every frame? Or indicates the LSTM evaluates different locations at different times? Need to be clarified. Also, from the title, the DBLSTM stands for Deep-BLSTM, how deep it is? BLSTM has already become a commonly used terminology, by adding “D” before it, the authors may want to answer what makes the proposed model special in depth?

Furthermore, some parts of this article are not detailed enough. The architecture and major parameters of neural networks and SVMs described in section 3.3.3 are not given, which makes the reimplementation of the new model difficult to some extent. And according to fig.3, the different hierarchical structures are treated equally (with the same weights), which indicates that different levels of structure are same important. Is this agree with the music theories?

Lastly, the authors may want to try their model with raw data inputs rather that extracted features, since the fact that NNs can learn by themselves and function like a feature extraction mechanism.

Minor Comments

1. Fig. 2: The figure should mark distinguish different V/As in different location of the model.
2. Section 2: Several mathematical annotations are not clearly defined. For example, R, R^M.