IAML Project#2 Music Mood Classification

Team 5 : 이 은, 한창진

1. Objective

본 프로젝트의 목표는 30초의 노래가 주어졌을 때 RNN을 이용한 Mood Classification 모델을 만드는 것이다.

2. Dataset

30초 길이의 노래 8499곡으로 label은 10개의 Mood(happy, film, energetic, relaxing, emotional, melodic, dark, epic, dream, love)로 이루어져 있다. 한 곡당 multi-label을 가질 수 있다. Training set은 그 중 6326개를 사용하였으며 각 mood당 727~1152곡으로 이루어져 있다. Validation set은 그 외 곡들로 2173곡으로 이루어져 있다. 각 mood당 비율은 training set과 유사하다.

3. Feature

Log-Mel-spectrogram Feature를 이용하여 학습을 진행하였다. Project#1에서 진행했던 것과 같이 Audio classification task나 tagging task에서 흔하게 쓰이는 feature이며, 다양한 model에서 안정적으로 우수한 성능을 보인다는 연구에 기반해 이를 사용하기로 하였다. 이외에 고려한 feature로는 Neural net을 이용하기 전 SVM 등을 이용한 mood classification에서 유용하다고 밝혀진 zero crossing rate, spectral centroid, spectral rolloff도 최종 모델에서 사용되진 않았으나 고려한 모델도 시도해보았다. [1]

4. Model Structure

Project#1에서 Convolutional Neural Network를 이용한 것과 유사한 방식으로 Audio data에 맞게 horizontal, vertical 필터 사이즈를 이용해서 각각 CNN – RNN 을 순서대로 쌓은 후 최종적으로 이 두 CNN-RNN 을 거친 feature들을 concat하여 fully connected layer을 통과시켰다.

Horizon filter로는 7개의 convolution layer와 6개의 max-pooling layer로 구성하였고 Vertical filter은 4개의 convolution layer와 3개의 max-pooling layer로 구성하였다. Convolution layer의 Activation function으로는 ReLU(Rectified Linear Unit)를 사용하였고 각 convolution layer 이후에

[1] Cyril Laurier(2011), Automatic Classification of Musical Mood by Content Based Analysis, Universitat Pompeu Fabra, pp.69-70.

Batch normalization을 사용하였다. 또한 ReLU를 위한 weight initialize 방법으로는 He initialization을 이용하였다. 이렇게 Convolution step에서는 Convolution – activation – batch normalization – max-pooling 과정을 거치도록 하였다. 그 후 각각 Bidirectional LSTM을 2회 통과시켰다. (각 unit = 128, 32) 이후 이 둘을 최종적으로 concat하여 fully connected layer에 dropout(keep prob=0.5)를 적용하였다. 이 때 softmax로 각 장르별 예측값을 output으로 한다.

이 외에 최종 모델에 사용하지는 않았으나, 해본 시도는 다음과 같다. 1) multi-label을 가질 수 있기에 마지막 layer에서 activation function을 softmax 대신 sigmoid도 시도해봤으나 softmax가 더 나은 결과를 보였다. 2)마지막 concatenation에 위의 '3. Feature' 에서 언급한 세 가지 zero crossing rate, spectral centroid, spectral rolloff(이하 hand crafted feature로 칭함)를 Fully connected layer 3개를 통과시킨 embedding을 같이 concat하여 Fully connected layer도 통과시켜보았으나 이 역시 좋지 않은 성능을 보였다. 또한 학습 결과가 training loss는 계속해서 낮아지나 validation loss가 높아지는 경향을 보여 parameter가 많아 생기는 over-fitting이 의심되었다. 3) 그래서 CNN block을 없애고 pure RNN만을 3회 통과시킨 후(Bidirectional LSTM, unit=128, 64, 32) hand crafted feature embedding과 concat 후 Fully connected layer을 통과시키는 방법도 진행해보았으나 over fitting issue는 해결되지 않았고 최종 모델에 비해 좋은 성능을 보이지 못했다.

아래 그림1은 model의 최종적인 전체 구조를 나타낸다.

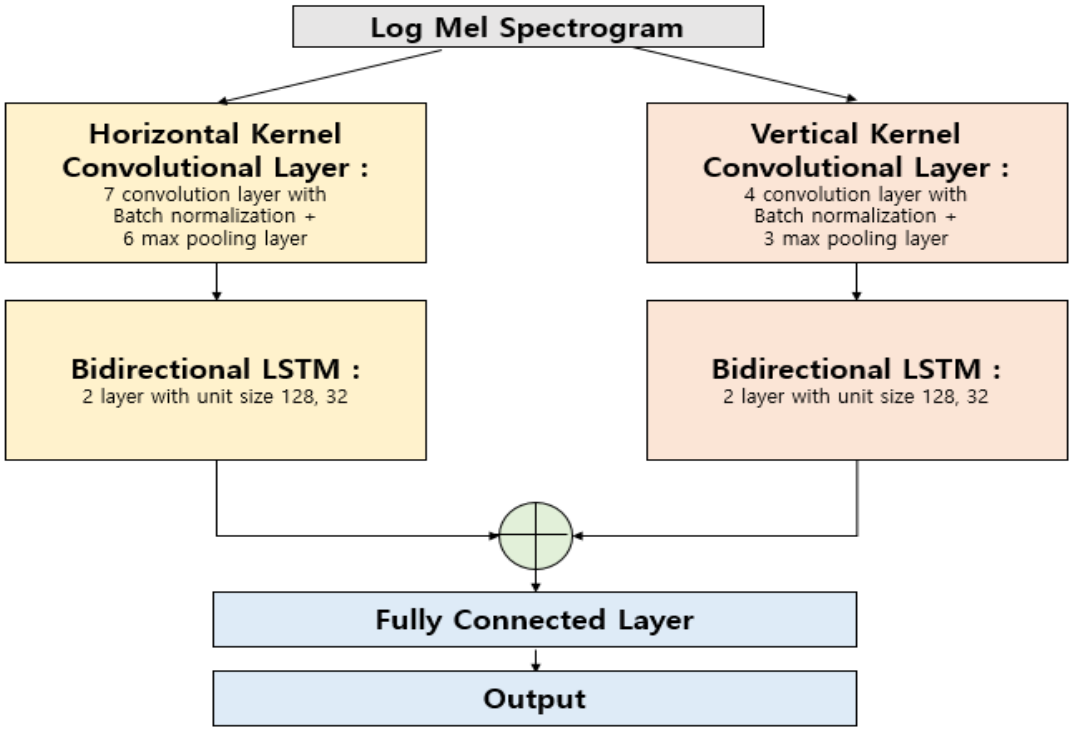

그림 1. 최종 모델 구조

5. Model Training

Optimizer로는 Adam optimizer를 이용하였고 learning rate의 초기값으로 0.001을 사용하였는데 learning rate decay 방식을 적용하여 매 training step 마다 learning rate가 0.3배가 되도록 조정하였다. Epoch은 50, Batch size는 32로 설정하여 training을 진행했다. Multi-label이 가능했기에 loss는 categorical이 아닌 BinaryCrossEntropy()를 사용하였다.

6. Result

본 모델을 사용한 결과 validation set에 대한 ROC-AUC score의 평균이 0.705임을 확인했다.