IAML Project#3 Singing Voice Melody Extraction

Team 4 : 윤 현, 한창진

1. Objective

본 프로젝트의 목표는 30초의 노래가 주어졌을 때 Deep learning model을 이용한 Singing Voice Melody Extraction 모델을 만드는 것이다.

2. Dataset

주어진 데이터셋은 30초 길이의 노래 7497곡이며, Training set은 그 중 5589개를 사용하였고 Validation set은 그 외 곡들로 1908곡을 사용했다. 각 노래의 label은 note 단위 40개(A2~B5 + non-voice)로 이루어져 있다. Frame 단위는 0.125초이며 30초의 노래 하나 당 240개의 순차적 label을 가진다.

3. Feature

Log-mel-spectrogram feature를 이용하여 학습을 진행하였다. 이전의 프로젝트에서 진행했던 것과 같이 mel-spectrogram 계열의 featue는 Audio classification task나 tagging task에서 흔하게 쓰이는 feature이며, 다양한 model에서 안정적으로 우수한 성능을 보인다는 연구에 기반해 이를 사용하기로 하였다. 이 때 log-mel-spectrogram이 mel-spectrogram에 비해 좋은 성능을 보인다는 연구에 기반해 log-mel-spectrogram을 사용하였다.[1] 그리고 이번 task에서 가장 중요한 것이 feature와 label 간에 time alignment를 시켜주는 것인데 이를 위해 Log-mel-spectrogram feature를 추출할 때 frame_step은 512, sampling rate은 4096으로 설정하여 melody length가 240이 되도록 맞춰주었다. 이외에 고려한 feature로는 Constant-Q-Transform이 있는데 mel-spectrogram보다 music genre classification에 우수한 성능을 보인다는 연구[2]에 기반하여 사용해보았으나, Log-mel-spectrogram을 사용했을 때보다 성능이 낮고 실험 시간이 오래 걸려 최종 모델의 input feature로

---

[1] Choi, Keunwoo & Fazekas, George & Cho, Kyunghyun & Sandler, Mark. (2017). On the Robustness of Deep Convolutional Neural Networks for Music Classification

[2] Lidy, T., & Schindler, A. (2016, September). CQT-based convolutional neural networks for audio scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop(DCASE2016) (Vol. 90, pp.1032-1048). DCASE2016 Challenge.

는 사용하지 않기로 하였다.

4. Model Structure

모델은 2016년 12월에 CNN+CRF모델로 music chord recognition 분야에서 state-of-the-art의 성능을 기록한 논문[3]을 참조하여 Convolutional Neural Network(CNN)로 구성했다. 총 8개의 convolution layer와 3개의 pooling layer(2개의 max-pooling, 1개의 avg-pooling)로 구성된 모델인데 6개의 convolution layer에는 activation function으로 Leaky-ReLU를 사용했고 나머지 2개에는 각각 ReLU와 linear function을 적용하였다. 또한 각 convolution layer 이후에 Batch normalization을 사용하였다. 이후 dropout(keep prob=0.5)을 적용한 후 Fully Connected Layer에 적용하였다. Fully Connected Layer의 경우 주어진 뼈대 코드와 유사하게 하나의 1 X (240 * 13) 길이의 FC Layer으로 적용 뒤 240 X 13으로 reshape하는 방식과 TimeDistributed Layer를 사용하여 길이 13의 FC Layer를 240개 만드는 방식을 비교하였다. 후자의 방법에서 더 좋은 성능을 얻을 수 있어 후자의 방법을 채택하였다. (전자의 경우, melody length를 240으로 맞춰주기 위하여 frame step을 2757, sampling rate을 22050 적용하였다.)

이 외에 pitch와 voice를 CRNN을 이용해 분리해낼 수 있는 모델 구조를 논문[4]에서 참고하여 구현해보았으나 위의 모델보다 좋은 성능을 보이지 못해 사용하지 않았다.

다음 장의 그림은 최종 모델의 전체적인 구조를 나타낸다.

5. Model Training

Epoch은 50, Batch size는 32로 설정하여 training을 진행했으며 Optimizer로는 Adam optimizer를 사용했다. learning rate의 초기값으로 0.001을 사용하였는데 learning rate decay 방식을 적용하여 매 training step마다 learning rate가 0.3배가 되도록 조정하였다. Loss function으로는 Sparse Categorical Cross Entropy 를 사용했다.

6. Result

본 모델을 사용한 결과 validation set에 대한 Accuracy가 0.627임을 확인했다.

---

[3] Filip, K. & Gerhard, W. (2016 December). A FULLY CONVOLUTIONAL DEEP AUDITORY MODEL FOR MUSICAL CHORD RECOGNITION. In Proceedings of IEEE 2016 workshop.

[4] Kum, S., & Nam, J. (2019). Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks.

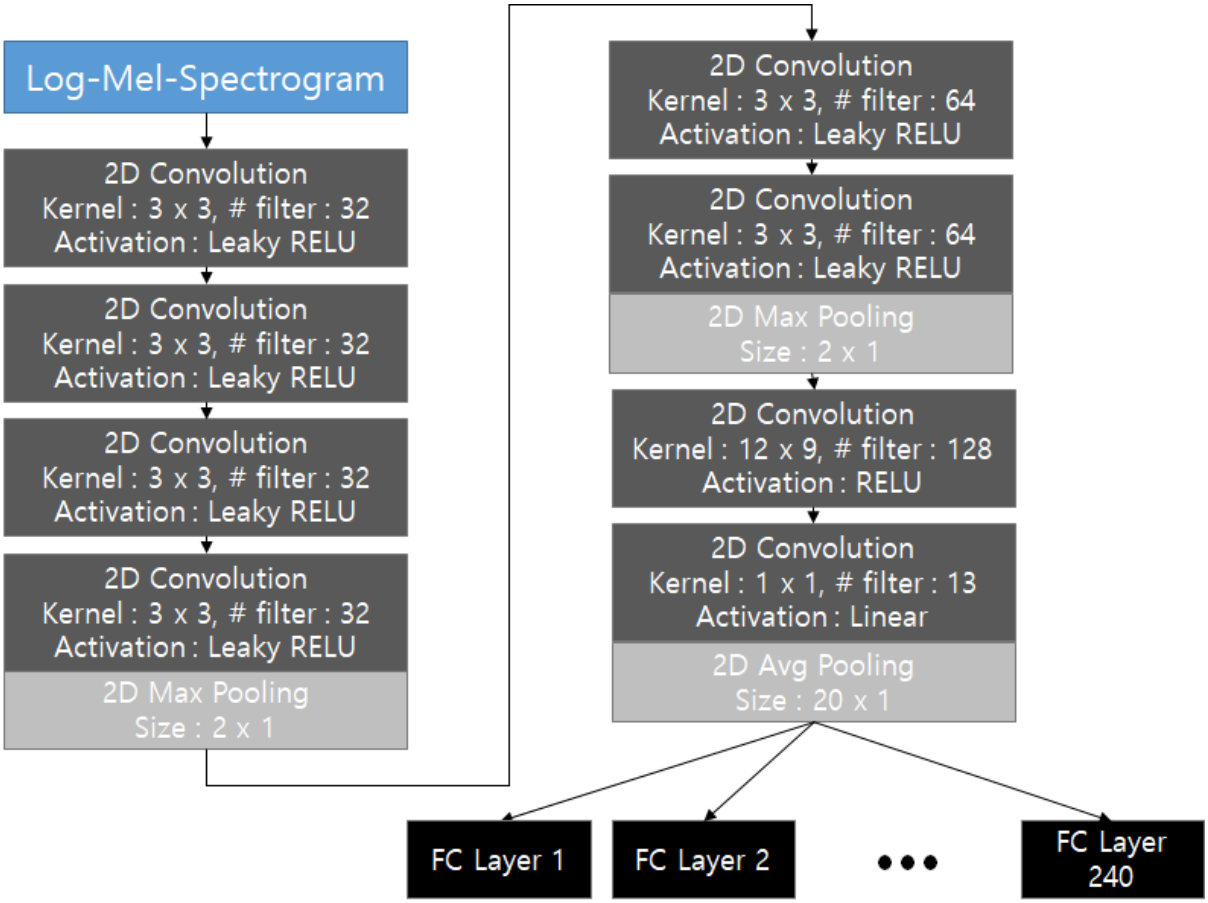

Figure 1. Model Structure

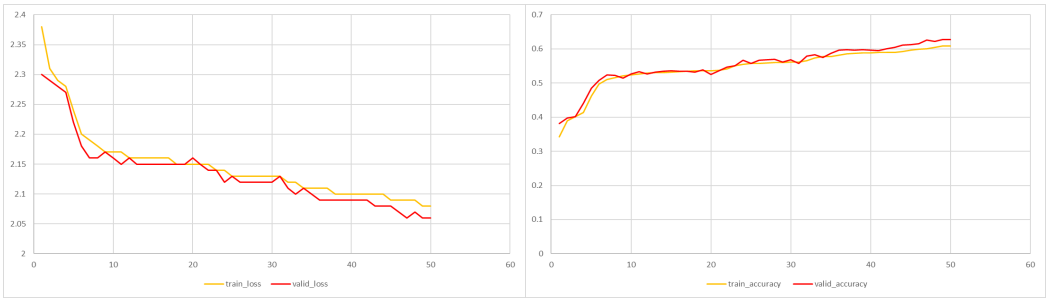

Figure 2. Loss and Accuracy