Multi-label Musical Instrument Recognition Classification

Bretan Cezar-Alexandru
Huy Truong

## Introduction

In this project, we chose to investigate the performance of Convolutional Neural Networks, given the task of tagging audio files containing fully-mixed musical pieces, based on the incorporated instruments. The tags should be applied based on the instrument-specific sounds that were detected at any point within an audio clip.

## Data Description

We used the OpenMIC-2018 open dataset [2], which consists of 20000 10-second audio snippets of music in OGG Vorbis format collected from Spotify, each clip having multiple out of a total of 20 possible labels assigned to it by numerous annotators of varying “trust levels”.

The dataset authors also provided CSV files with the audio filenames forming a training and testing 75-25 data split and aggregated labels via binary numpy dumps.

## Feature Extraction

The audio files from the dataset have varying sampling rates, so we resampled all of them to 32kHz. We observed the values in the regular Mel-Spectrograms extracted from the waveforms range from 0.0 to ~500, so, at first, we applied a log10(1+|N|) operation to them in order to clamp the values to a narrower range to potentially improve the numerical stability for the model.

Unlike the original implementation, we separated the feature extraction from the training, and dumped the log-mel-spectrogram and label pairs via h5py files.

However, when we began experimenting with transfer learning, the results with this extraction setup were unexpected and we thought that it must be related to the fact that the pre-trained models used a slightly different extraction methodology.

The second extraction setup we experimented with clamped the range of the Mel-Spectrogram values to a ~-60 to ~30 range, representing values on the dB scale.

## Model Selection & Training Setups

We used the CNN14 proposed network from paper [1], with a modified final linear classifier layer for outputting 20 instrument labels instead of 527 sound classes. This model seemed suitable for our task since the dataset used in the original paper [3], does contain some music-related classes and the model itself is also used for audio tagging.

The authors also provided some pre-trained models, that we did end up using for transfer learning.

We ran training using the first extraction methodology with different amounts of data, and found massive improvements in the Recall metric the more data we used. More specifically, when using 50 batches of 32, the Recall couldn’t break past 60%, but upon using all of the ~15000 training snippets, it broke past 70%.

Afterwards, we loaded the weights from a pre-trained model and replaced the final linear layer such that it outputs a 20-dimensional vector of probabilities.

The results were unexpectedly poor, at around 35% Recall, but once we switched to the second feature extraction methodology, the Recall reached 66% when training using only 50 batches from the training split and reached 75% when using the full training split.

## Results

We ran a series of experiments with different features, prediction thresholds, learning rates, extracted features, data amounts and with or without fine-tuning a pre-trained model. All testing results were run on the full testing split.

- **Config I**: Log-Mel-Spectrograms obtained via log10(1+|N|), training done from scratch using full training split, learning rate = 1e-4;
- **Config II**: Log-Mel-Power Spectrograms obtained via conversion to dB range, training done from scratch using full training split, learning rate = 1e-4;
- **Config III**: Log-Mel-Power Spectrograms obtained via conversion to dB range, training done by fine-tuning using 100 batches of 32 from the training split, learning rate = 1e-4;
- **Config IV**: Log-Mel-Power Spectrograms obtained via conversion to dB range, training done by fine-tuning using full training split, learning rate = 2e-5;
- **Config V**: Log-Mel-Power Spectrograms obtained via conversion to dB range, training done by fine-tuning using full training split, learning rate = 1e-4;
- **Config VI**: Same model as **Config V**, prediction threshold = 0.3;
- **Config VII**: Same model as **Config V**, prediction threshold = 0.7.

| | Config I | Config II | Config III | Config IV | Config V | Config VI | Config VII |
|---|---|---|---|---|---|---|---|
| Recall | 71.16 % | 75.43 % | 72.81 % | 78.18 % | 78.54 % | 83.62 % | 73.12 % |
| Accuracy | 96.63 % | 96.57 % | 96.75 % | 97.23 % | 97.12 % | 96.63 % | 97.33 % |
| Precision | 77.29 % | 73.70 % | 78.81 % | 80.85 % | 79.09 % | 71.03 % | 86.07 % |
| Test Loss | 0.092 | 0.092 | 0.091 | 0.074 | 0.081 | 0.081 | 0.081 |

We also put together a script for compiling the pairs of audio filenames and the predicted tags into a single text file, such that the system's outputs are easily parsable and verifiable.

Upon some manual listening tests, our observations match what the objective measurements suggest. Config IV is the most consistent performer, but by lowering its prediction threshold, it does often output more labels, but with a significantly increased false positive rate.

## Conclusion

Through this project, we demonstrated the feasibility of musical instrument audio tagging and the efficacy of transfer learning for this specific task.

The results don’t quite breach the performance threshold of what’d be required for such a system to be described as “satisfactory” for production usage, but results could be improved by employing attention mechanisms or residual connections and using augmentation techniques.

One more important takeaway from our experiments would be that the Accuracy metric can be very deceiving when it comes to multi-label classification.

[1] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2880-2894.

[2] Eric J. Humphrey, Simon Durand, Brian McFee. OpenMIC-2018

[3] J. F. Gemmeke et al., “Audio Set: An ontology and human-labeled dataset for audio events,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 776–780.