Executive summary

Objective of the study

The purpose of the report is to investigate the performance of convolutional neural networks (CNN) over one and two-dimensional data representation of audio clips in a music genre classification task. We are studying whether a designed CNN model (potentially combined with a RNN) is able to extract useful features from two forms of audio data (samples and spectrograms) and evaluate the encoding capability of each representation based on the performance of the model trained on such setup.

Proposed approach

The dataset used for this experiment is the MagnaTagATune (MTAT) dataset. The first 10 most popular labels were used for the multi-class multi-label classification. A detailed analysis of the dataset is performed to reveal connections between labels and between content and labels.

A collection of eight model architectures from four major classes are proposed to be trained on the (MTAT) dataset:

- Simple 1D CNN;
- 1D CNN + RNN/LSTM/GRU;
- Chunk 1D CNN (with GlobalMaxPooling reduction);
- 2D CNN + RNN/LSTM/GRU.

All 1D models use raw samples as input, while the 2D models use spectrograms.

All models are trained using the same learning setup. The models are optimized according to cross-entropy loss and validated using binary accuracy.

The performance discussion takes into account the per-genre precision, recall, AUC-ROC and AUC-PR.

Conclusion and further directions

Statistical evidence if found that some major types of music tags (slow vs fast, classical vs techno) could be accurately separated even without employing machine learning techniques. However, the nuances of genres are deeply entangled in the data, thus dimensionality reduction projections fail to identify any helpful cluster structure.

The deep learning models tested has similar results in both 1D and 2D forms, achieving around 80-90% accuracy. This confirms the efficiency of general purpose architectures and their converging capabilities even in their simplified form (i.e. not engineered in the same amount as state-of-the-art solutions in the field). Both 1D and 2D CNN-GRU have comparable performance with literature models like Musicnn and Harmonic CNN.

This work can be improved by trying out more related architectures, adjusting hyper-parameters, and using music-specialized Mel spectrograms instead of ones that handle general signal processing.