Music Tagging using Deep Learning on Mel-Spectrogram Image

TRAN, Thanh Cong
Student ID: 2210421

Problem

- Music tag classification: Predict the tags of music clips
  - Genre: Blue, Rock, Pop, …
  - Instruments: Bass, Drum, Acoustic Guitar, …

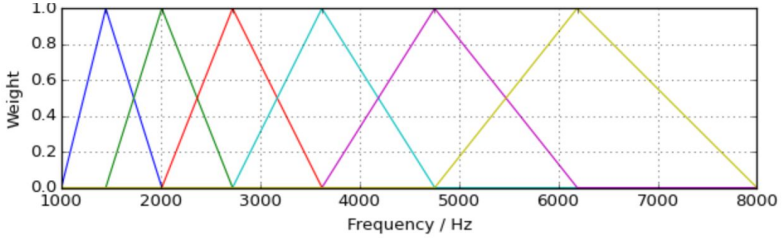
- Approach:
  - Convert the waveform representation (Time-domain) into Mel-Spectrogram representation (Time-Frequency domain)
  - Train a deep neural network on the Mel-Spectrogram image of music clips to classify tag of each clip

Mel-Spectrogram Representation

- Waveform to Mel-Spectrogram:
  - Normalize the amplitude of waveform to unit norm
  - Apply Short-time Fourier Transform to the normalized waveform to get the Spectrogram

$$X_i[k] = \sum_{n=1}^{N} x_i[n] * e^{-j2\pi kn/N}$$

  - Generate the Mel filter bank



$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

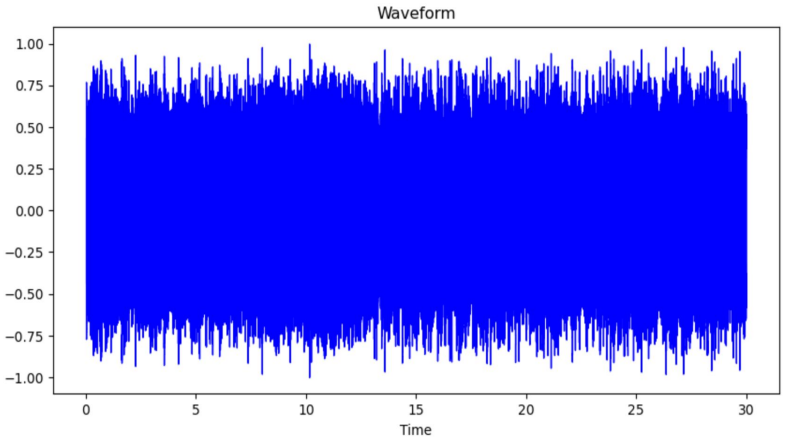
Mel-Spectrogram Representation

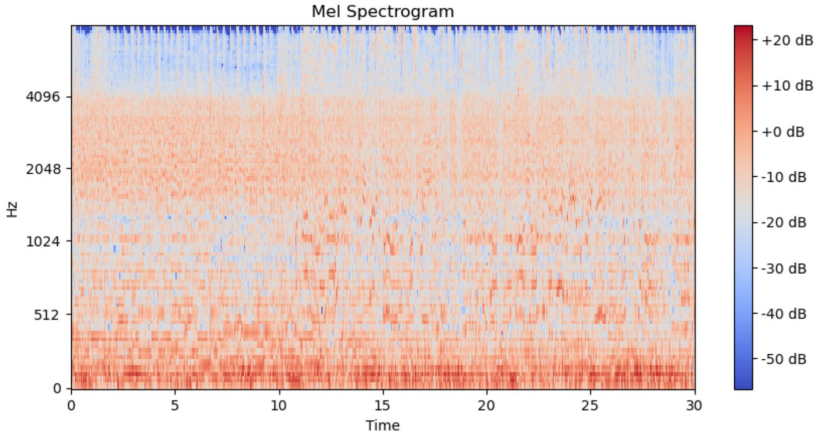
- Waveform to Mel-Spectrogram (cont):
  - Apply the Mel filter bank to each STFT window to generate the Mel-Spectrogram Representation

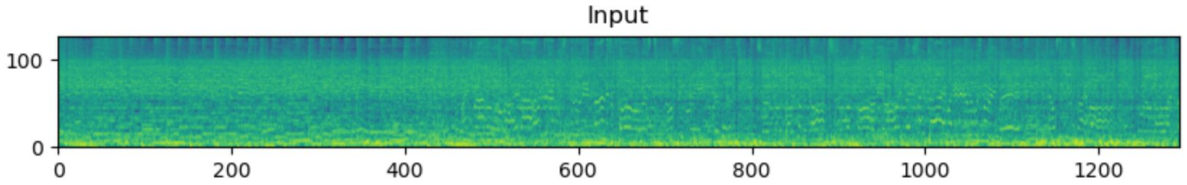
$$\tilde{X}_i[m] = \sum_{k=1}^{K} H_m[k] * X_i[k]$$

  - Transform the amplitude of the Mel-Spectrogram to Decibels unit
  - Convert each Mel-Spectrogram representation into a 2D image to form the input data

Mel-Spectrogram Representation

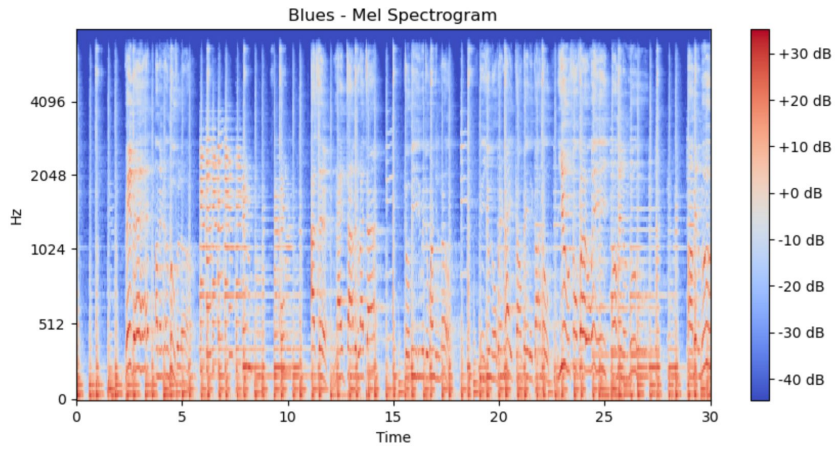


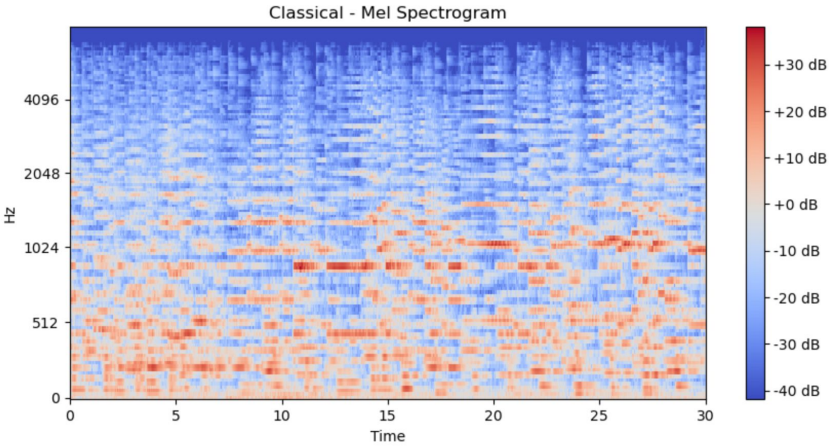




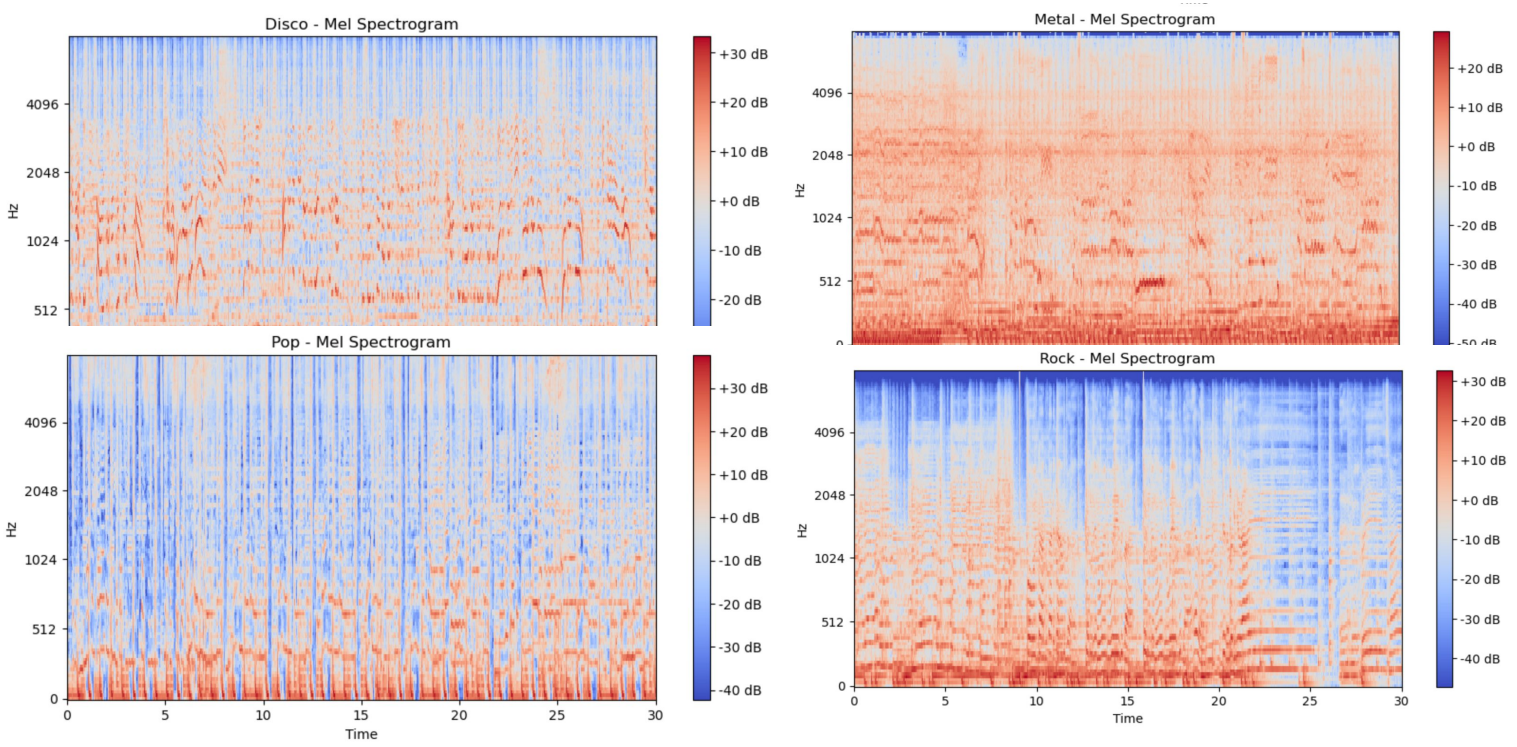
Dataset: GTZAN

- GTZAN: Music Genre Classification Dataset
- Consists of 1000 30-second mp3 files with 10 genres
- Single-label classification: Each genres has 100 mp3 files



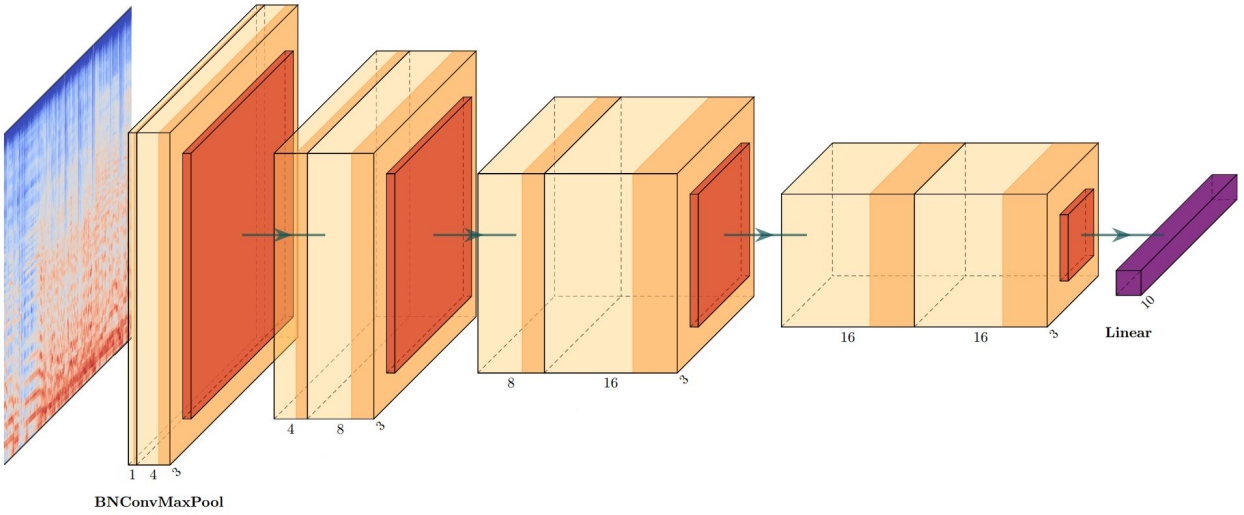


Dataset: GTZAN



Deep Learning Architecture

- Deep convolution network: consists of convolution and linear layers



Experiment

- Implementation:
  - Programming Language: Python 3
  - Waveform to Mel-Spectrogram: librosa and numpy
  - Deep Learning: Pytorch
- Short-time Fourier Transform configuration:
  - Sample rate: 22050 Hz
  - Window Size: 2048
  - Hop Length: 512
- Number of Mel bands: 128

Experiment

- Train / Test split: 80/20
- Metric: Accuracy

Blues	Classical	Coutry	Disco	Hiphop
90	95	75	75	90

Jazz	Metal	Pop	Reggae	Rock
85	95	85	75	60

- Average accuracy: 81.5 %

Thank you