Music Genre Classification using CNNs

1. Finding a phenomenon and a question to ask about it

Introduction

Audio Deep Learning is the sub field of Deep Learning where the data being used is Audio. Its major applications are Speech recognition systems and Music Industry. Recently the Google introduced a noise cancellation in their Google Meets. This is implemented through Neural network where the input and outputs are audio. Audio has major impact in our lives.

2. Understanding the state-of-the-art

The simplest audio classification problem is the Music Genre classification. The data set which will be used is the GTANZ .

The audio files in the data set are of the extension wav. Usually Librosa library is used in audio signal processing. The simplest method of audio classification is using the only Time domain analysis of signals which are passed to NN to classify the genre of the audio.

The research has progressed from only Time domain to Both Frequency domain and time domain. MFCCs (Mel-Frequency Cepstral Coefficents) is one technique used in pre-processing data.

3. Determining basic ingredients

- What is a spectrogram?

Spectrograms are derived from the spectrums. The audio in time domain analaysis is the plot of Amplitude versus time. This can be converted to frequency using Fourier transformations. Librosa provides two types of transformations FFT and STFT. For generating spectrums FFT is used and for spectrograms STFT is used. Spectrograms depict frequency on y -axis and time on x-axis and colour bands show the amplitude in dB.

Ideal audio features for audio classification:

1. Time- Frequency representation
2. Perceptually-relevant amplitude representation

3. Perceptually-relevant Frequency representation

- Do spectrograms fulfill all three ideal features?

They provide with the appropriate representation for the first two points but there seems a bit inaccurate representation of Frequency in the spectrogram as human do not perceive pitch as exactly as frequency. They do have a correlation but not in linear form. Researchers have suggested a different scale for this called as Mel Scale as from experiments it is most accurate resemblance of the frequency as pitch.

Mel scale is given by m = 1127 ln(1 + f/700)

As the Mel scale is the most accurate , this scale is used to generate Mel-Spectrograms which will be method of data pre-processing.

- Could Spectrograms be treated as images which would further processed using Computer Vision techniques?

They can be treated as images and processed with the help of Convolutional Neural Networks (CNNs). The image processing has long history of advancements which can be used for Audio Deep learning.

4. Formulating specific, mathematically defined Hypothesis

Hypothesis:

Part 1: To Implement a CNN architecture for Music Genre Classification using Mel-spectrograms

Part 2: To detect multiple genres present in a mixture of genres in a song(This can be compared to semantic segmentation problems in Computer Vision.)

Note: Part 2 of Hypothesis depends on the outcome of the experiment of Part 1. The model will only planned for Part 1.

5. Selecting the Toolkit

Toolkit to be used is PyTorch library's torchaudio for data processing and Models from PyTorch which are usually used for classification problems in CV like ResNet, Densenet,etc

6. Planning the Model

Data augmentation is widely used in the Computer Vision. It can be used in audio deep learning as well. The proposed data augmentations are:

1. Time Stretching
2. Pitch shifting
3. Adding Background Noise
4. Artificially increasing the audio files.

After data augmentation, Mel sepctrograms of the audio files will be obtained which will used as images which could be further processed using CNNs.

From the ImageNet challenge we have a lot of types of CNNs which could be used for the classification.

The proposed CNN architecture is ResNet but may be changed with keeping hypothesis in mind.

7. Implementing Model

8. Completing the model

9. Testing Model

The model will be evaluated based on the following factors:

1. Validation accuracy
2. Training Time