Music Genre Classification

Based on MUSIC GENRE RECOGNITION WITH DEEP NEURAL NETWORKS by Albert Jimenez, Ferran José DSAP course, Master MET, ETSETB

INTRODUCTION:

- What we want to do?

Recognize the music genre of a song

- What we have?

GTZAN database: 10 genres (pop, reggae, rock,..) 100 songs per genre

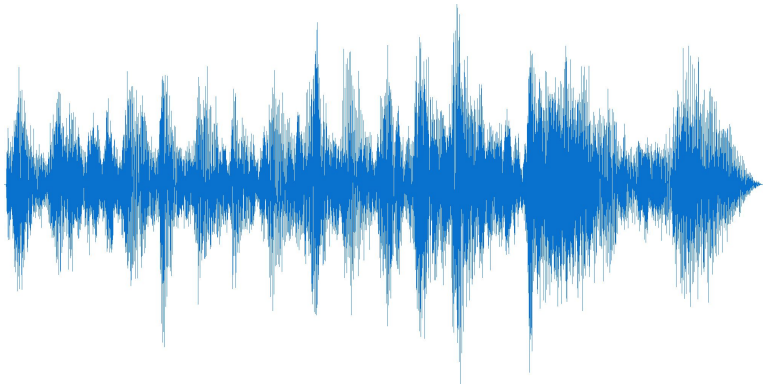
- How to do it?

Convolutional networks filled by raw audio data or features

Input: Data representation

Raw audio

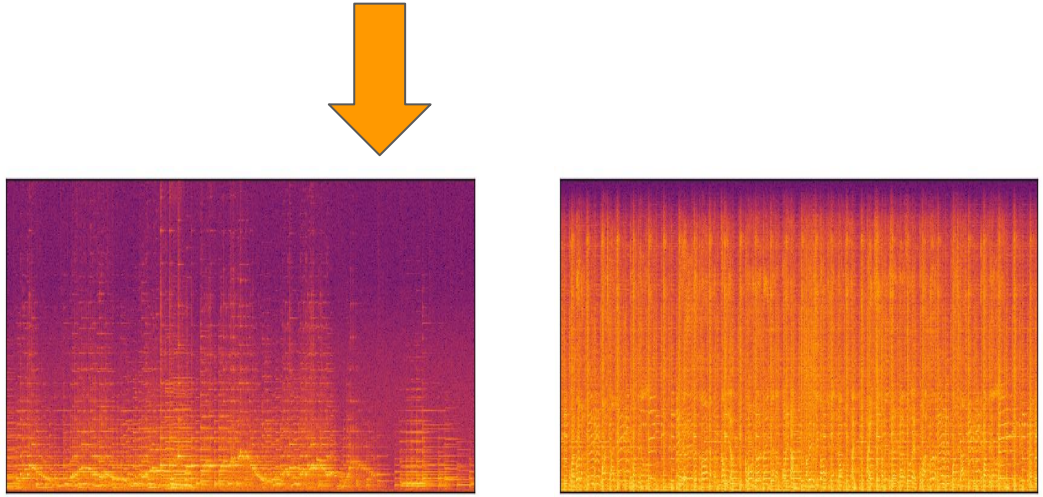
The pure signal isn't very handy though - mainly because it's quite heavy. Feeding it to a CNN was an option, but since it's computationally expensive, we abandoned that idea.



Input: Data Representation II

We need to be able to detect frequencies (tones) over short period of time (for example to notice the chords used in a song), but also need to look on the song as a whole.

Spectogram



Genre: classical

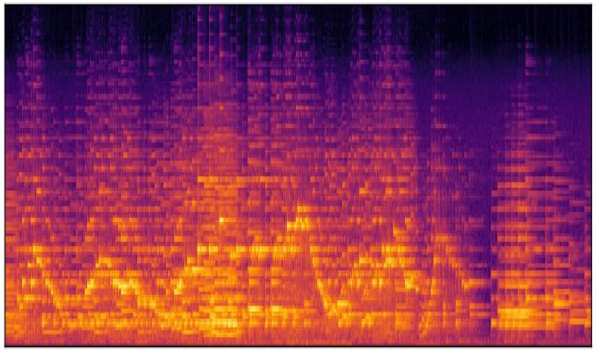
Genre: blues

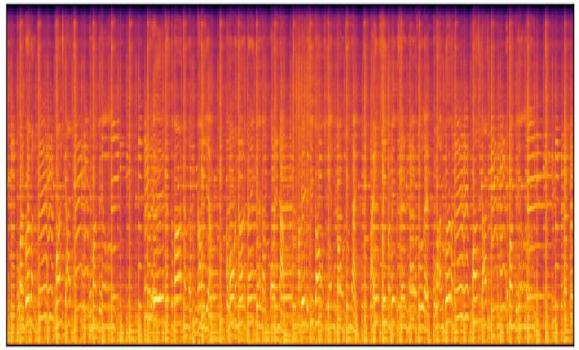
Data Representation:

Better Spectograms → Better Results

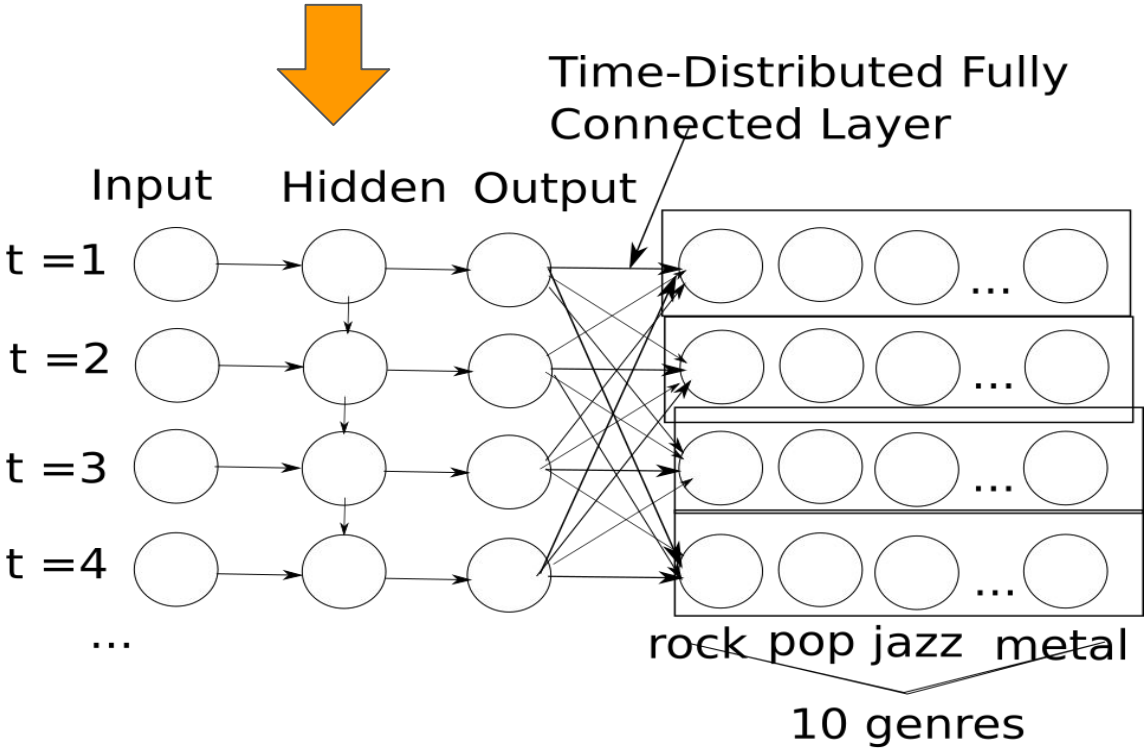
MFCC :

A mel-spectrogram is a spectrogram transformed to have frequencies in mel scale, which basically is a logarithmic scale, more naturally representing how human actually senses different sound frequencies.





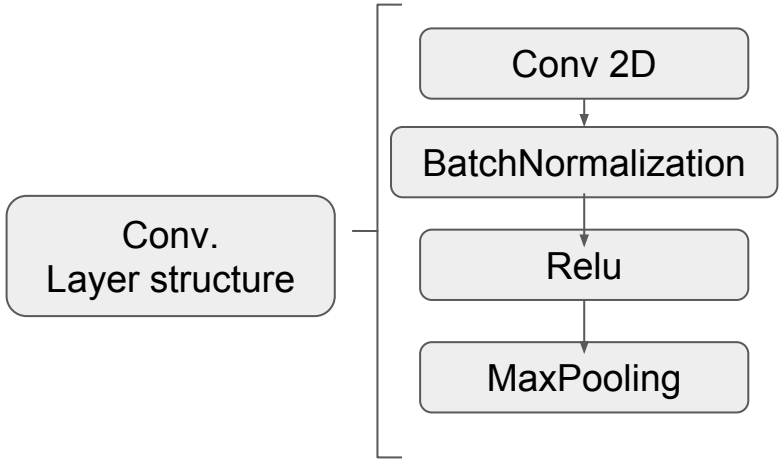
Neural network architecture:

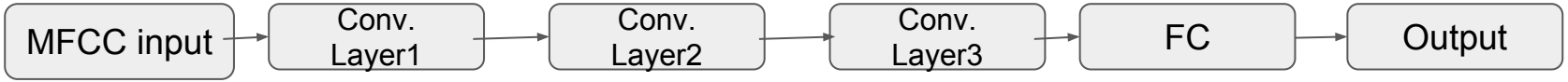


DNN structure:

Types of transformations used:

- 2D convolutional
- Batch normalization
- Non-linear transform: relu
- MaxPooling2D
- Fully connected
- Dropout
- Softmax

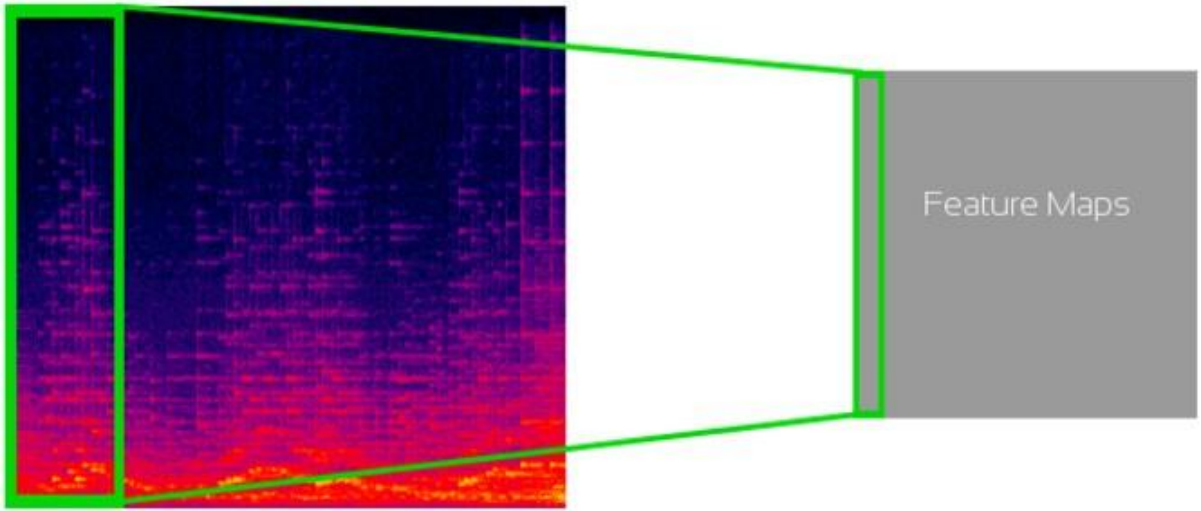




Convolutional Neural Network

The convolutional layers are used to extract the features from the MFCC matrixes.

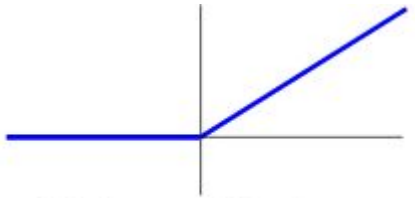
Every convolution layer look at small period of time in order to extract valuable info and create a feature map. (still being a sequence across the time)



Activations

ReLU (Rectified Linear Units):

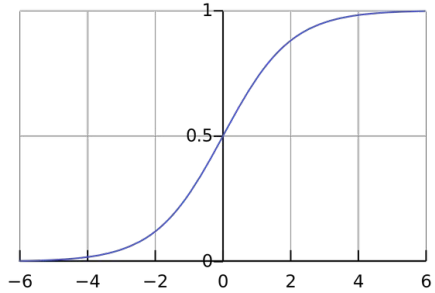
- Non linear funciton
- Efficient
- Rapid convergence in training
- Usually used in hidden layers



$$\phi(z_i) = \max(0, z_i)$$

Softmax:

- Used in output layer
- Multi-class Classification
- Used to represent categorical distribution

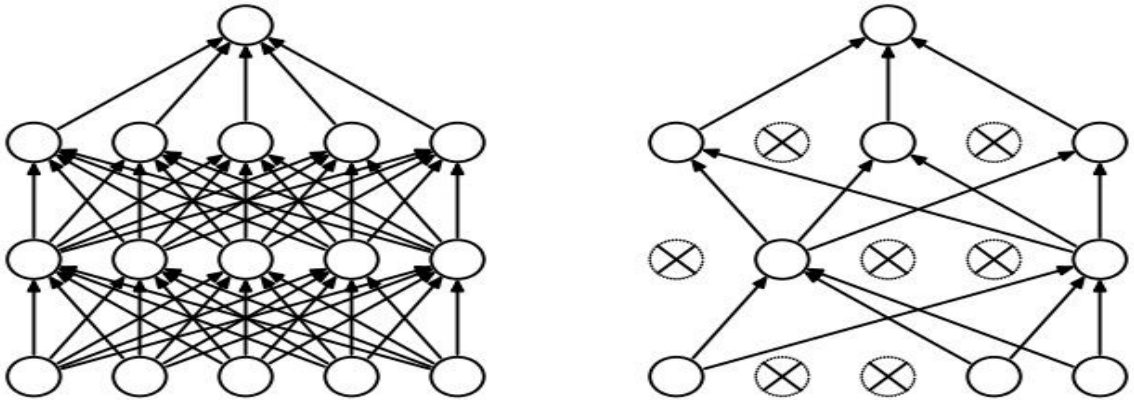


$$\phi(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)}$$

Dropout

Dropout is applied to avoid overfitting

It consists on putting some outputs to 0 (with probability P) while amplifying the other ones



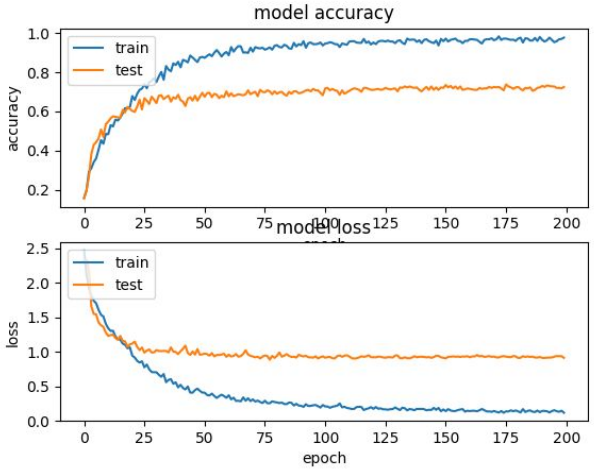
(a) Standard Neural Net

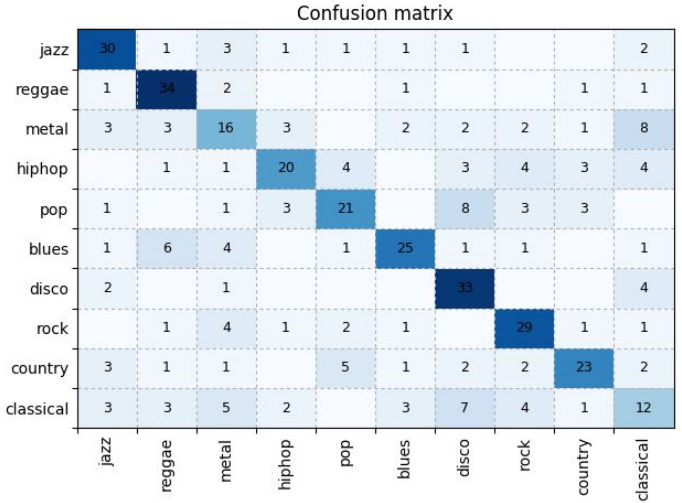
(b) After applying dropout.

Results:

-We trained all the net since the weights for the feature extractor CNN are not available anymore.

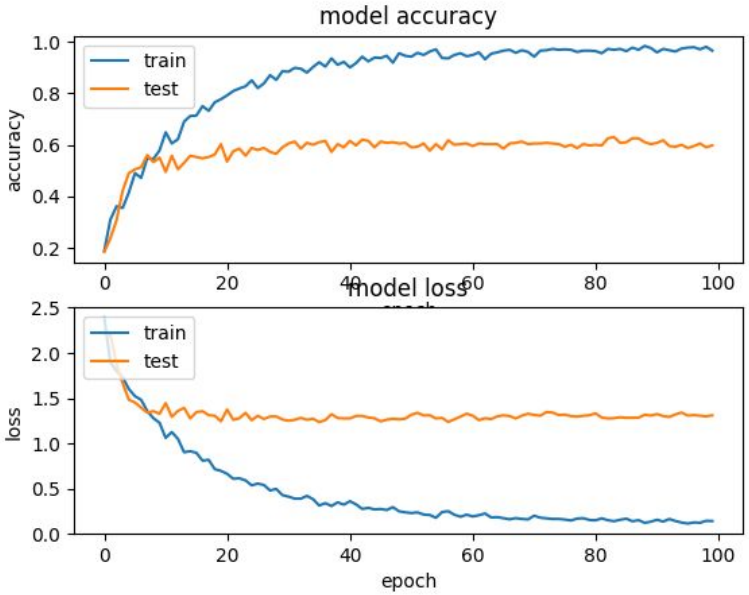
Experiment 1: 26 mfccs/frame,

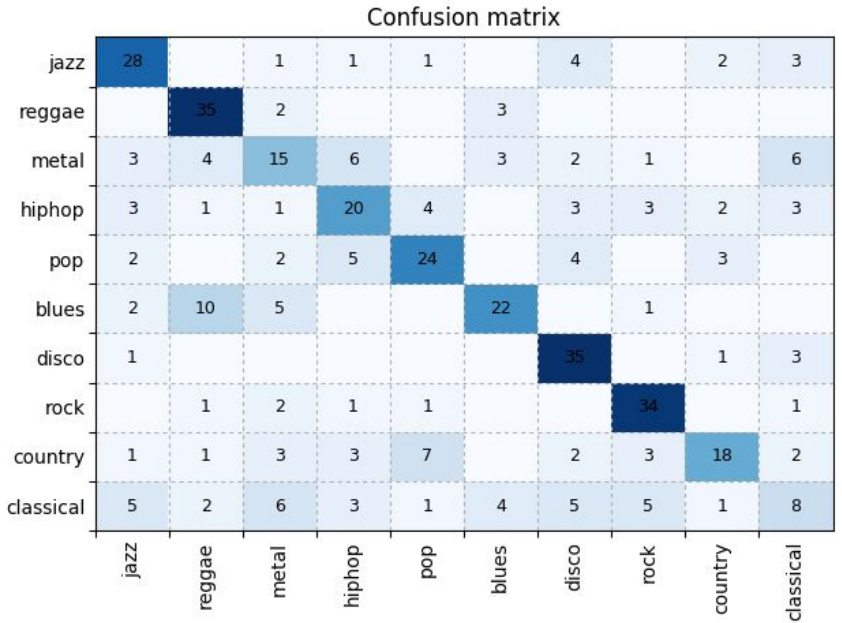




Results II

Experiment 2: 18 mfccs/frame,





Convolutional Layers Outputs:

Every convolutional block extract different features.

Some examples provided by Piotr Kozakowski & Bartosz Michalak using DeepVis Toolbox

Conv Block 1	Rap
Conv Block 2	Jazz
Conv Block 3	Pop

Conclusions and discussion:

- Our model achieved 61% test accuracy, it may not appear impressive compared to the state-of-art in recognizing music genre on GTZAN using deep learning approach was 84% in 2013
- Taking into account the amount of data used (10 genres*100 songs each), the amount of time for the project and the simplicity of the model we think results are good enough for such a hard task.

References:

- MUSIC GENRE RECOGNITION WITH DEEP NEURAL NETWORKS by Albert Jimenez, Ferran José DSAP course, Master MET, ETSETB
- https://github.com/jameslyons/python_speech_features Jimefor MFCC coeficients extraction
- Automatic Tagging Using Deep Convolutional Neural Networks by Keunwoo Choi, George Fazekas, Mark Sandler
- http://deepsound.io/music_genre_recognition.html
- https://github.com/keunwoochoi/music-auto_tagging-keras/blob/master/music_tagger_cnn.py Initial network model (5 conv layers, more classes, more data)


what's your question huh?
mememaker.net