MUSIC GENRE CLASSIFICATION

Areeb Khan Shabih A20469525
Carmen Acero Vivas A20472656

ILLINOIS INSTITUTE OF TECHNOLOGY

MOTIVATION AND PROBLEM DESCRIPTION

- Music Information Retrieval
  - Music Generation
  - **Music Genre Classification**
  - Recommender Systems
  - Music Source Separation

Music Genre Classification:

- Purpose is to distinguish one genre from another
- Challenging task since boundaries between genres are ambiguous
- Requires to extract meaningful features from audio
- Useful for music streaming companies Spotify and iTunes

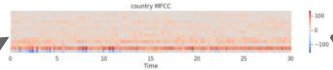
PROPOSED SOLUTION

Raw data

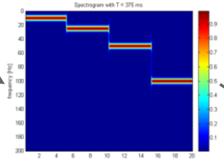
GTZAN Dataset
10 genres
100 audio clips per genre
30 seconds per audio clip

FEATURES EXTRACTION

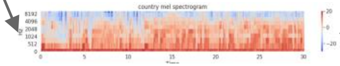
Chroma features



MFCC



STFT



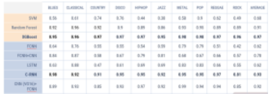
Mel spectrogram

MODELS

SHADOW LEARNING
- SVM
- Random Forest
- XGBoost

DEEP LEARNING
- FCNN
- FCNN + CNN
- LSTM
- C-RNN
- CNN (VGG16) + FCNN

MODEL EVALUATION AND RESULTS



DEMO

DATA PRE-PROCESSING

HYPERPARAMETER SELECTION

SAMPLE RATE

Defines the number of discrete data instances used per second to represent the analog sound in digital form. This number is set to **22050** after literature research.

WINDOW SIZE AND SAMPLE SIZE

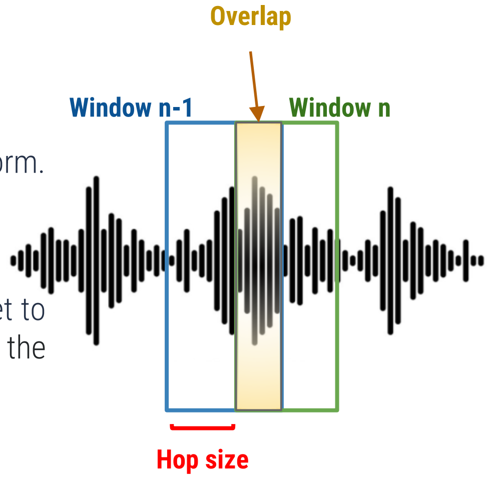
In order to do fourier transform, a window size and number of samples has to be defined. Windows size is set to **0.1** and number of samples size is set to **2048**. Optimal window size is the one where is expected that the properties of the signal chunk do not vary too fast.

OVERLAP

Defines how much the window size overlaps with each others. Higher the overlap, better information, but higher computation cost.. Set to **50%**

HOP SIZE

Amount the samples we shift to the right each time we take a new sample. Hop size to **512.**



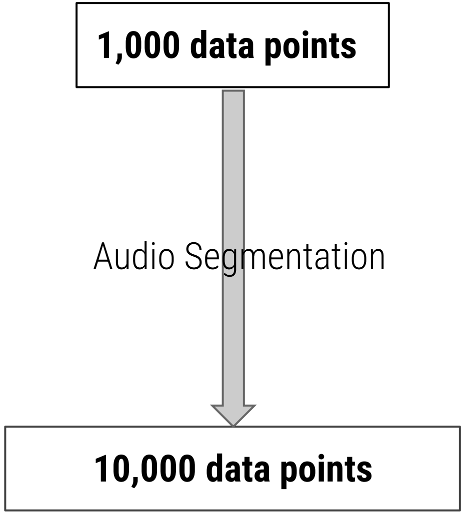
DATA PRE-PROCESSING

Data Augmentation

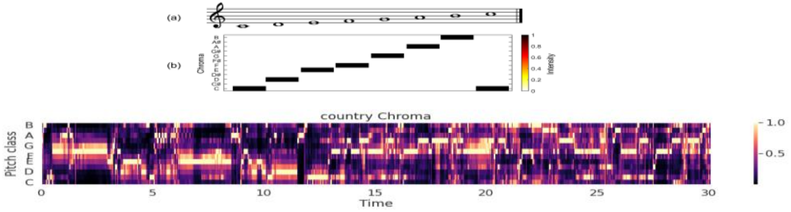
AUDIO CLIP SEGMENTATION

Each audio clip has a duration of 30 seconds. We divide each clip into 10 segments, of 3 seconds.

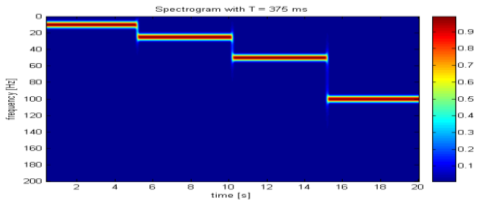
Advantage: Reduce the dimension of audio in time domain and increase the training dataset.



FEATURE ENGINEERING



Chroma features

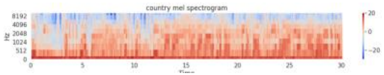


STFT Short time Fourier transform

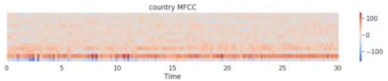
- Imparts information about pitch trend of the music signal
- Represent the **12 semitones** (pitch classes) versus time.
- We observe 12 boxes stacked on top of each other, where the color intensity represent the contribution

- Sequence of fourier transforms on a windowed signal
- Provides time localized frequency information
- Captures frequency information when frequency components of a signal vary with time

FEATURE ENGINEERING



Mel spectrogram



MFCC

- Representation of the spectrogram in the Mel scale

.

- A **spectrogram** is the representation of frequencies vs time

- The **mel scale** is a non linear transformation, and scales the frequency to match what humans can hear

- DNN Learns complex representations from the images

- **Mel-frequency cepstrum**, take the logarithm of Mel's frequency and then discrete cosine transformation

.

- Compresses the bands of Mel's spectrogram to 12-13 MFCC coefficients

- Coefficients are uncorrelated and work well with linear models

MODELS

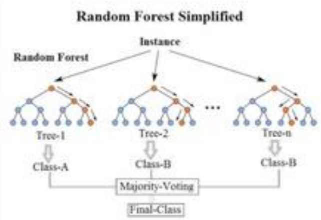
MODELS

SHADOW LEARNING
- SVM
- Random Forest
- XGBoost

DEEP LEARNING
- FCNN
- FCNN + CNN
- LSTM
- C-RNN
- CNN (VGG16) + FCNN

Support Vector Machine classifies by constructing a hyper-plane based on support vectors (extreme values of each class). Able to learn non-linear decision boundaries using **Kernel trick**.

Random Forest classifies by building independent decision trees. Each decision tree is trained and the prediction is made by an average voting.



XGBoost is a decision-tree-based gradient boost algorithm that outperforms random forest. Unlike Random Forest, the successive decision trees built are not independent and each tree learns from the errors of previous tree. Incorporates pruning and regularization in cost function to avoid over-fitting.

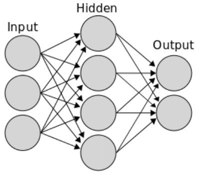
MODELS

MODELS

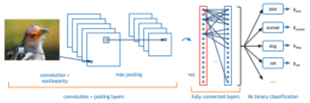
SHADOW LEARNING
- SVM
- Random Forest
- XGBoost

DEEP LEARNING
- FCNN
- FCNN + CNN
- LSTM
- C-RNN
- CNN (VGG16) + FCNN

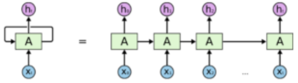
Fully Connected Neural Networks are a type of artificial neural network where all the nodes/neurons in one layer are connected to the neurons in the next layer



Convolutional Neural Networks is a class of deep neural networks, commonly applied to image recognition. The neurons represent 'template' which is applied to the image and it creates a feature map that summarize the presence of that feature. **VGG16** is a CNN proposed in 2015, highly popular and widely used.



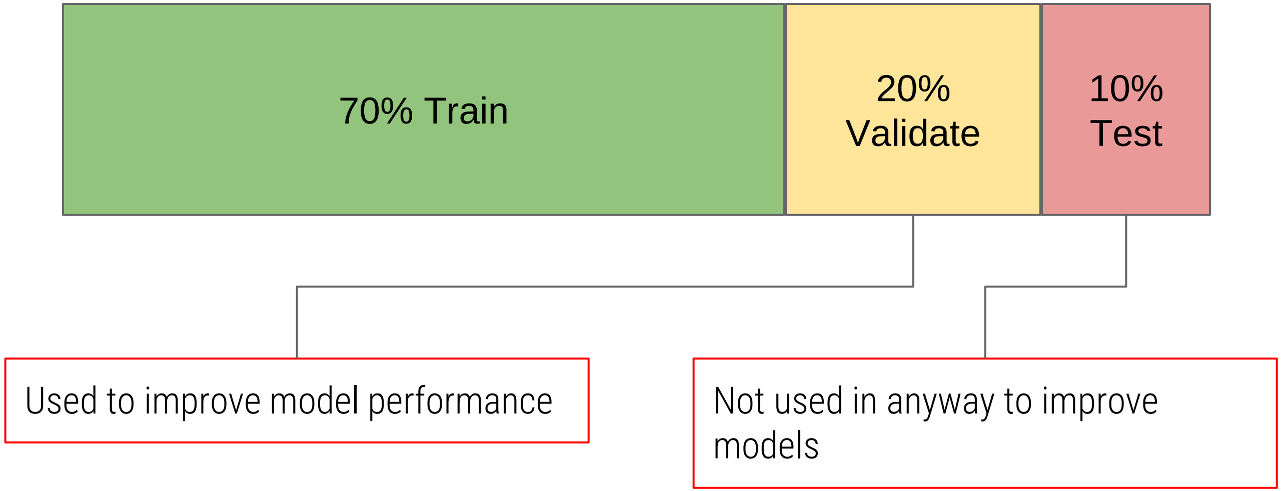
Long-Short Term Memory is an artificial recurrent neural network. This type of architecture are well suited to classify and predict time series data, since they take into account time dependencies.



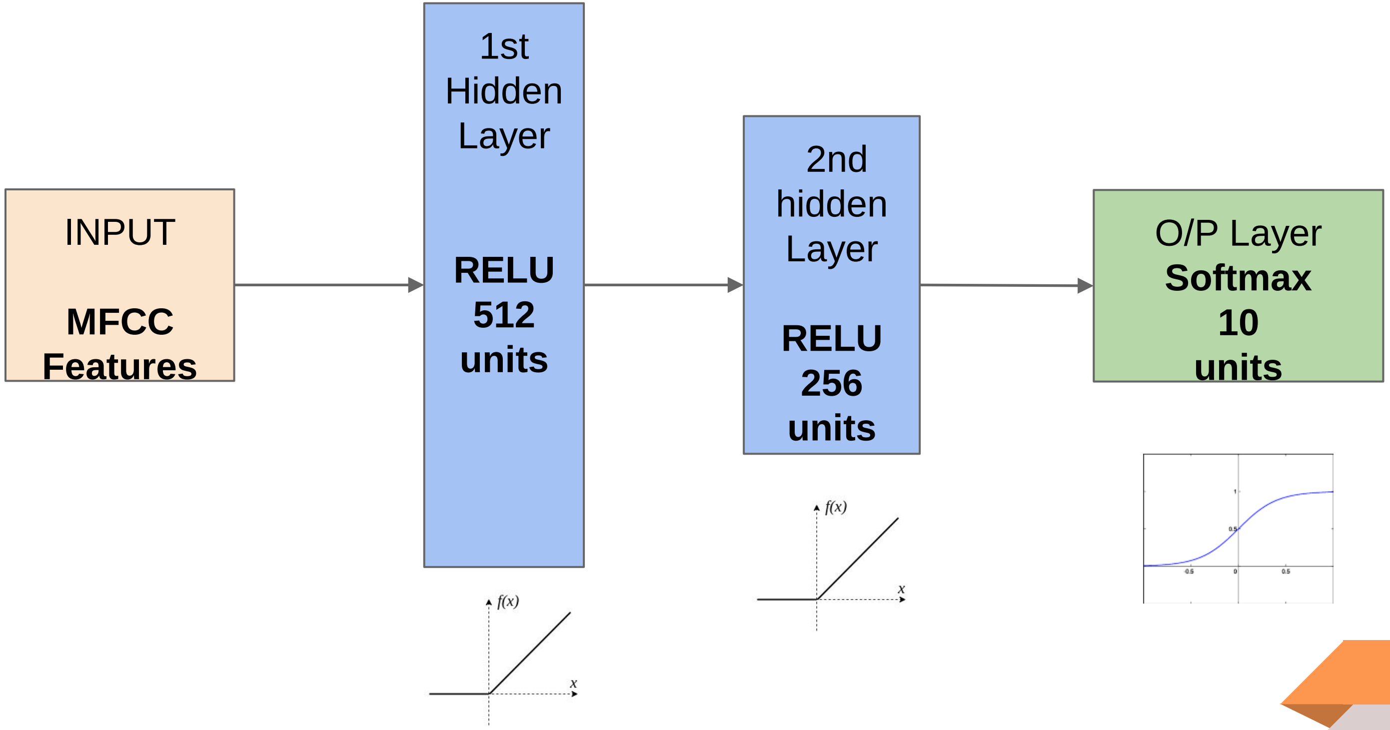
An unrolled recurrent neural network.

Convolutional Recurrent Neural Network involves a CNN followed by a RNN. It generates better results especially towards audio signal processing

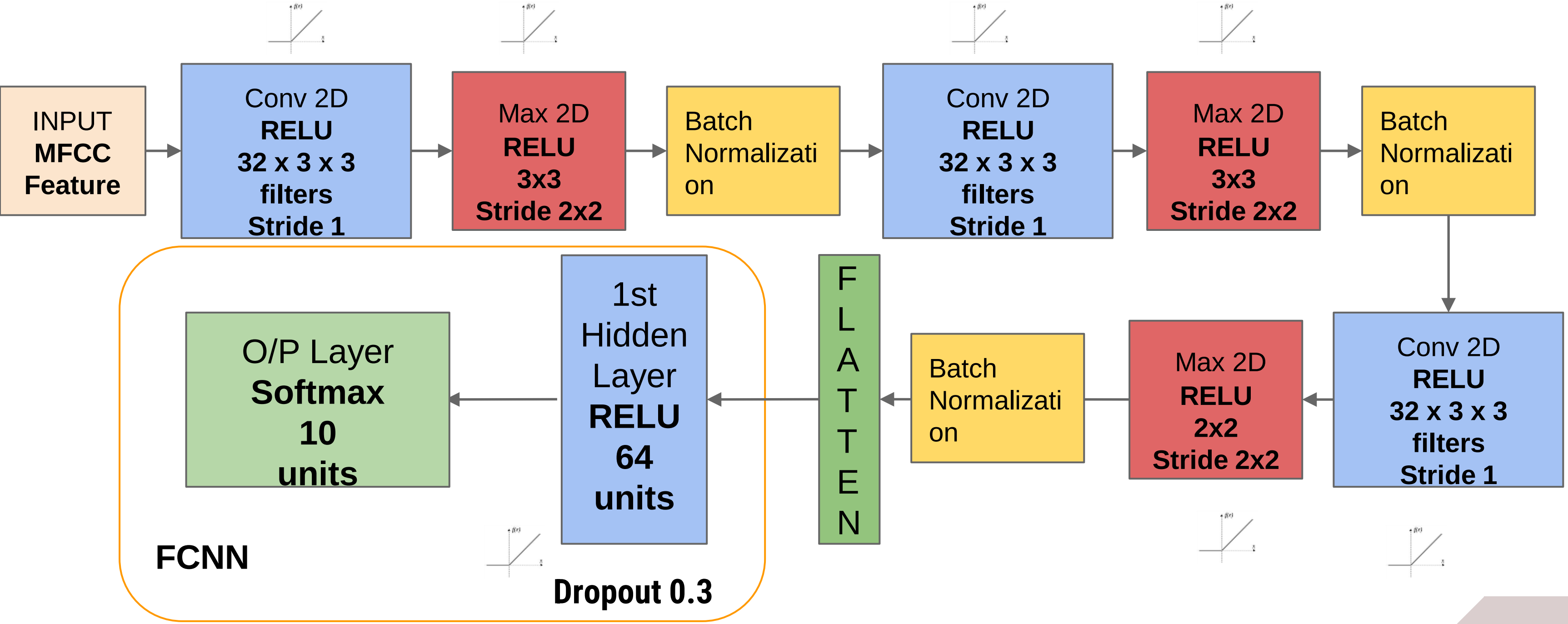
MODELS - Train, Validate and Test



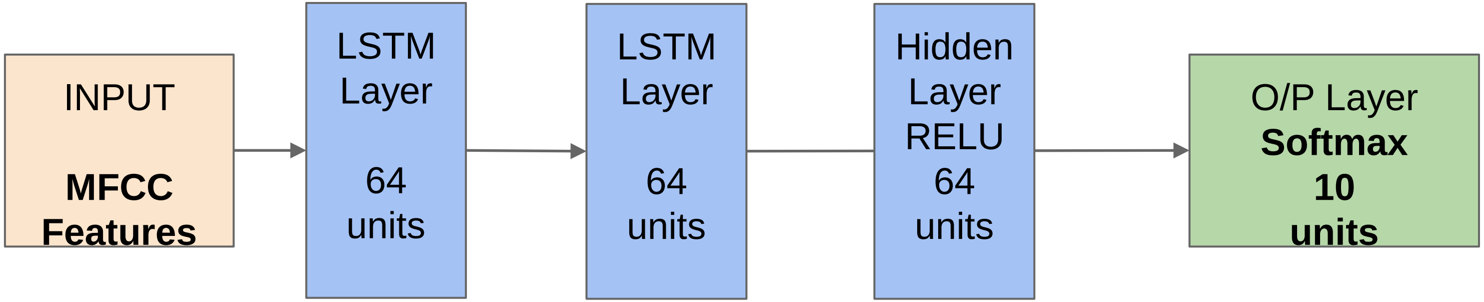
FCNN Architecture



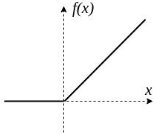
CNN+FCNN Architecture

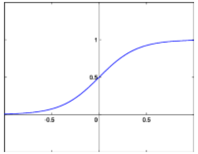


RNN Architecture

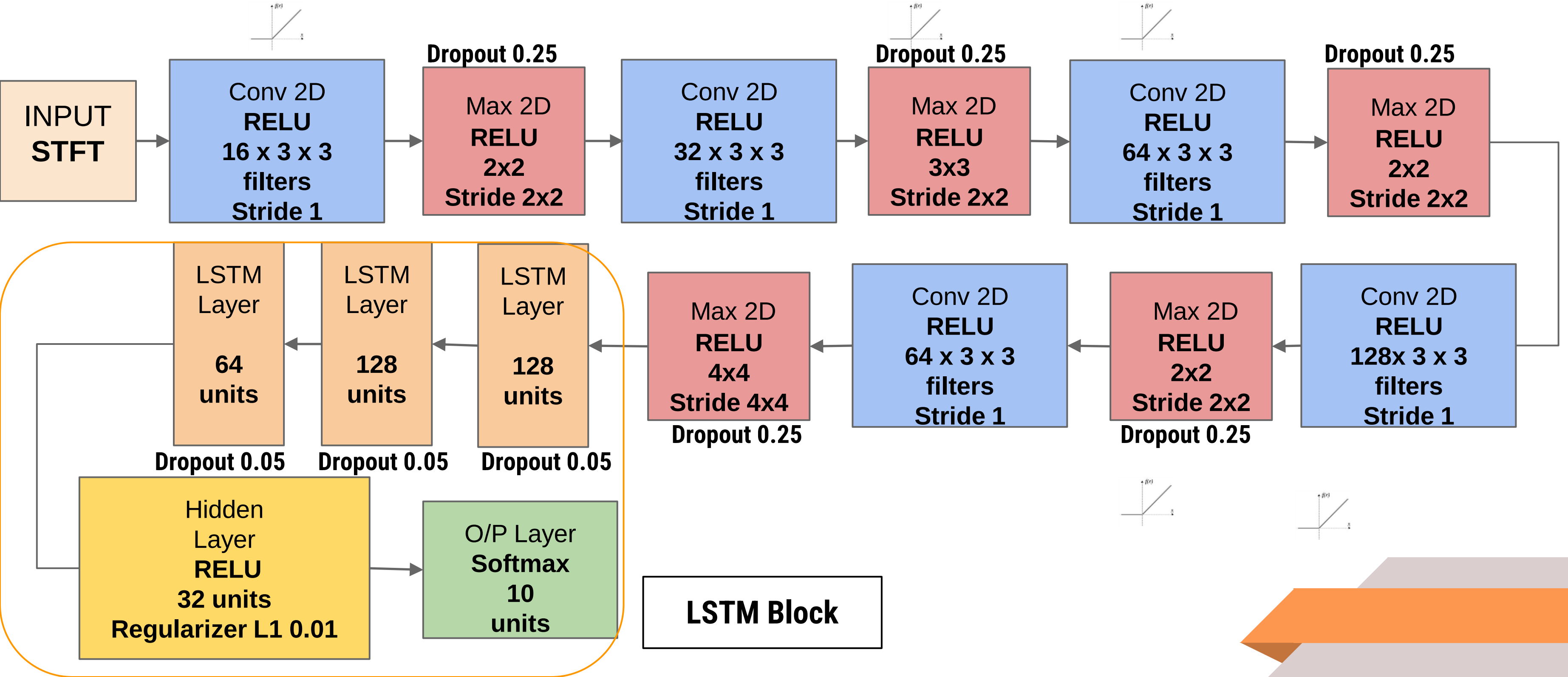


Dropout 0.3





CRNN Architecture



VGG 16 Architecture

INPUT

Mel's Spectrogram

→

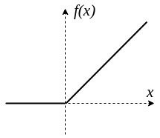
CNN VGG-16 Architecture

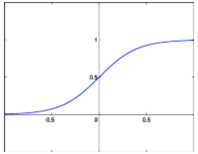
→

Hidden Layer
RELU
32
units

→

O/P Layer
Softmax
10
units





RESULTS: Accuracy & F1-scores

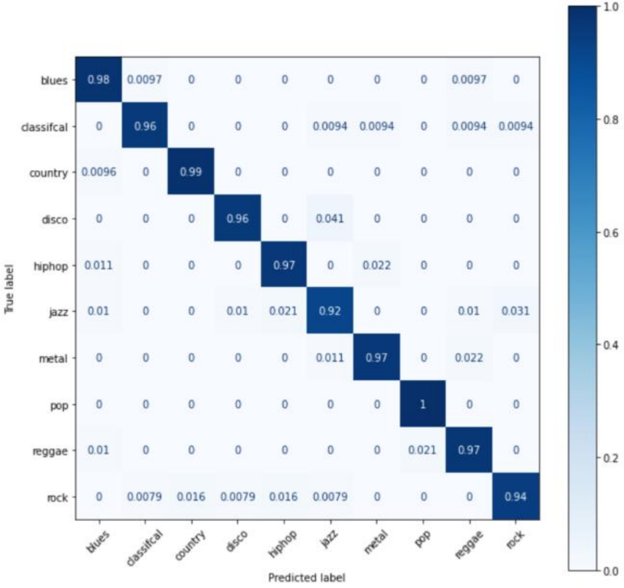
Algorithm	Input	Accuracy	F1-score
SVM	MFCC	0.60	0.60
Random Forest	MFCC	0.91	0.91
XGBoost	**MFCC**	**0.97**	**0.97**
FCNN	MFCC	0.62	0.61
FCNN+CNN	MFCC	0.71	0.70
LSTM	MFCC	0.62	0.61
C-RNN	**STFT**	**0.93**	**0.93**
CNN (VG16)+ FCNN	Mel’s Spectrogram	0.92	0.92

RESULTS: F1-scores

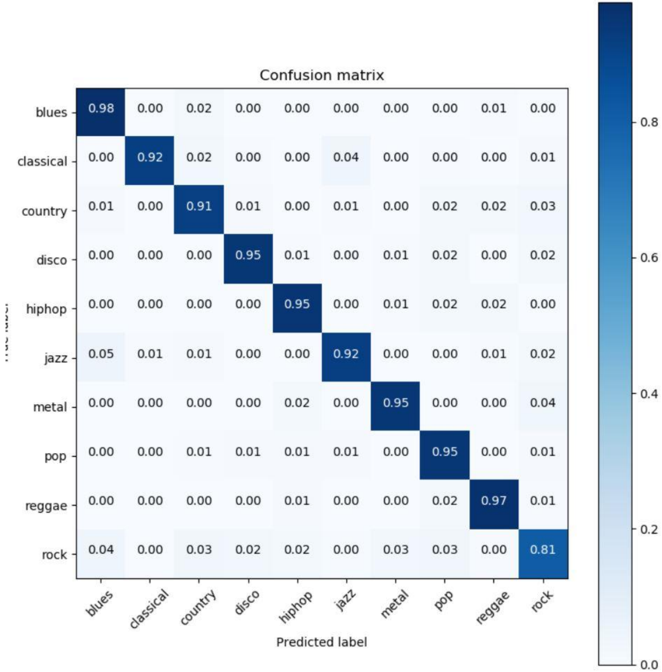
	BLUES	CLASSICAL	COUNTRY	DISCO	HIPHOP	JAZZ	METAL	POP	REGGAE	ROCK	AVERAGE
SVM	0.56	0.61	0.74	0.76	0.44	0.38	0.58	0.9	0.62	0.49	0.60
Random Forest	0.92	0.96	0.92	0.9	0.89	0.86	0.93	0.95	0.89	0.89	0.91
XGBoost	**0.95**	**0.96**	**0.97**	**0.97**	**0.97**	**0.95**	**0.98**	**0.98**	**0.97**	**0.96**	**0.97**
FCNN	0.64	0.76	0.55	0.55	0.54	0.59	0.79	0.79	0.51	0.42	0.62
FCNN+CNN	0.84	0.87	0.58	0.67	0.79	0.81	0.68	0.67	0.66	0.57	0.70
LSTM	0.63	0.88	0.47	0.61	0.69	0.69	0.83	0.83	0.66	0.55	0.62
C-RNN	**0.98**	**0.92**	**0.91**	**0.95**	**0.95**	**0.92**	**0.95**	**0.95**	**0.97**	**0.81**	**0.93**
CNN (VG16)+ FCNN	0.89	0.93	0.85	0.93	0.97	0.92	0.99	0.94	0.94	0.85	0.92

RESULTS: Winning Models

XGBoost



C-RNN



RESULTS: Why XGBoost?

XGBoost

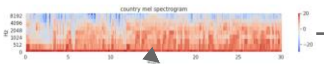
- Ensemble tree method
- Builds decision trees sequentially and not independently
- **Gradient Boosting** - Model learns from its mistakes and gives more weightage to wrong predictions
- Leverages patterns in residuals and make weak learners stronger
- Minimizes a regularized objective function to avoid overfitting
- Prune decision trees to avoid overfitting
- Extremely fast and optimized

RESULTS: Why C-RNN?

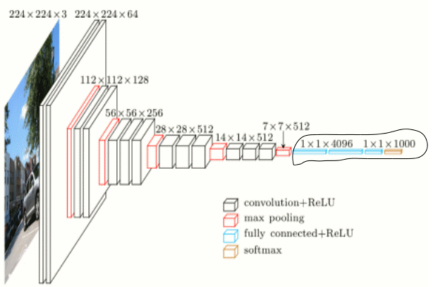
CRNN

- CNN learns complex representations/features from the image
- CNN sends the set of derived features to RNN
- RNN analyzes features in order, captures temporal information and discover important links between features

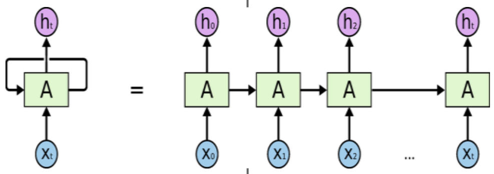
Image



CNN

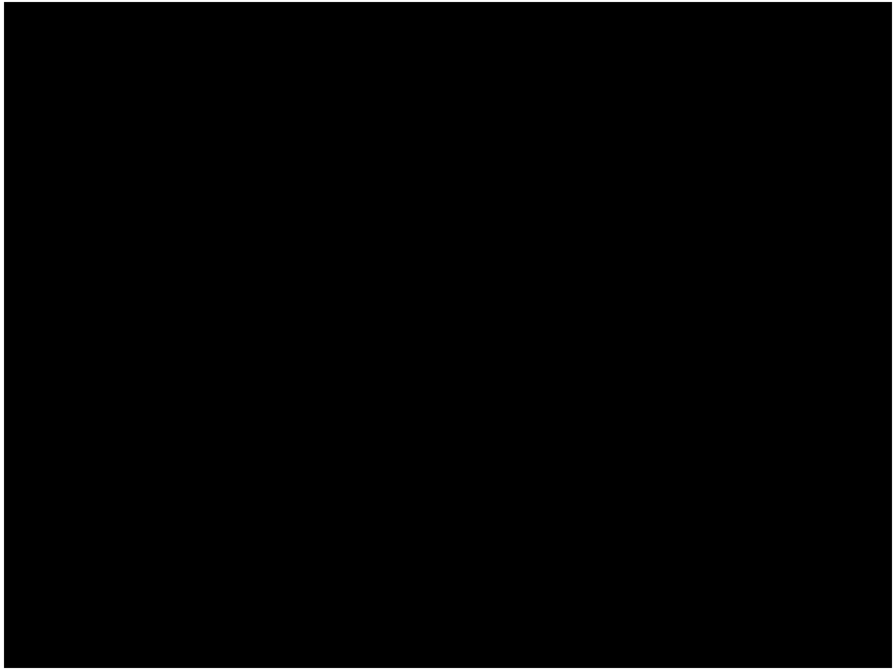


RNN



Output

DEMO



REFERENCES

PAPERS:

[1] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10:293 – 302, 08 2002. doi:10.1109/TSA.2002.800560.

[2] Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. CoRR, abs/1306.1461, 2013. URL http://arxiv.org/abs/1306.1461.

[3] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis, 2017.

[4] Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The million song dataset. pages 591–596, 01 2011.

[5] J. R. Castillo and M. J. Flores. Web-based music genre classification for timeline song visualization and analysis. IEEE Access, 9:18801–18816, 2021. doi:10.1109/ACCESS.2021.3053864.

[6] Chi Zhang, Yue Zhang, and Chen Chen. Songnet: Real-time music classification. Standford, 2018.

THANKS!