The University of Texas at Austin
Oden Institute for Computational Engineering and Sciences

MAY 4, 2022

SONIC SEPARABILITY

Applying Algorithms to Decipher a Song's Genre

Matt Goldberg, Ever Olivares, and Graham Pash
The University of Texas at Austin

Presentation Outline

- Goals
- Data Sources
- Classifier Methods
- Multimodal Modeling
- Transfer Learning
- Results and Concluding Remarks

Motivation + Goals

- Automatic music classification is a fundamental problem for
  - Music indexing
  - Content-based music retrieval
  - Music recommendation
  - Online music distribution
- Create models to classify music genres
- Explore types of music data used for prediction, such as audio signal

ABOUT OUR DATA

Data Sources

- GTZAN[1]
  - Source: kaggle.com
  - Genres: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock
  - 1000 WAV files
- **Free Music Archive (FMA)**[2]
  - Source: GitHub
  - Genres: Pop, Rock, Instrumental, Folk, Hiphop, Electronic, International, Experimental
  - 106574 MP3 files
  - 8000 track balanced subset
  - Echonest/Spotify provided feature vectors for a subset of full data

[1] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing 10.5 (2002): 293-302
[2] Defferrard, Michaël, et al. "Fma: A dataset for music analysis." arXiv preprint arXiv:1612.01840 (2016).

Data Extraction/Transformation/Load (ETL)

- Preprocessing of data is often overlooked but is crucial to a ML pipeline
- Processing of raw audio handled with librosa
- Scikit-learn & TensorFlow used for modeling
- Data comes in different formats (.wav, .mp3, vector, …)
- 30 second clips are chopped into 5 second intervals and processed

Level of abstraction of audio features

- High-Level:
  - Examples: instrumentation, keys, chords, melody, rhythm, tempo, lyrics, genre, mood…
- Mid-Level:
  - Examples: pitch and beat related descriptors such as note onsets, fluctuation patterns, MFCCs…
- Low-Level:
  - Examples: amplitude envelope, energy, spectral centroid, spectral flux, zero-crossing rate…
- Partially synonymous with David Marr's Tri-Level Hypothesis for Computer Vision [2]

[1] Knees, P, and Schedl, M. "Music similarity and retrieval: an introduction to audio and web based strategies" (2016)
[2] Marrs, D " Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information" (1982)

Mel Spectrogram (librosa)

- Time-domain to Time/Frequency domain
- A stack of FFTs
- From STFT:
  - Converts amplitudes from Hz to Db
  - Converts frequencies to Mel-Scale (log-scale)

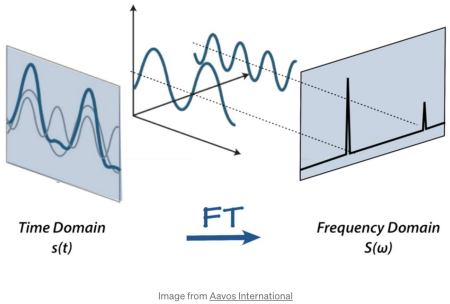
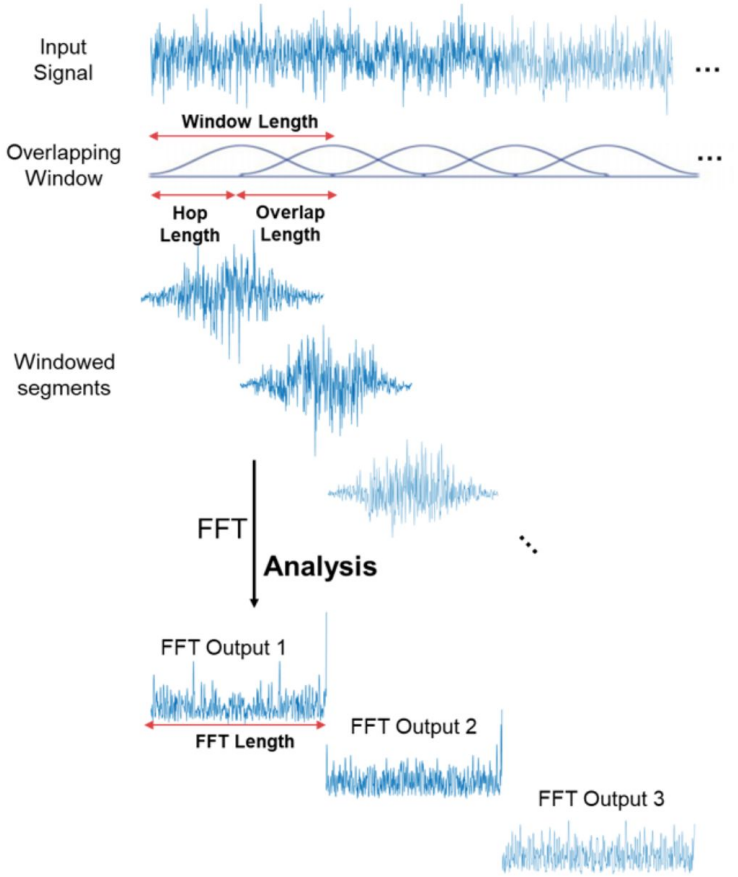


Image from Aavos International



[1] https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Discrete Fourier Transform -> Short Time Fourier Transform

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

Mel-Scale

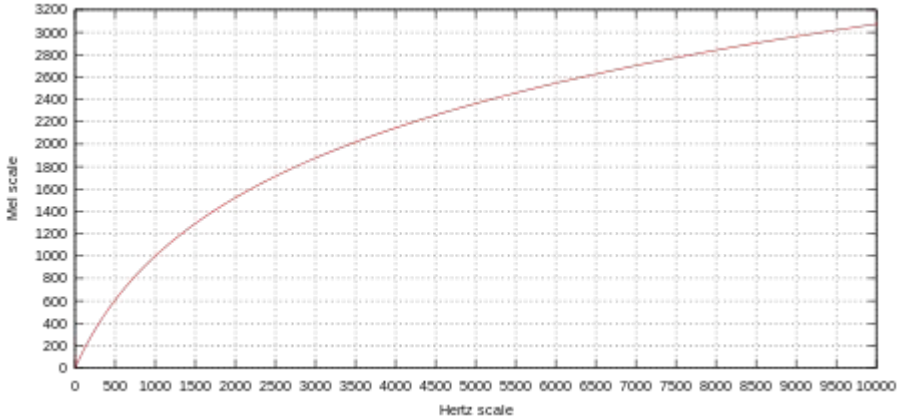
Mel-scale is a scale that relates the perceived frequency of a tone to the actual measured frequency.

$$\text{Mel}(f) = 1125 \cdot \ln(1 + \frac{f}{700})$$

- *It scales the frequency in order to match more closely what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies).*
- *This scale has been derived from sets of experiments on human subjects.*

Mel-Frequency Cepstral Coefficients (MFCCs) (librosa)

- Audio feature of choice for speech recognition / identification (1970s)
- Used in music processing (2000s)
- Believed to encode timbral information, since it represents short-duration musical textures [1].



[1] T. L. Li and A. B. Chan. Genre classification and the invariance of mfcc features to key and tempo. Volume Part I, MMM'11, pages 317–327, Berlin, Heidelberg, 2011. Springer-Verlag

Computing the cepstrum

Time Domain Signal

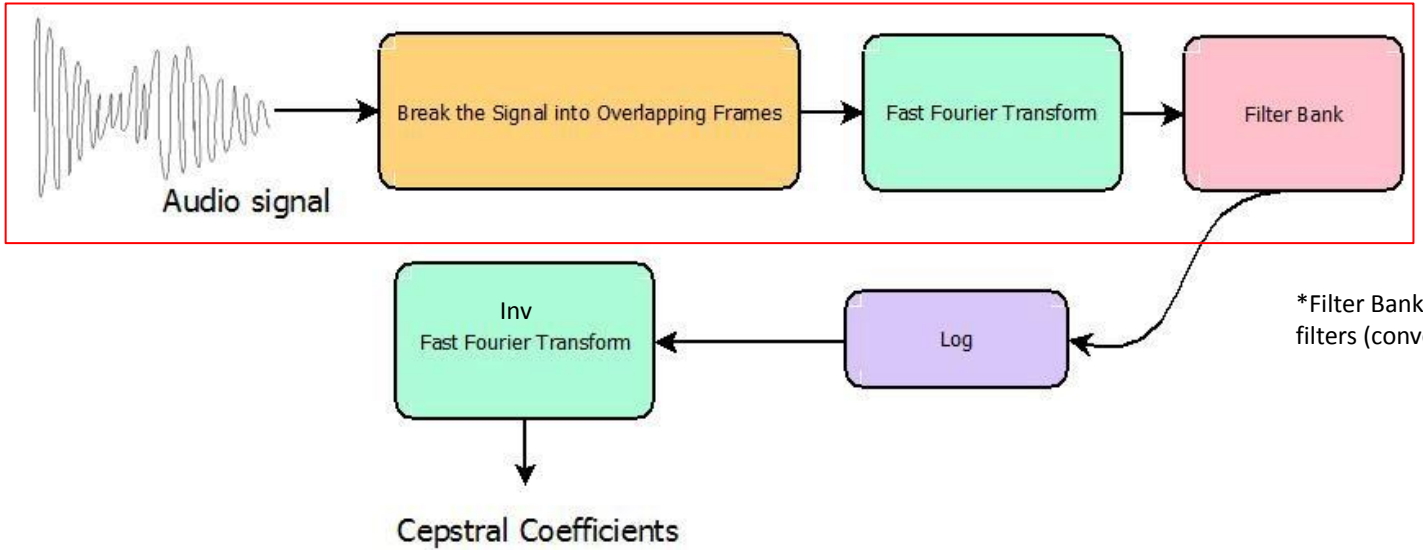
Cepstrum

Log Spectrum

Spectrum

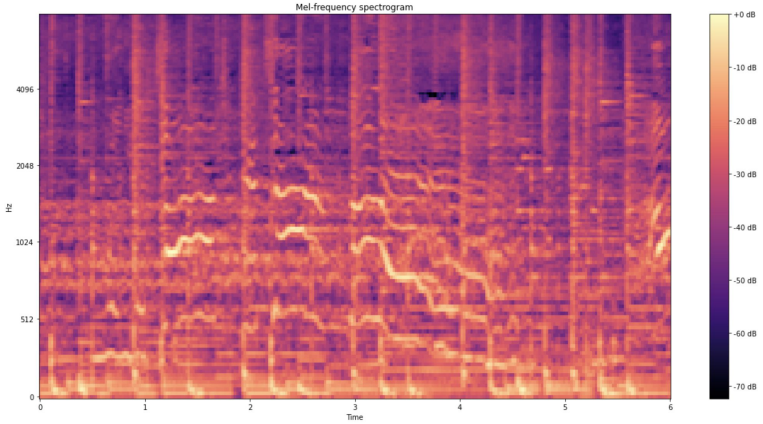
$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

Mel-Spectrogram to Mel-Frequency Cepstral Coefficient

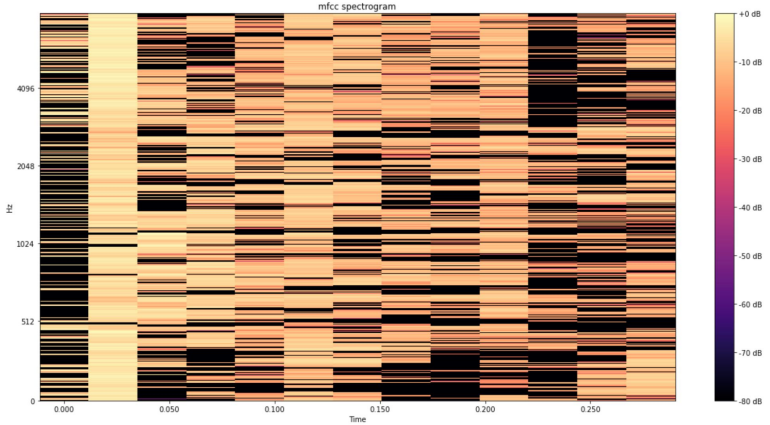


*Filter Bank refers to the mel filters (converting to Mel scale)

Mel-Spectrogram



MFCC



Visualizing high-dimensional data: T-SNE

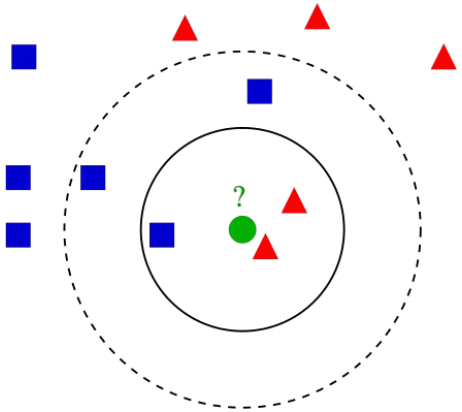
"Similarity is measured by the conditional probability of picking a datapoint to be a neighbor if picked in proportion to their probability under a Gaussian distribution centered at one point"[1]

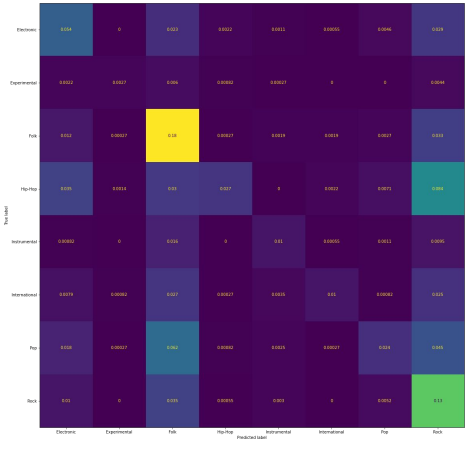
[1] van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE". Journal of Machine Learning Research. 9: 2579–2605.

BENCHMARKS

FMA Benchmarks

- **k-Nearest Neighbors**
- Support Vector Machine
- Random Forest
- XG Boost
- Multi-Layer Perceptron
- Convolutional Neural Network



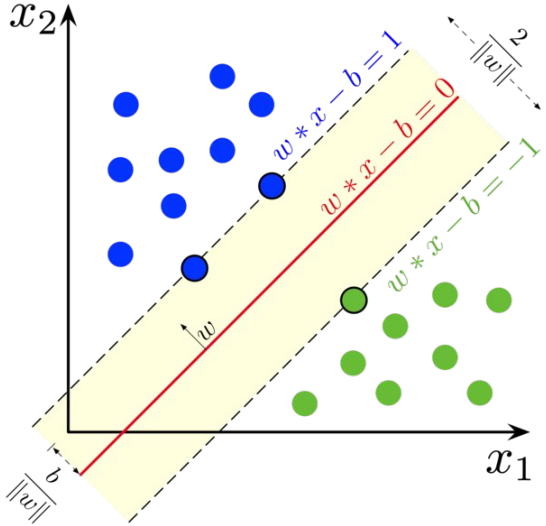


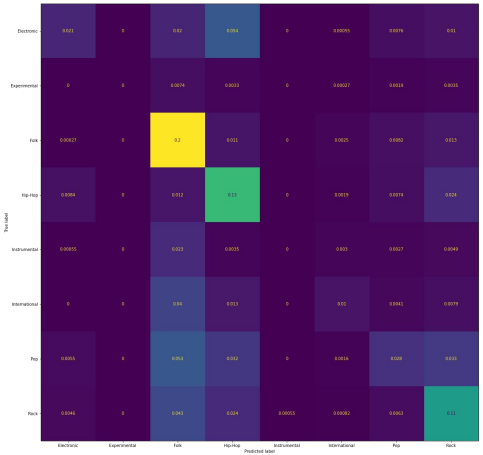
Hyper Parameter Tuning	
Parameter	**Range**
# Neighbors	3, 5, 7, 10, 15
Leaf Size	15, 30, 45

[1] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

FMA Benchmarks

- k-Nearest Neighbors
- **Support Vector Machine**
- Random Forest
- XG Boost
- Multi-Layer Perceptron
- Convolutional Neural Network



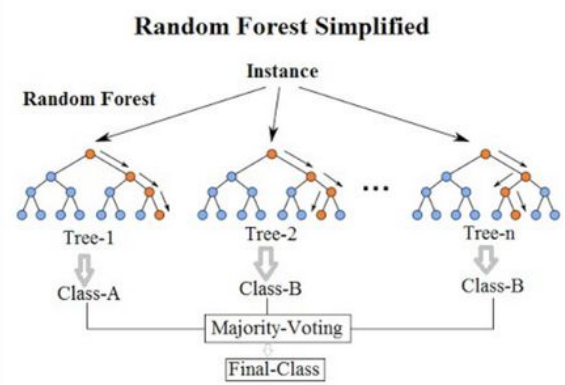


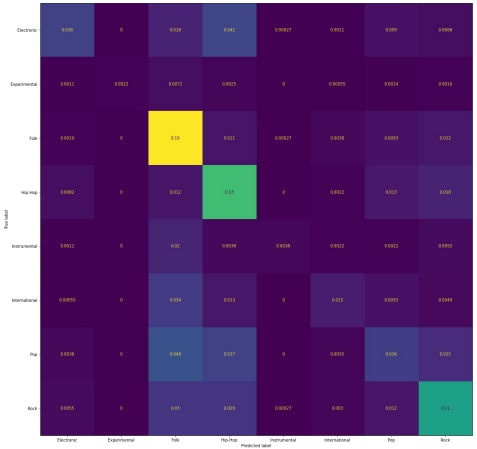
Hyper Parameter Tuning	
Parameter	**Range**
Kernel	Linear, Poly, RBF

[1] https://en.wikipedia.org/wiki/Support-vector_machine

FMA Benchmarks

- k-Nearest Neighbors
- Support Vector Machine
- **Random Forest**
- XG Boost
- Multi-Layer Perceptron
- Convolutional Neural Network



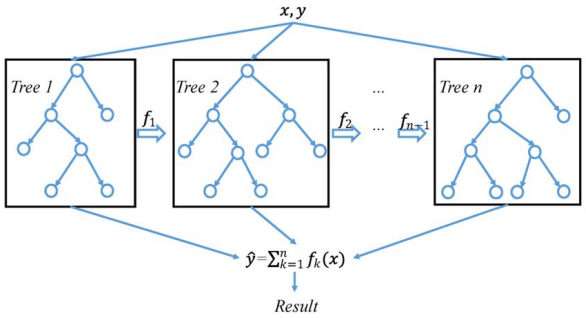


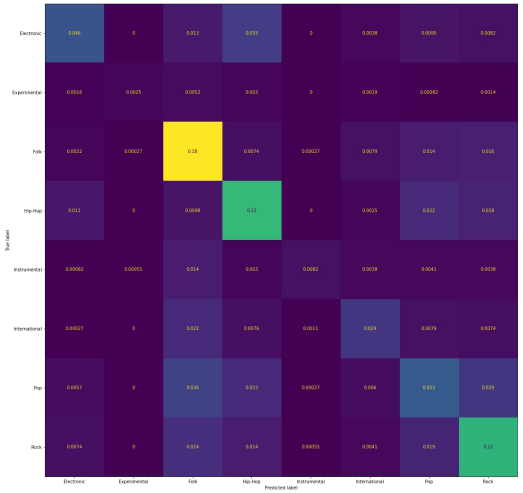
Hyper Parameter Tuning	
Parameter	**Range**
# Estimators	50, 100, 200
Max Depth	10, 30, 50, 70, None

[1] https://en.wikipedia.org/wiki/Random_forest

FMA Benchmarks

- k-Nearest Neighbors
- Support Vector Machine
- Random Forest
- **XG Boost**
- Multi-Layer Perceptron
- Convolutional Neural Network



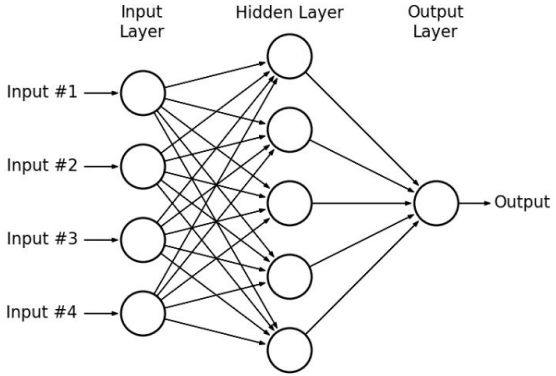


Hyper Parameter Tuning	
Parameter	**Range**
# Estimators	50, 100, 200
Max Depth	10, 30, 50, 70, None

[1] Wang, Yuanchao & Pan, Z. & Zheng, J. & Qian, L. & Mingtao, Li. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science. 364. 10.1007/s10509-019-3602-4.

FMA Benchmarks

- k-Nearest Neighbors
- Support Vector Machine
- Random Forest
- XG Boost
- **Multi-Layer Perceptron**
- Convolutional Neural Network



Model Training

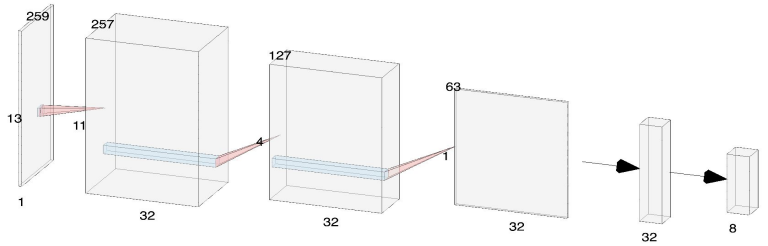
- Early Stopping
- Learning Rate Scheduler
- ADAM Optimizer[1]
- Sparse Categorical Entropy
- 100 Epochs
- Batch Size of 32
- 49/21/30 Train/Val/Test Splits

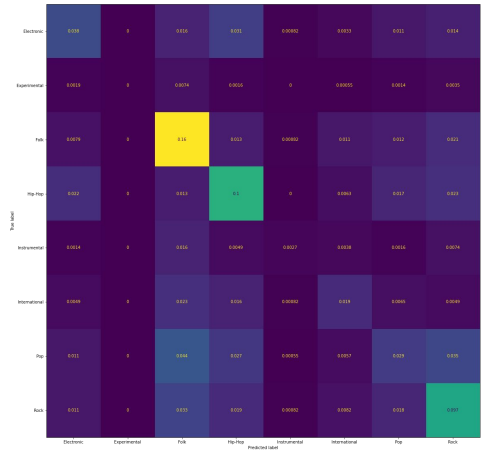
Model Architecture	
FC Layer	64, ReLU, L2 Regularization
FC Layer	32, ReLU, L2 Regularization
FC Layer	16, ReLU, L2 Regularization
Output	Softmax $\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$

[1] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

FMA Benchmarks

- k-Nearest Neighbors
- Support Vector Machine
- Random Forest
- XG Boost
- Multi-Layer Perceptron
- **Convolutional Neural Network**





Model Training

- Early Stopping
- Learning Rate Scheduler
- ADAM Optimizer[1]
- Sparse Categorical Entropy
- 30 Epochs
- Batch Size of 32
- 30% Dropout
- 2x2 Strides
- 49/21/30 Train/Val/Test Splits

[1] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

FMA Benchmarks

- k-Nearest Neighbors
- Support Vector Machine
- Random Forest
- XG Boost
- Multi-Layer Perceptron
- Convolutional Neural Network

Method	Accuracy
KNN	43.9%
SVM	49.4%
RF	52.7%
XGB	56.1%
CNN	**60.3%**

MULTIMODAL DATA

Spotify Engineered Features

Feature	Description	Values
Acousticness	Measure of whether the track is acoustic	0 to 1.0
Energy	Measure of intensity and activity - fast, loud, noisy	0 to 1.0
Instrumentalness	Whether or not a track has vocals	0 to 1.0
Loudness	Overall loudness of track	-60 to 0 dB
Danceability	Suitability of a track for dancing - tempo, rhythm stability, beat strength	0 to 1.0
Speechiness	Presence of spoken works	0 to 1.0
Valence	Musical positiveness conveyed by track	0 to 1.0
Liveness	Presence of an audience in the recording	0 to 1.0

[1] https://www.therecordindustry.io/spotify-audio-features/

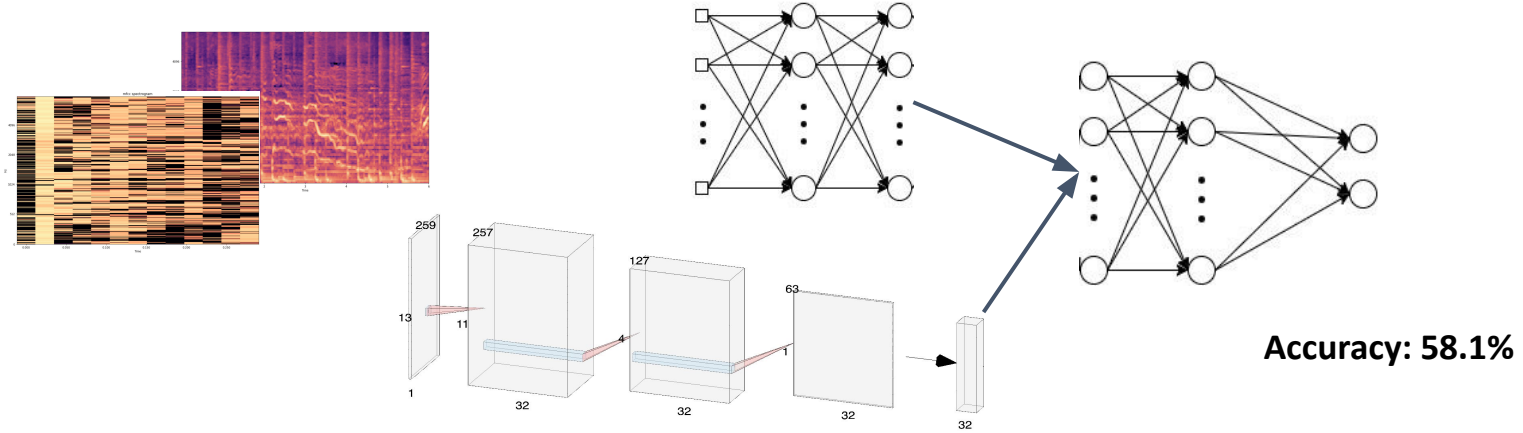
Benchmarks on Spotify Data

- k-Nearest Neighbors
- Support Vector Machine
- Random Forest
- XG Boost
- Multi-Layer Perceptron

Method	Accuracy
KNN	65.4%
SVM	27.5%
RF	**99.7%**
XGB	**99.7%**
MLP	23.0%

Multimodal Deep Learning Approaches

- Multimodal refers to different modalities of data: image, MFCC, vector, audio
- Would we see better performance by enriching our CNN?
- 8 Spotify/Echonest Features
- Concatenate features from CNN (MelSpec or MFCC) and Spotify



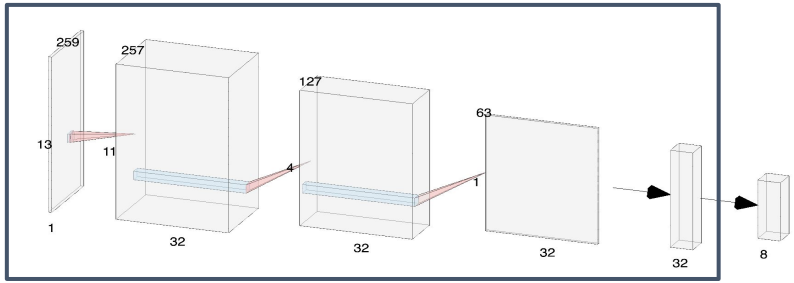
Accuracy: 58.1%

TRANSFER LEARNING

Transfer Learning from FMA to GTZAN

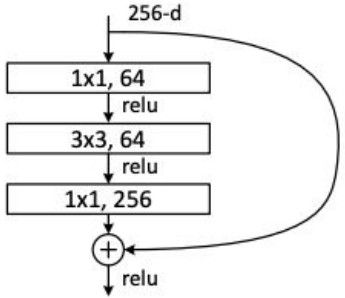
- Map GTZAN data to FMA dimensionality
- Freeze pretrained FMA model
- Replace last layer to predict GTZAN genres
- Retrain last layer **only** with GTZAN dataset
- Assess performance

Frozen Pretrained FMA CNN



Method	Accuracy (MFCC)	Accuracy (Mel)
KNN	36.7%	31.7%
SVM	52.9%	38.7%
Random Forest	55.4%	55.4%
XG Boost	58.5%	64.3%
MLP	51.3%	10.9%
CNN	**73.4%**	**65.5%**
FMA Transfer	38.8%	

Transfer Learning with ResNet50



- ResNet50[1]
- Can we treat spectrograms as images?
- Spectrogram data manipulation:
  - Resize: spectrograms need to be upsampled to ImageNet dimensions
  - Map intensities (in dB) to grayscale
  - Map grayscale to RGB $\in \{0, 255\}$
- Accuracy: 12.6%

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

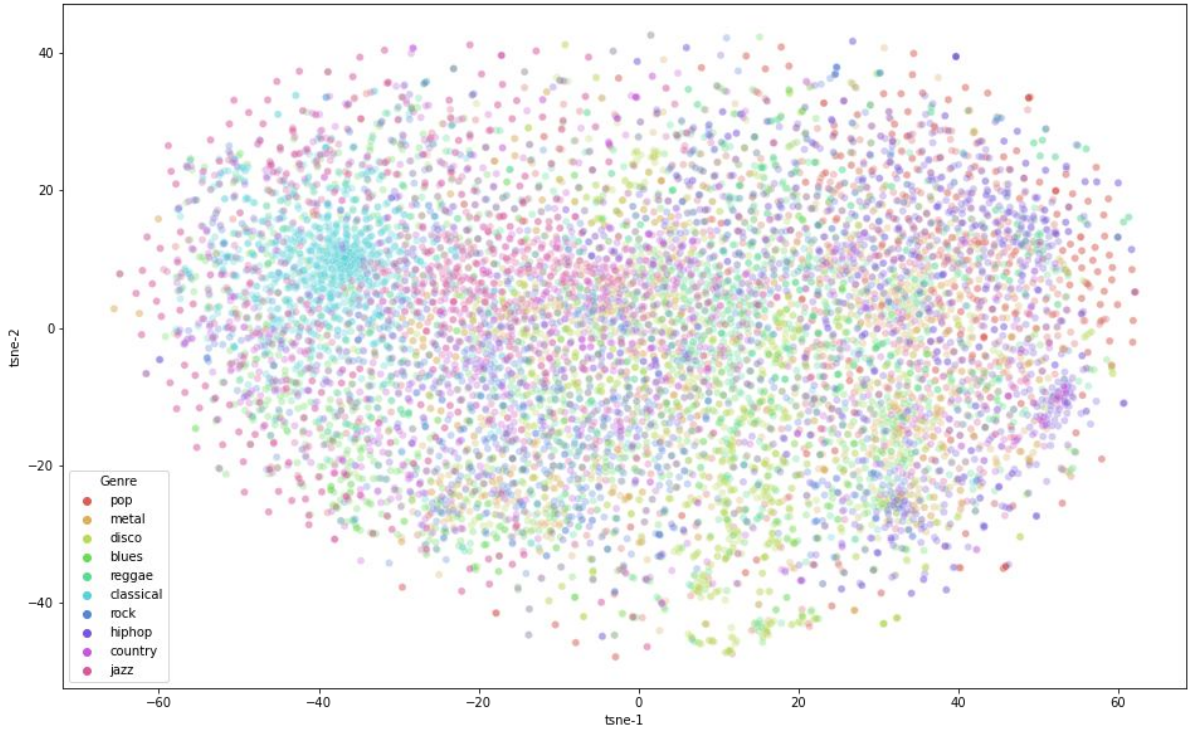
Conclusions + Future Work

- Using Free Music Archive data are we able to build classification models to determine what is the genre of a song
- Key takeaway: use models that are “aware” of your data structure
- Spotify’s derived features can be significantly more predictive than our engineered features (MecSpec, MFCCs)
- Fine-tuning the entire transferred network may increase performance
- Other data modalities may increase prediction accuracy
    - CNN applied to album artwork
    - Language models applied to lyrical data

QUESTIONS + DEMO

BACKUP SLIDES

GTZAN T-SNE


40
20
0
−20
−40
tsne-2
−60
−40
−20
0
20
40
60
tsne-1
Genre
pop
metal
disco
blues
reggae
classical
rock
hiphop
country
jazz