Latent Feature Extraction for Musical Genres from Raw Audio

Woody Wang
wwang153@stanford.edu

Arjun Sawhney
sawhneya@stanford.edu

Vrinda Vasavada
vrindav@stanford.edu

Motivation

While style is not well-defined for music, the genre of a piece of music is highly related to its acoustic properties. Current attempts at musical style encoding boast extensive feature engineering and static definitions of components of style. Learning encodings directly from raw audio instead has significant applications in musical style transfer and audio processing.

Task Definition: We seek to transform raw audio samples to genre encodings without explicit feature engineering.

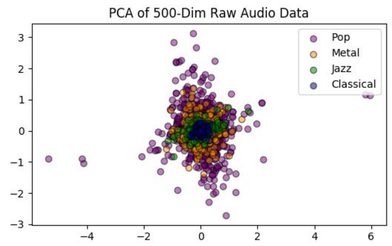
Dataset Information and Feature Engineering

GTZAN Dataset [1]

- 400 songs (30 seconds each) labeled as classical, jazz, metal, and pop

Inputs and Feature Engineering

- No explicit feature engineering, per task description
- Sampled one second clips at 22.05 kHz
- Downsampled to 500-dim inputs using average pooling



Model Infrastructures

Vanilla Autoencoder

Encoder: 3 hidden layers, learn $f(x) : \mathbb{R}^{500} \rightarrow \mathbb{R}^{64}$, where x is downsampled input
Decoder: 3 hidden layers, learn $g(x) : \mathbb{R}^{64} \rightarrow \mathbb{R}^{500}$, where x is encoder output
$\mathcal{L}_{reconstruction} = ||x - g(f(x))||_2^2$, where x is downsampled input

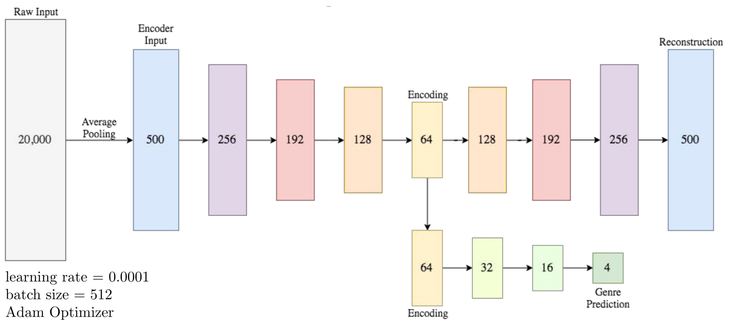
Two Layer Neural Network

Hidden layer: 128-dim, tanh activation
$\mathcal{L}_{cross-entropy} = -\sum_{i=0}^{3} y_i \ log(\hat{y}_i)$

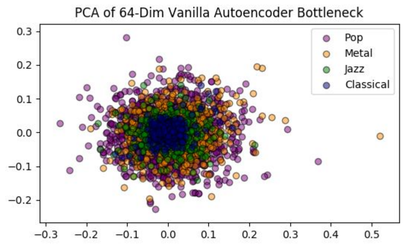
Deep Softmax Autoencoder (Final Architecture)

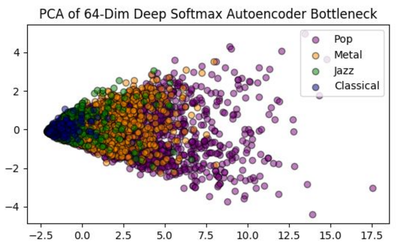
Simultaneously train a deep autoencoder and multi-class classifier using the 64-dim encoding as input to the classifier

$\mathcal{L} = \gamma||x - g(f(x))||_2^2 - (1-\gamma)\sum_{i=0}^{3} y_i \ log(\hat{y}_i)$, where reconstruction weight $\gamma = 0.9$

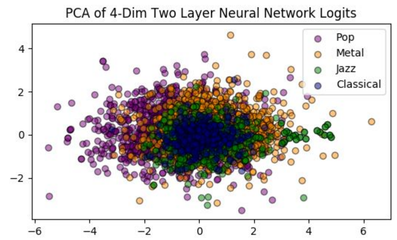


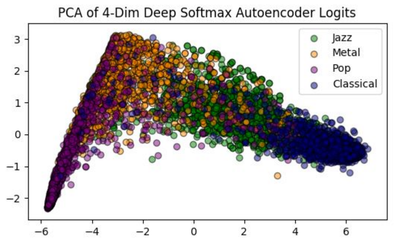
Qualitative Results





- The vanilla autoencoder results are as expected, since it is unsupervised and has no incentive to learn a distinguishable representation of genre
- When supervised with a genre classifier, the Deep Softmax Autoencoder's encodings display promising separation and smoothness in the latent space





- Motivated by neural style transfer on images, we experiment with visualizing the classifiers' logits as a form of genre encoding
- As expected, due to optimization objective, we see a clearer distinction between each class in the visualization of the classifiers' logits when accuracy is high

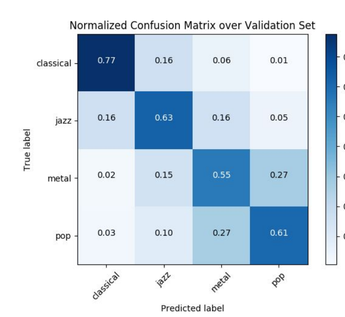
Summary:

- Compared to the 4-dim encodings, the 64-dim encodings have the potential to capture more variance within each genre
- The 4-dim encodings and 64-dim encodings serve different purposes: particular tasks may require the expressivity of the 64-dim or the conciseness of the 4-dim

Error Analysis

- We see a noticeable overlap between the pop class and the remaining three classes
  - When listening to random samples, we see that pop songs can easily be mistaken for the other three genres
- The pop genre encodings have the largest variance
  - This is corroborated by the variance of the pop songs in the raw data PCA
  - When listening to exclusively pop samples, there seems to be less of a distinct style within the genre
- Classical and jazz music have similar instrumentation, which might explain the proximity of their genre encodings

Quantitative Results (Final Architecture)



Classification Accuracies	Training Set (6000 Examples)	Development Set (2000 Examples)
Two Layer Neural Network	52.0%	38.1%
Deep Softmax Autoencoder	94.9%	64.1%

Deep Softmax Autoencoder	Precision	Recall	F1 Score
Classical	0.783	0.775	0.779
Jazz	0.606	0.627	0.616
Metal	0.515	0.554	0.5337
Pop	0.670	0.608	0.638

- Model has the most difficulty discerning between metal and pop
- Classical music exhibits the highest precision, recall, and F1 score, likely due to its distinct style

Future Work

- Replace the autoencoder with a β-TCVAE to learn disentangled representations via a mutual information gap (MIG) metric
- Increase number of classes in dataset to test generalizability of model
- Experiment with using learned latent representations as style encodings for music style transfer
- Interpolate components in the latent space to measure interpretability of latent representations

References

[1] G. Tzanetakis et al. Musical Genre Classification of Audio Signals in IEEE, 2002.
[2] S. Dai et al. Music Style Transfer: A Position Paper in arXiv, 2018.
[3] H. Bahuleyan. Music Genre Classification using Machine Learning Techniques in arXiv, 2018.
[4] I. Simon et al. Learning a Latent Space of Multitrack Measures in arXiv, 2018.