Music Genre Classification

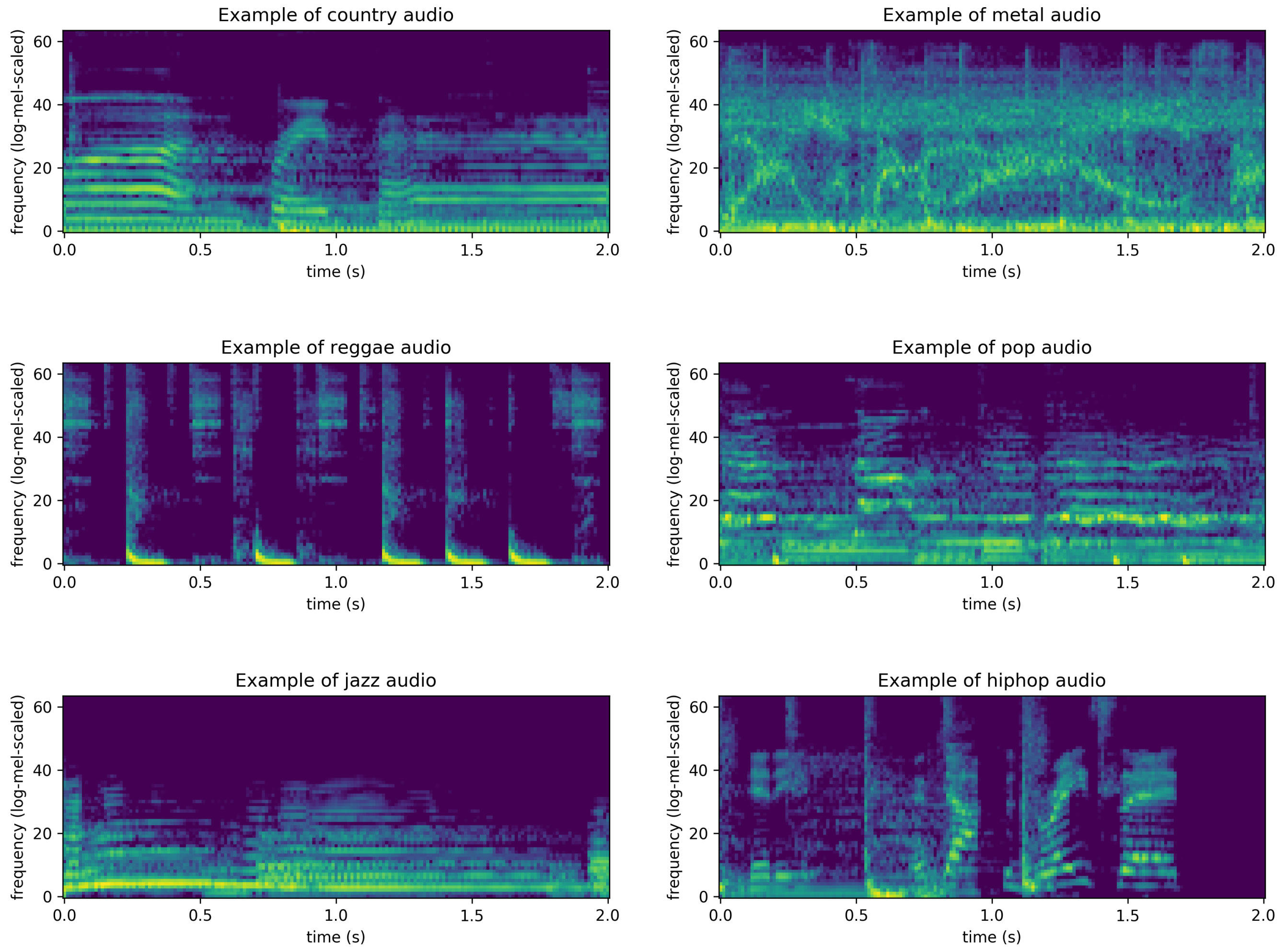
Derek Huang, Eli Pugh, Arianna Serafini
Stanford University

Data

We used the GTZAN genre collection dataset, which features 1000 samples of raw 30s data. However, since this raw audio was sampled at 22050HZ, we could reasonably use 2 seconds of data at most to keep our feature space relatively small (44100 features). To compromise, we augmented our data by randomly sampling four 2-second windows to produce 8000 samples. While this dataset has its flaws, its widespread use makes it easy to compare our work across the field.

Data Processing

- Initially ran our models on our raw audio data (amplitudes), which take the form of 44100 length arrays, but found that preliminary accuracy was lower than hoped for in all models.
- Decided to use mel-spectrograms, which are time vs. mel-scaled frequency graphs. Similar to short-time Fourier transform representations, but frequency bins are scaled non-linearly in order to more closely mirror how the human ear perceives sound.
- We chose 64 mel-bins and a window length of 512 samples with an overlap of 50% between windows. We then move to log-scaling based on previous academic success. Used the Librosa library – see examples below.

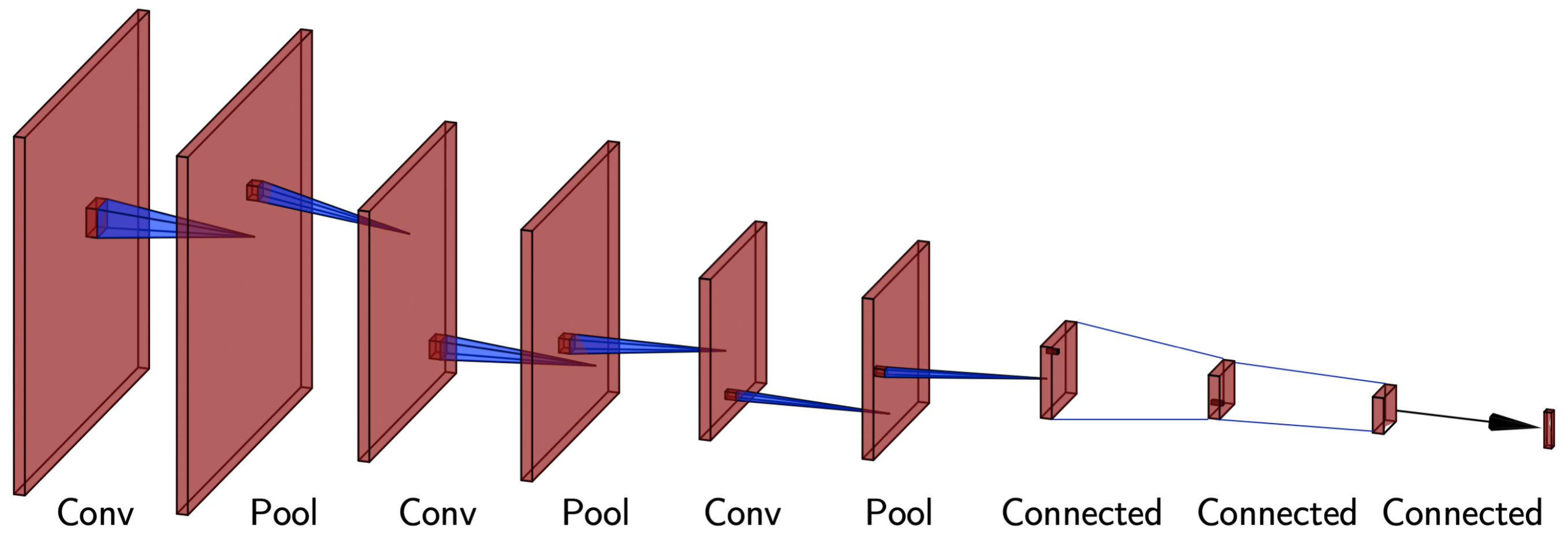


Motivation

Genre classification is an important task with many real world applications. As the quantity of music being released on a daily basis continues to sky-rocket, especially on internet platforms such as Soundcloud and Spotify, the need for accurate meta-data required for database management and search/storage purposes climbs in proportion. Being able to instantly classify songs in any given playlist or library by genre is an important functionality for any music streaming/purchasing service, and the capacity for statistical analysis that music labeling provides is essentially limitless.

Models

- **Support Vector Machine:**
  For the sake of computational efficiency, we first perform PCA on our data to reduce our feature space to 15 dimensions. Then we create an SVM model with an RBF kernel. This models offers us a baseline accuracy with which to compare our more complicated deep-learning models.
- **K-Nearest Neighbors:**
  We first perform PCA to reduce our feature space to 15 dimensions. We use $k = 10$ and distance weighting. Computation is deferred until prediction time.
- **Feed-forward Neural Network:**
  Our standard feed-forward neural network contains six fully-connected layers, each using ReLU activation. We use softmax output with cross-entropy loss, and Adam optimization.
- **Convolutional Neural Network:**
  As before, we use Adam optimization and ReLU activation. Structure is as illustrated below.



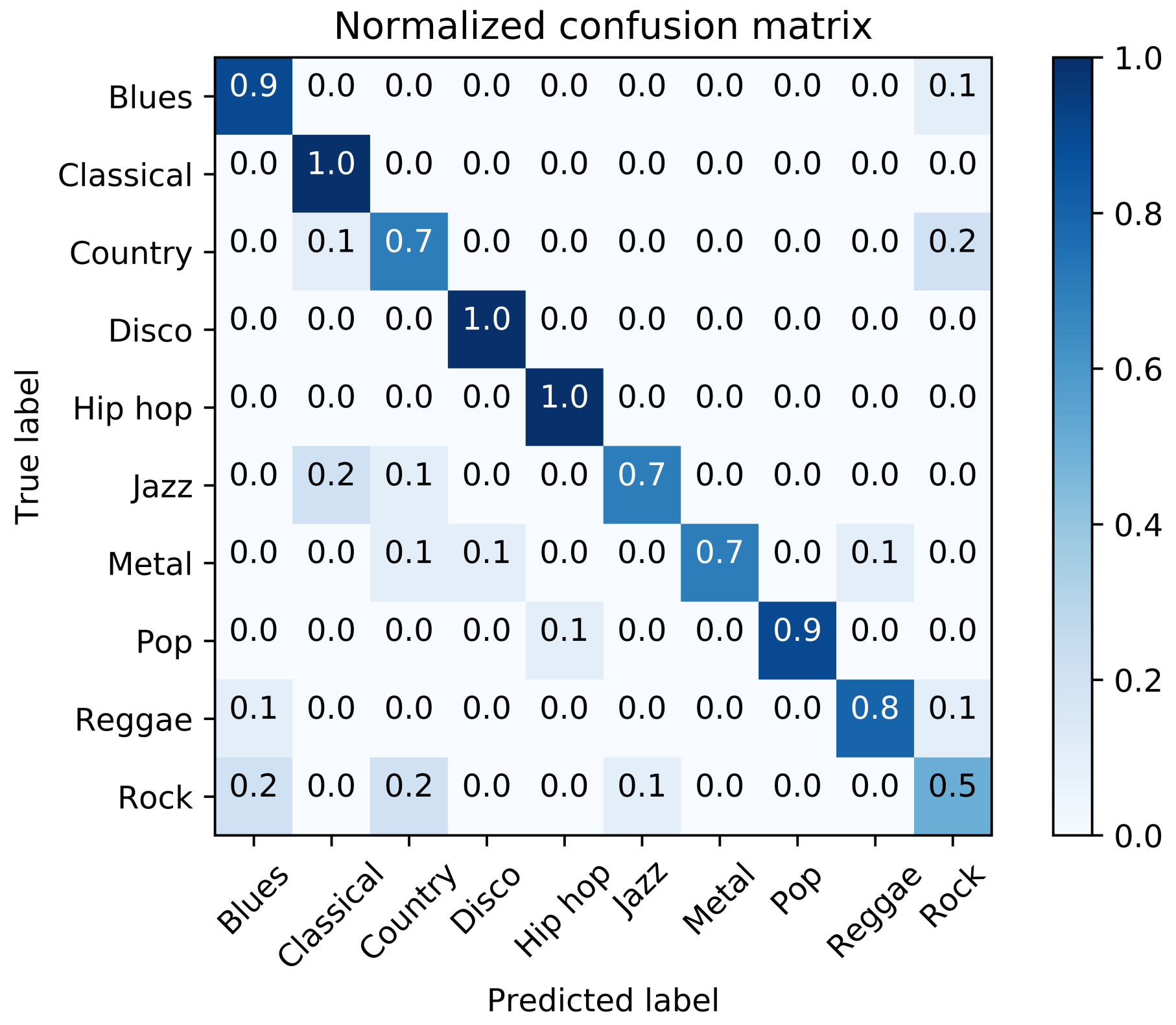
Convolutional layer:
$z_{k,l} = \sum_{j=1}^{n} \sum_{i=1}^{m} \theta_{i,j} \, x_{i+ks,j+ls}$

Loss function:
$CE = -\sum_{x \in X} y(x) \log \hat{y}(x)$

Results

The confusion matrix to the right visualizes results from our CNN.

Model Accuracy:	Train	CV	Test
Support Vector Machine	.97	.60	.60
K-Nearest Neighbors	1.00	.52	.54
Feed-forward Neural Network	.96	.55	.54
Convolution Neural Network	.95	.84	.82


Normalized confusion matrix

True label \ Predicted label	Blues	Classical	Country	Disco	Hip hop	Jazz	Metal	Pop	Reggae	Rock
Blues	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Classical	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Country	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.2
Disco	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Hip hop	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Jazz	0.0	0.2	0.1	0.0	0.0	0.7	0.0	0.0	0.0	0.0
Metal	0.0	0.0	0.1	0.1	0.0	0.0	0.7	0.0	0.1	0.0
Pop	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.9	0.0	0.0
Reggae	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.1
Rock	0.2	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.5

Discussion

For this project, we used traditional machine learning methods as well as more advanced deep learning methods. While the more complex models took far longer to train, they provided significantly more accuracy. In real world application, however, the cost/benefit of this tradeoff needs to be analyzed more closely.

We also noticed that log-transformed mel-spectrograms provided much better results than raw amplitude data. Whereas amplitude only provides information on intensity, or how "loud" a sound is, the frequency distribution over time provides information on the content of the sound. Additionally, mel-spectrograms are visual, and CNNs work better with pictures.

Future Work

While we are generally happy with the performance of our models, especially the CNN, there are always more models to test out – given that this is time series data, some sort of RNN model may work well (GRU, LSTM, for example). We are also curious about generative aspects of this project, including some sort of genre conversion (in the same vein as generative adversarial networks which repaint photos in the style of Van Gogh, but for specifically for music). Additionally, we suspect that we may have opportunities for transfer learning, for example in classifying music by artist or by decade.

References

Mingwen Dong. *Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification.* CoRR, Feb 2018, `http://arxiv.org/abs/1802.09697`

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

Piotr Kozakowski & Bartosz Michalak.*Music Genre Recognition.* Oct 2016, `http://deepsound.io/music_genre_recognition.html`

Fall 2018 CS229 Poster Session

Emails: huangda@stanford.edu, epugh@stanford.edu, aserafini@stanford.edu