# Classifying Album Genres by Album Artwork

Chris Koenig | koenig97@stanford.edu
CS230: Deep Learning

## Predicting

As vast digital music libraries have grown rapidly, proficient music genre classification algorithms have been developed to keep pace. However, little has been done to investigate the relationship between album artwork style and music genre. In this work, we utilize CNN frameworks to take album cover inputs and classify them by genre in single-label and multi-label approaches. We find that while the multi-label model surpasses the AUROC score of previous work, both models struggle to achieve high performance. This suggests the diverse range of album art in genres poses inherent challenges but also illuminates paths forward to overcome these challenges.

**224x224 album artwork image** → **Music genre(s) for album**

**Input** **Output**

## Data

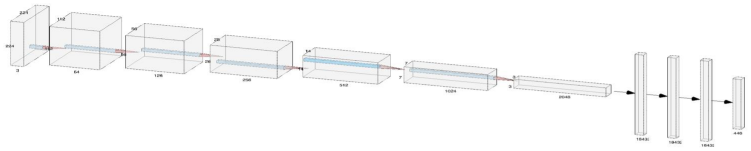
Our dataset is derived from the public Multimodal Music dataset [2] and consists of 31471 224x224 album images with multiple genre labels spread across 446 genres in a 4-level genre taxonomy. In the single-label case, the dataset is reduced to 18584 224x224 album images with single genre labels spread across 12 high-level genres (Folk, Jazz, etc.).

| Genre | % of albums | Genre | % of albums |
|---|---|---|---|
| Pop | 13.62 | Polynesian Music | 0.00052 |
| Rock | 8.57 | Elegies & Tombeau | 0.00052 |
| Alternative Rock | 4.29 | Fantasies | 0.00052 |
| World Music | 3.34 | Marches | 0.00052 |
| Dance & Electronic | 2.48 | Nocturnes | 0.00052 |
| Jazz | 2.36 | Madrigals | 0.00052 |
| R&B | 1.84 | Tangos | 0.00052 |
| Metal | 1.76 | Tierra Caliente | 0.00052 |
| Dance Pop | 1.74 | Islamic | 0.00052 |
| Indie & Lo-Fi | 1.47 | Armenia | 0.00052 |

## Features

We only use the raw pixel inputs of the images without any alteration other than simple data augmentation techniques to the single-label dataset by horizontally flipping the images with 50% probability and applying random brightness and saturation. These features are appropriate for this task because CNNs are highly proficient at learning directly from image input data as opposed to preprocessed image data via methods such as PCA and LDA. In terms of feature learning in the network, we do not implement transfer learning because we would like the network to make predictions based on the high-level style components of the album covers as opposed to specific objects or items in those images (such as a guitar, etc.), and past work in the field [2] has indicated that transfer-learning based models genre predictions are largely informed by the specific objects the model identifies in the images as opposed to broader image styles.

## Models



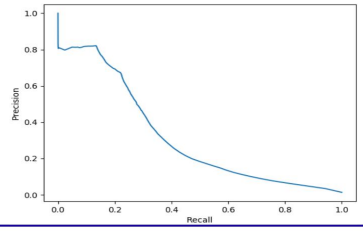
**ResNet-101 Baseline** - ResNet-101 pretrained on ImageNet and fine-tuned via transfer learning, published by Oramos et al.

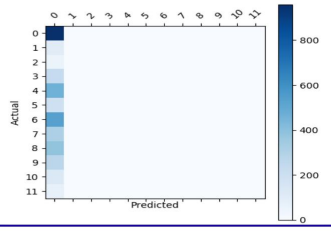
**Multi-Label Model** - CNN similar to VGG-16 with six convolutional blocks consisting of a 2D convolution, batch normalization, Relu activation, and max pooling. Each convolutional block utilizes filters of size 3 with stride 1 and same padding, and each max pooling layer has size and stride of 2. The convolutional blocks are followed by 3 fully connected layers, each applying batch normalization, Relu activation, and dropout. Finally is the output layer with one node for each of the 446 genre labels in the dataset. Because of the multi-label nature of the problem, we utilize the sigmoid cross entropy loss

**Single-Label Model** - Same architecture as above except consisting of five convolutional blocks instead of six and two fully connected layers instead of three. In addition, the final output layer is reduced to output 12 labels rather than 446. In addition, as each album is labeled with one genre, we replace sigmoid cross entropy with the softmax cross entropy loss.

## Results

The multi-label dataset was split according to a 80/10/10 train/dev/test breakdown, yielding a train set size of 25176 and test set size 3147. The single-label dataset was split according to a 60/20/20 breakdown. The mutli-label model achieved an AUROC score of 0.315 and AUPRC score of 0.899; the single-label model achieved an accuracy of 24.7%. The multi-label PR-curve and single-label confusion matrix are below.




## Discussion

From the Multi-Label results in the PR-curve, we see that although the model achieved a high AUROC score that surpassed the 0.743 AUROC score of the ResNet-101 baseline from [2], it was limited to a poor AUPRC score. Furthermore, across multiple prediction thresholds the model exhibits high precision but low recall. This seems to indicate that the model only makes predictions in which it is very confident, but its predictions are usually correct. Given the vast number of genres in the dataset - 446 - this seems reasonable, as there are a large number of highly specific genres with low supports that would be difficult for the model to learn well enough to make confident predictions.

From the confusion matrix we see that our single-label model struggled to gain traction on the task. Despite testing many architectures - from the very simple (2 convolutional layers and a single fully connected layer) to our final much more powerful model - the confusion matrix clearly demonstrates that the model reverts to a simple majority class classifier, predicting 'Alternative Rock' for all inputs. This means it most definitely overfit the training date despite a high dropout rate (0.5) and L2 regularization. The success of simple approaches such as that in [1] in which images are preprocess with PCA and LDA before kNN perhaps suggests that image-based genre classification is an inherently difficult task due to the diversity of album covers in genres and thus benefits from preprocessing that identifies the most critical features of the data.

## Future Work

1. Explore training a model to predict only the high-level genre (rock, rap, country, etc.) of an album, then develop separate, more specialized models to classify the more specific sub-genres.
2. Leverage public streaming service APIs to develop a large dataset exclusively for album cover image-based music genre classification.
3. Use a pretrained model to obtain style encodings for each image from the early layers of the network, and then classify these encodings rather than classifying the images directly.

## References

[1] Tyler Dammann and Kevin Haugh. Genre Classification of Spotify Songs using Lyrics, Audio Previews, and Album Artwork. 2017.

[2] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features. In International Society for Music Information Retrieval Conference, 2017.