CS230

Classifying Album Genres by Album Artwork

Christopher Koenig
Department of Computer Science
Stanford University
koenig97@stanford.edu

Abstract

As vast digital music libraries have grown rapidly over the past decade, highly proficient music genre classification algorithms have been developed to keep pace. While multi-modal models that incorporate album artwork images exist, little work has been done to explicitly investigate the relationship between album artwork style and music genre. In this work, we utilize CNN frameworks to take album artwork inputs and classify them by genre. We investigate both single-label and multi-label classification approaches and compare the results between the two schemes. Ultimately, we find that the CNN models struggle to gain traction in both approaches, suggesting the inherent challenges the diverse range of album art poses while also illuminating the way forward to overcome these challenges.

Code: github.com/koenig125/album-artwork-classification

1 Introduction

Music genres enable the categorization of musical items that share common characteristics. The problem of music genre classification has been thoroughly explored in the machine learning community. However, work in this field has focused primarily on textual and audial features for classification rather than visual features, namely the album artwork that accompanies music track releases. This has held true even as the improving ability and accessibility of deep learning methods have made incorporating visual features in machine learning models increasingly feasible. Thus while there is a wide body of literature and large public datasets available - the Million Song Dataset, for example - for music genre classification via textual and audial features, there is relatively little scholarship and resources focused on music classification via visual features such as album artwork or music videos. This project seeks to address that gap by exploring convolutional neural network architectures (CNNs) to conduct music genre classification utilizing solely album artwork.

The problem of music genre classification via album artwork is important not because of classification accuracy - highly effective music genre classification models have existed for some time. Rather, the problem is important because of the potential relationship that image-based classification models can reveal between album artwork and music genres. If music genre can be accurately classified from album artwork, this would indicate the presence of genre-level album artwork styles that reflect a mapping between the audial features of a genre's music and the visual features of that genre's artwork. This mapping would then provide a rich area for further exploration. For example, it would be prudent to investigate which visual features - tone, color, texture, the presence of particular objects, etc. - most powerfully influence the identity of a genre's artwork style. Furthermore, as features of visual art evoke emotion and meaning like any other art, the characteristics of a genre's album artwork style could provide insight into the emotional resonance that the music in a genre communicates.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Most existing music genre classification approaches are uni-modal and single-genre, relying either on direct audio sources such as audio recordings [2] or textual sources such as song lyrics [3, 11]. In recent years, deep learning CNN approaches have been developed for spectrograms, which provide a visual representation of the audio signal [4, 5, 7, 8, 15]. However, few approaches explore image-based classification tasks through album cover images, and some of the few that do use only traditional machine learning methods [10]. Furthermore, most multimodal approaches combine audio and lyric features [9, 13] as opposed to image features. As far as we are aware, there is only one published multimodal approach involving album cover images that integrates deep learning [14].

Oramos et al. [14] implement a multimodal approach by training three separate models on an album's audial, textual, and visual features derived from the MuMu (Multimodal Music) dataset, consisting of approximately 150k songs and 30k albums. Specifically, they utilize a ResNet-101 pretrained on ImageNet for the cover art classification component. They then concatenate the output from each model's final layer into a single feature vector, which is fed into a Multi Layer Perceptron for final classification. Their results indicate that multimodal models involving audial, textual, and visual features achieve best performance, confirming the utility of the cover images. In addition, the model trained on the cover images alone achieved a 0.743 auc score, which - while lower than the other modalities in isolation - still indicates the model learned to differentiate genres. In addition, Dammann and Hugh [6], in a CS229 final report, found that simple k-Nearest Neighbors on album artwork image data processed with PCA and LDA outperformed separate models trained on lyric and audio data. They conclude "perhaps the biggest takeaway from this project is that there is a huge possibility for more exploration of album artwork in regards to genre and other information".

3 Dataset and Features

3.1 Multi-Label Dataset

For this project we use the MuMu dataset - or Multimodal music - introduced by Oramos et al. [14], which was synthesized from the Amazon Reviews dataset [12] and the Million Song dataset [1]. MuMu contains audio tracks, text reviews, and album cover URLs for 31471 albums with multiple genre labels for each album. We leverage the images and genres from MuMu to form our dataset of 31471 300x300 album images with genre labels. There are a total of 446 genres represented in the dataset structured in a hierarchical taxonomy that is four levels deep. The first level is the most broad (Pop, Rock, etc.), and the specificity of the genres increases with each level. Albums are labeled with the full branch for their genre(s). For example, an album that is labeled as Traditional Pop also comes with Pop and Oldies genre labels as well. Overall, each album has an average of roughly 6 labels. However, the genres are highly imbalanced, with the genres in the first level of the taxonomy being much more common. Finally, no data preprocessing is applied to the images other than resizing to 224x224 resolution. As the dataset is not large, we split the dataset into a 80/10/10 train/dev/test split.

3.2 Single-Label Dataset

For the single-label task, we consolidated the multi-label tags by choosing a small subset of genres to be the single-label tags. The natural choice of genres were those from the first level of the music genre taxonomy. However, 'Pop' and 'Rock' were too broad and frequently co-occurred with every other label, and were thus eliminated. The 12 final labels were then 'Metal', 'Alternative Rock', 'Dance Electronic', 'Rap & Hip-Hop', 'RB', 'Jazz', 'Folk', 'Country', 'Latin Music', 'Reggae', 'Classical', and 'Christian'. Once chosen, we created a second dataset for the single-label task by filtering out albums tagged with more than one genre from the list above or none of the genres listed above. This resulted in a new dataset of 18584 images of 224x224 resolution, each with a single genre label. Because of the reduction in the number of images, we split the dataset into a 60/20/20 train/dev/test split to prevent overfitting and ensure high confidence in the overall performance of our system. Finally, we applied simple data augmentation techniques to the new dataset by horizontally flipping the image with 50% probability and applying random brightness and saturation.

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

Figure 1: Most/Least Represented Genres for Multi-Label

Genre	% of albums
Alternative Rock	25.35
Jazz	14.49
Dance & Electronic	12.55
Metal	10.36
Latin Music	7.74
R&B	7.05
Country	6.54
Folk	5.58
Rap & Hip-Hop	3.57
Christian	2.8
Reggae	2.34
Classical	1.64

Figure 2: Representation of Genres By Percent in Single-Label Dataset

4 Methods

4.1 Baseline

Because the work reported by Oramos et al. [14] is the only deep learning approach to music genre classification through cover images, we use their results as a baseline. Their ResNet-101 model achieved a macro-averaged 0.743 AUC ROC score using solely album cover images as input. We refer to this score throughout our analysis for the multi-label approach. However, Oramos et al. do not offer precision or recall statistics for the performance of their model despite the highly imbalanced nature of the dataset, so we have no baseline in this regard. Because the MuMu dataset is relatively new, the authors are not aware of image-based deep learning genre classification results for the single-label task. The single-label results reported in this work are thus the first of their kind.

4.2 Multi-Label Approach

For the multi-label approach, we leverage a fairly large 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 to predict class probabilities independently, defined as follows:

$$\mathcal{L}_{ML} = \sum_{i=1}^{C} y_i * -log(sigmoid(\hat{y}_i)) + (1 - y_i) * -log(1 - sigmoid(\hat{y}_i))$$

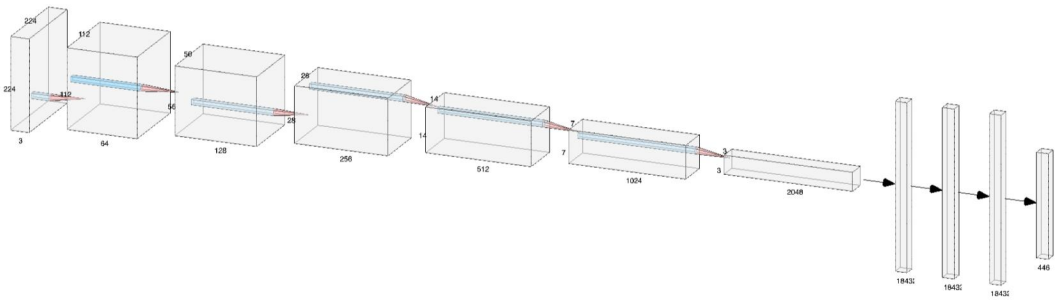

Figure 3: CNN Architecture for Multi-Label Classification

4.3 Single-Label Approach

For the single-label approach, we leverage a similar - but slightly smaller - architecture 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:

$$\mathcal{L}_{SL} = \sum_{i=1}^{C} y_i * -log(softmax(\hat{y_i})) + (1 - y_i) * -log(1 - softmax(\hat{y_i}))$$

5 Experiments/Results/Discussion

5.1 Metrics

For the multi-label task, we report the area under the ROC curve (AUROC) as a direct comparison to the baseline results from Oramos et al. [14]. In addition, because the dataset is highly imbalanced - both in terms of positives/negatives (ie every album is only labeled with 1-10 genres out of 446) and in terms of class skew (as demonstrated in Figure 1) - we also report the area under the precision-recall curve (AUPRC) as well as the micro-averaged precision and recall scores for a threshold of 0.5. For the single-label task, we report accuracy as the classes are more balanced than in the multi-label case.

5.2 Hyperparameters and Results

The primary hyperparameters for our algorithm are the learning rate, the number of channels, the dropout rate, and the l2-regularization rate. We first tuned the learning rate and number of channels by performing a log-scale random search. We settled on 64 channels for both models because both models exhibited high bias with lower numbers of channels, necessitating an increase in model power. To combat the overfitting that resulted from the increased size of the network, we also performed log-scale random searches for both the l2-regularization rate and dropout rate, leading to the choice of 0.001 regularization rate and 0.8 dropout rate (meaning 20% of activations were dropped) for the multi-label model and 0.1 regularization rate and 0.5 dropout rate for the single-label model.

AUROC	AUPRC	Precision (threshold=0.5)	Recall (threshold=0.5)
0.315	0.899	0.683	0.209

Figure 4: Multi-Label Results

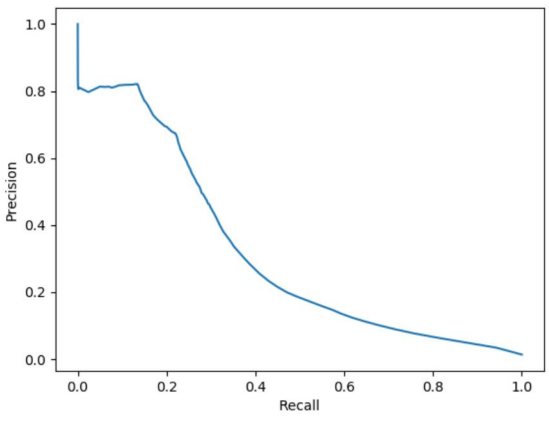

Figure 5: Multi-Label Precision-Recall Curve

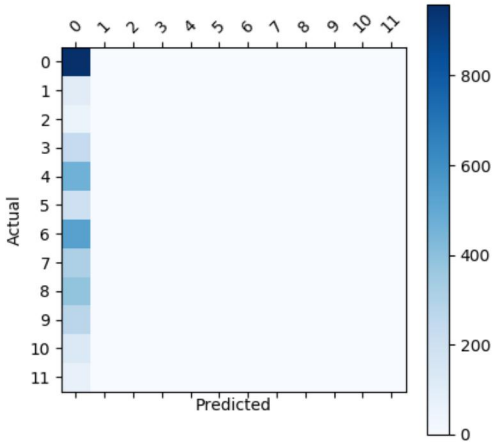

Figure 6: Single-Label Confusion Matrix

5.3 Analysis

5.3.1 Multi-Label

From the Multi-Label results in Figure 4, we see that although the model achieved a high AUROC score, it was limited to a poor AUPRC score. In addition, in Figure 5 we see that 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.

Comparing our model to our baseline, we find that the AUROC score of our model comfortably surpasses that of the ResNet-101 pretrained on ImageNet [14]. While this is not an entirely equivalent comparison - Oramos et al. report the macro-averaged AUROC score while we report the micro-averaged - it is a surprising result nonetheless given the power that transfer learning provides. One possible interpretation could be that a model pretrained on ImageNet is equipped to recognize many different object types, but the genre of an album maps more on the general style (color, aesthetic, etc.) of an album cover than any specific object on a per-genre basis. However, it is also important to note that the AUROC score is likely not a reliable metric for this task due to the unbalanced structure of the datset. Specifically, the vast majority of labels in the problem are negatives because each album is on average labeled with 6 out of 446 genres, so the specificity component of the AUROC score - $TrueNegatives/(TrueNegatives + FalsePositives)$ - will be dominated by the high number of true negatives in the dataset. Because of this, it is possible to develop a model that performs well with regard to AUROC while still performing poorly with regard to precision and recall. However, without the precision/recall metrics of the baseline model, we can't know if this was indeed the case.

5.3.2 Single-Label

In Figure 6 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. It also performed far below the results reported by Dammann and Haugh [6] for their simple k-nearest-neighbor approach. They achieved over 90% classification accuracy by compressing images using PCA and LDA and then using the compressed data in their kNN algorithm. Again, these results are not directly comparable as they were on a smaller dataset of 4000 images spread across only 4 distinctive genres - Christian, Metal, Country, and Rap. However, the success of their simple model perhaps suggests that image-based genre classification is an inherently difficult task due to the diversity of album covers within genres and thus benefits from preprocessing such as PCA and LDA that identifies the most important features of the data. However, conducting techniques such as PCA and LDA prior to passing them as inputs would disable investigation into what components of image styles or features the system is learning from, which is what we hoped to enable through this work.

6 Conclusion/Future Work

While our models in both approaches struggled to achieve top performance, they nevertheless reveal significant insight into the relatively unexplored realm of image-based music genre classification. The struggles of both our multi-label model as well as the baseline to gain strong traction on the problem suggests that perhaps the best path forward would be to train a model to predict only the high-level genre (rock, rap, country, etc.) of an album as these genres are often mostly orthogonal, and then develop separate, more specialized classifiers to classify the more specific sub-genres after the initial disaggregation. Another definite limitation in this work was the size of the MuMu dataset. While sufficiently large, our models could benefit from much more data. However, the MuMu dataset is one of the few public datasets mapping album covers to album genres. Therefore an area for future work is to develop a large dataset exclusively for image-based genre classification. Finally, an interesting path to explore would be to use a model pretrained on ImageNet 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. This could be useful because it would focus the system intensively on the style of the album covers, which is precisely what we want to investigate with respect to genre.

7 Contributions

Christopher Koenig conducted all work for this project, including topic field research, dataset gathering and cleaning, model design and development, analysis of results, and drawing conclusions.

References

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In ISMIR, 2011. (http://millionsongdataset.com/)

[2] Dmitry Bogdanov, Alastair Porter, Perfecto Herrera, and Xavier Serra. Cross-collection evaluation for music classification tasks. In ISMIR, 2016.

[3] Kahyun Choi, Jin Ha Lee, and J Stephen Downie. What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 453–454. IEEE Press, 2014.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. ISMIR, 2016.

[5] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. *arXiv preprintarXiv:1609.04243*, 2016.

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

[7] Sander Dieleman, Philemon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In ISMIR, 2011.

[8] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pages 6964–6968. IEEE, 2014.

[9] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal music mood classification using audio and lyrics. In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on, pages 688–693. IEEE, 2008.

[10] Janis Libeks and Douglas Turnbull. You can judge an artist by an album cover: Using images for music annotation. IEEE MultiMedia, 18(4):30–37, 2011.

[11] Rudolf Mayer, Robert Neumayer, and Andreas Rauber. Rhyme and style features for musical genre classification by song lyrics. In ISMIR, 2008.

[12] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 43–52. ACM, 2015.

[13] Robert Neumayer and Andreas Rauber. Integration of text and audio features for genre classification in music information retrieval. In European Conference on Information Retrieval, pages 724–727. Springer, 2007.

[14] 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.

[15] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on, pages 1–6. IEEE, 2016.