Album Covers Deserve Some Attention

Natalie Greenfield
Stanford University
450 Jane Stanford Way
natgreen@stanford.edu

Ngorli Paintsil
Stanford University
450 Jane Stanford Way
ngorlip@stanford.edu

Abstract

Album covers are more than just a decoration. They often provide a visual narrative that complements and even enhances the auditory experience of music albums. Before digital music, album cover art had a much greater presence and presumable influence on listeners choosing which music to listen to. As modern day music recommendation approaches are developed, including album cover art could be a valuable asset. We explore if cover art can be an effective proxy for determining genre, the primary way to classify music. Some work has been done to approach this, but has been generally unsuccessful due to unbalanced datasets. Our approach uses a balanced dataset with 18 genres and uses two different approaches to classify the data. First, we used a DenseNet architecture, DenseNet-201, pre-trained on ImageNet. The other architecture we explored is ViT-16/B, a Vision Transformer model pre-trained on ImageNet. The results for both models were positive. The DenseNet achieved an accuracy of 31.44% on the test set and ViT-16/B achieved an accuracy of 33.00%. While the models we explored still struggled to classify similar genres, their overall performance showed significant improvement over previously explored methods.

1. Introduction

Album covers are an important part of the creative process of producing music. Historically when buying physical copies of music was more popular, album cover art had a greater influence on attracting listeners. Furthermore, there tends to be certain characteristics of album cover art that are generally associated with certain genres. For example, country albums typically depict a cowboy aesthetic and rural landscapes while Jazz albums typically have depictions of instruments and musicians playing those instruments. Today album cover influence is much less as music selection is dominated by various recommendation algorithms that do not take album covers into account. Rather, artists, lyrics, and melody are used by these algorithms to classify songs and calculate similarity scores. Spotify claims that their algorithm uses the ”characteristics of the content itself, such as its genre, release date, podcast category, etc.” This allows them to ”identify which content has similar characteristics and might be enjoyed by similar listeners.” [17]

Exploring if a relationship can be determined between an album's genre and it's cover art could produce a new method for music classification which could possibly be used in music recommendation albums. In a recent study of Spotify music genres, specifically their music genre database, it was found that there are over 6,000 different genres, as of 2023. [19] The study found a lot of overlap in this vast number of genres. In this case that relationship can be determined, album covers could act as a proxy for their contained music and be an interesting addition to music recommendation algorithms. This would have a valuable application to the music industry but also individuals in using an alternative metric to classify album genres especially for niche genres and sub-genres. To test the existence of this relationship, this study will have 18 different genres across a wide range of sounds and styles, with a wide range of artists and years within the album genres.

1.1. Problem Statement

Formally, the problem that this study is attempting to answer is single-label album genre classification using only the album cover art of that particular album. The input to our models are images of album covers. We then use a DenseNet and a Vision Transformer to output a predicted album genre.

2. Related Work

2.1. Multi-label Classification

There has been considerable work on adjacent problems already, with many different approaches being taken on a diverse array of datasets. First, papers that used a multi-label classification approach will be analyzed. One such paper is ”Multi-Label Music Genre Classification from Audio, Text, and Images, Using, Deep Features”. [14] This is a prominent paper in which Oramas et al created the MuMu dataset, which contains 31,461 images classified into 250 genres. The dataset was used by many other subsequent papers. The study found that using only image based classification to be the worst of the three modes they tried, however it did indeed show classification potential. Building off of this paper, the same researchers once again worked on this problem in another paper. Using a ResNet, they found some improvements, but still saw the classification task to be most successful when in tandem with other modalities such as music reviews and audio tracks.[12] In ”Genre Classification via Album Cover”, multi-label classification was attempted using VGG-16, with 3 unfrozen convolution blocks. The fine-tuning of this model showed success with an area under the ROC curve of 0.7844 on the test set. [1] [16]

The paper ”An Audio-Visual Approach to Music Genre Classification through Affective Color Features” approached this classification problem not with album covers but instead by using images from music videos, extracting the different features and characteristics present in a frame. This process of visual analysis and a SVM achieved about 50% accuracy. [15] Another paper that approached classification without Neural Networks was ”You Can Judge an Artist by an Album Cover: Using Images for Music Annotation”. In this study, Libeks et al created feature vectors based on the color characteristics (RGB, HSV, and more criteria) in the album covers. This paper reported an area under the curve accuracy of over 63% for several genres. [10] [8]

2.2. Single-Label Classification

A rather unique approach compared to the rest of the literature was taken by the researchers in ”Bridging Music and Image via Cross-Modal Ranking Analysis”. This paper generated semantic meanings for the different album covers and then used a SVM classifier trained on this data. Given that this paper explored generating music image pairs and not genre classification, the results are not applicable, though the paper does provide an interesting and novel approach. [21]. Another study that provided relevant information for solving this problem is ”Relationship between album cover design and music genres”. Although this paper did not classify album genres by album covers, it’s findings asserted that albums covers have strong relationships with their genres. Dorochowicz et al looked at similarities with factors such as size of text, cover compositions, and whether or not the cover had a photograph in addition to other factors. [3]

For single-label classification, the first work that does so is ”Genre Classification of Spotify Songs using Lyrics, Audio Previews, and Album Artwork”. In this study, the dataset only contained 4 genres, with 1000 images for each genre. Despite implementing an RNN, they found that a k-NN approach was the most successful, with a test accuracy of over 90%. We found these results to be promising, but not hold to much weight because of such a small dataset. Furthermore, Christian, country, metal and rap are albums that have practically no overlap in their musical style and aesthetic, so this was not a real test of a models capabilities. [2] In the paper ”Classifying Album Genres by Album Artwork”, the larger MuMu dataset that was referenced earlier was used. MuMu is a very unbalanced dataset, and even with some adjustments made by the researchers to try and balance it for single-label work, 2 of the 12 genres used made up 25% of the dataset. Using a 5 layer CNN, the results of this paper were unsuccessful, with the researchers stating that the model was always predicting the genre that appeared most frequently in the dataset. [7] The struggles that Koenig had in this paper was the motivator for choosing a large dataset that had the same amount of album covers in each genre.

The single-label classification paper that will serve as a baseline for the work in this paper is ”Predicting Musical Genre from Album Cover Art”. In this study, a balanced dataset of 8800 album covers was used. Among the different pre-trained architectures implemented, AlexNet performed the best, achieving an accuracy of 18%. [9] [8] Unlike other research, this paper benefited from using a balanced dataset. As a result, the classifier did not just predict the most common genre and was able to learn and differentiate features. Furthermore, they used a similar size of dataset in both number of genres and number of images per genre. Their success in using a balanced dataset with a variety of genres inspired the choice of dataset for this study.

3. Dataset

The dataset utilized for this study is the ”20k Album Covers within 20 Genres” on Kaggle, curated by Michael Kerr [6]. This dataset comprises 20,000 images, with 1,000 images allocated to each of the 20 genres. The images stretch over a wide range of years within the respective genres. In contrast to most research in this field, which employs the larger yet unbalanced MuMu dataset, we have selected this balanced dataset for our analysis. Previous studies have indicated that the imbalance in the MuMu dataset impedes classification accuracy, often resulting in models biased towards the majority class. [13] [7] The use of a balanced dataset in our study aims to mitigate this issue. However due to intense overlap between the genres Doom Metal, Death Metal, and Heavy Metal we adjusted the dataset to only include Heavy Metal and not the other metal genres, bringing it to a total of 18 genres and 18,000 images. Table 1 displays the 18 genres included in the dataset.

A 80/10/10 split was used for training, validation, and test sets, respectively. Thus for each genre, 800 album cov-

Table 1. List of Genres

Genre	
Blues	Classical
Country	Drum N Bass
Electronic	Folk
Grime	Heavy Metal
Hip-Hop	Jazz
Lo-Fi	Pop
Psychedelic Rock	Punk
Reggae	Rock
Soul	Techno

ers were part of the training set, bring the total size of the training set to 14,400 images, with 180 images for validation and test sets. The only data preprocessing required was to resize and crop images to 224x224 before any training. The resolution of the images prior to any preprocessing is 512x512. A sample image from the dataset shown in Figure 1. We also explored several kinds of data augmentation. While we tried several types of color augmentation, these all resulted in worse performance by the DenseNet and Transformer models, suggesting that color contributed significantly to the features learned by these models. We also implemented random cropping, horizontal flipping, and rotation. These transformations put the items in the album cover in different physical places on the cover and effectively helped our models generalize to our validation and test set.

Figure 1. Sample Album Cover from the Dataset

4. Methods

4.1. Baseline

The baseline for this classification task will be the model used in ”Predicting Musical Genre from Album Cover Art” given the similarities in dataset size and number of genres. [9] This model used multiple neural networks in the approach for the classification task. The best performing model in this study was the transfer learning approach using AlexNet. The model was pre-trained on ImageNet and fine tuning was performed by un-freezing the last layer to train on their dataset. The architecture of AlexNet from the original paper is an eight layer net. Some key novelties introduced in this paper were ReLU nonlinearity and local response normalization. ReLU nonlinearity allowed for large increases in training speed while local response normalization assisted in the generalization of the model. The first five layers are convolutional and the last three are fully connected. The output of the last layer is fed into a softmax classifier. The kernels of the second, fourth and fifth convolutional network are connected to the kernel maps which reside on the same GPU in the previous layer. The kernels of the third convolutional layer are connected to all of the kernel maps in the second layer. For the fully connected layers, those neurons are connected to all of the neurons that are in prior layer. [8] In this study 10 genres were used: ambient, dubstep, folk, hip-hop, jazz, metal, pop, punk, rock, and soul. The test set contained 100 album covers per genre. The architecture above described above achieved a maximum test accuracy of 18%, out performing randomness by 8%. The other models used in this paper were Resnet18, Resnet34, and Resnet152. These models all achieved accuracies ranging 16% to 17%.

4.2. AlexNet

To provide a baseline for our own dataset, the AlexNet architecture used in the baseline paper was implemented. Figure 2 is an image of the AlexNet architecture. As stated in the prior section, a pretrained AlexNet was used with a unfrozen final layer for model fine tuning. The detailed architecture of AlexNet is as follows. First is a layer with 96 kernels of size 11x11 and a stride of 4. Next is a convolutional layer with 256 kernels, size 5x5. The next three convolutional layers are connected without any normalization or pooling layers. The third layers has 384 kernels of size 3x3. The fourth layer has the same number of kernels with the 3x3 size. The fifth layer has 256 kernels of size 3x3. Finally, each of the three fully-connected layers at the end have 4096 neurons each.

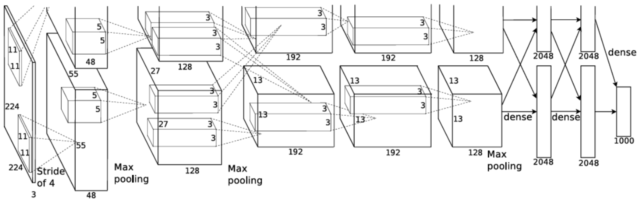


Figure 2. Alexnet Architecture [8]

4.3. DenseNet

The first architecture that we implemented was a DenseNet. Specifically the DenseNet-201 model from PyTorch that is pretrained on ImageNet.

The paper that developed this architecture is ”Densely Connected Convolutional Networks”. [5] The major novelty in the DenseNet architecture lies in how flow is maximized within the model. This is achieved by connecting all layers directly to each other. By making sure each layer obtains inputs from all prior layers and still passing on its own feature maps, the feed-forward nature is still preserved. This dense connectivity allows for greater flow and higher performance. DenseNets are also advantageous for their high parameter efficiency and the regularization effects on smaller datasets that arise from the dense connections. This last feature is particularly beneficial for this study given the relatively small size of the chosen dataset.

The paper proposed four different DenseNet architectures, but for the scope of this paper, only DenseNet-201 will be considered. The first layer in the DenseNet is a 7x7 convolutional layer with a stride of 2. Next is a 3x3 max pooling layer with a stride of 2. The main portion of the architecture is the three sets of dense blocks followed by transition layers. A dense block consists of 1x1 convolutional layer followed by a 3x3 convolutional layer with the dense connectivity structure highlighted earlier. These two layers are repeated multiple times to create the block. The first block is repeated 6 times, the second block 12 times, and the third block 48 times. After each block, the transition layer is a 1x1 convolutional layer followed by a 2x2 average pool layer with a stride of two. Following the final transition layer, there is another dense block repeated 32 times. The final two layers in the architecture are the 7x7 global average pool and a fully connected softmax. Note that every convolutional layer described is a batchnorm-ReLU-convolution.

4.4. Vision Transformer

The second architecture that we explored is a vision transformer. Specifically, a ViT-B/16 model from PyTorch that is pre-trained on ImageNet.

The two most relevant papers that contributed to the creation of this model are ”Attention is All You Need” and ”An Image Is Worth 16X16 Words: Transformer For Image Recognition At Scale”. [20] [4] The former paper proposed the original transformer architecture which was intended for use in language translation tasks. This model leverages scaled dot-product attention to compute attention scores. Given values(V), queries(Q) and keys(K), self-attention can be computed by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The transformer model consists of an encoder block and a decoder block. The encoder uses multi-headed self attention and fully connected layers interspersed with activation to encode relationships between elements. The decoder block also uses multi-headed self-attention as well as regular attention to focus on the encoder outputs. Multi-headed attention, which projects inputs into multiple subspaces, performs attention and the concatenates the results, allowing the model to capture different aspects of the relationships between inputs. Both parts use positional encodings to maintain information about the sequence order.

The ViT-B/16 model that we used is described in the second paper. This model expands upon the original transformer model to make it viable for use with images. This involves dividing images into 16-by-16 pixel patches which are then encoded into a vector through linear transformation. A positional encoding is given to each patch to maintain spatial information and then the sequence of vectors are inputted to the transformer described above. The model leverages the transformer’s ability to handle long-range dependencies, making it effective for understanding complex visual data, such as the images seen on album covers.

4.5. Loss

All of the models that we explored used multi-class cross entropy loss. Given N samples and C classes, cross entropy loss can be calculated by:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log(\hat{y}_{i,y_i}) \tag{2}$$

$\hat{y}_{i,y_i}$ is the predicted probability that the model will output the true class y_i for the ith sample. Cross entropy loss is effective as it uses probabilistic interpretation which enables it to penalize predictions that are confident but wrong.

4.6. Consideration of Alternate models

We will briefly discuss other methods that were considered or attempted and their shortcomings. The other architecture that was considered for this project was using an InceptionNet, specifically an InceptionNet V3. The inceptionNet was a result of the work done by Szegedy et al in the paper ”Rethinking the Inception Architecture for Computer Vision” [18]. The key features of the InceptionNet V3 are its label smoothing and its ability to classify multi-scale features more efficiently because of its inception modules, which apply multiple types of convolutions and pooling operations in parallel. As a result, we felt that using an InceptionNet could be an potential successful approach for model classification. Ultimately, this approach was abandoned as it only received a test accuracy of 20.56% after a considerable amount of hyper parameter tuning.

5. Experiments and Results

5.1. Metrics

We selected three metrics to analyze our model: accuracy, top-k accuracy and confusion matrices. The paper that created our baseline model used accuracy as it's performance metric so including this as our first metric allows for direct comparison of our model to the baseline. Through observing many images from the dataset, we determined that album covers often have features that overlap with several genres despite only having a single label. Due to this overlap, we wanted to include both top-3 and top-5 accuracy. Our models use softmax to produce the output, so we can analyze if the correct class was in the top-3 or top-5 most probable classes. We felt including these could allow for additional insights such as if the model was considering the correct genre but ultimately selected an adjacent genre. The last metric we selected was confusion matrices. This was a common metric across many of the papers we analyzed as it effectively shows which genres are classified the most and least frequently, as well as what other genres they may be misclassified as. It also clearly shows when models are only majority class classifiers and not effectively picking up on album cover features, which was an issue noted in the paper ”Classifying Album Covers by Album artwork” which also attempted single-label album cover classification by genre.[7]

5.2. Alexnet

Using AlexNet as a baseline showed poor results. Following the architecture described in the paper and outlined in the baseline section, the same hyperparameters as the original paper were used. They reaserchers used a learning rate of 0.01, a batch size of 10, and a SGD optimizer with a momentum of 0.9. Finally, a softmax classifier with cross-entropy loss was used. A final test accuracy of 5.78% for Top-1, 17.00% for Top-3, and 28.22% for Top-5 was achieved. The confusion matrix is shown in figure 3. Looking at the confusion matrix, the model was barely classifying any of the data. Instead, Electronic was predicted 98.6% of the time. The only other genres that were correctly classified at least once were Grime, Lo-Fi, and Psychedelic Rock. It is also worth noting that of the 26 predictions that were not Electronic, 11 of them (42%) were Lo-Fi.

5.3. DenseNet

To tailor the DenseNet-201 model to our data we made several modifications. The last three layers were all unfrozen, allowing for fine-tuning on the dataset. The optimizer used was an AdamW optimizer with a learning rate of 1e-4, betas = (0.9, 0.999), epsilon of 1e-8 and a weight decay of 1e-2. We also noticed that the model was overfitting the training data so we added data augmentation, which performed cropping, horizontal flipping, and rotation. We also added dropout which was fine-tuned to 0.7. For data augmentation, cropping, horizontal flipping, and rotation were all used. Finally, a stepLR scheduler was used with a step size of 3 and a gamma of 0.1.

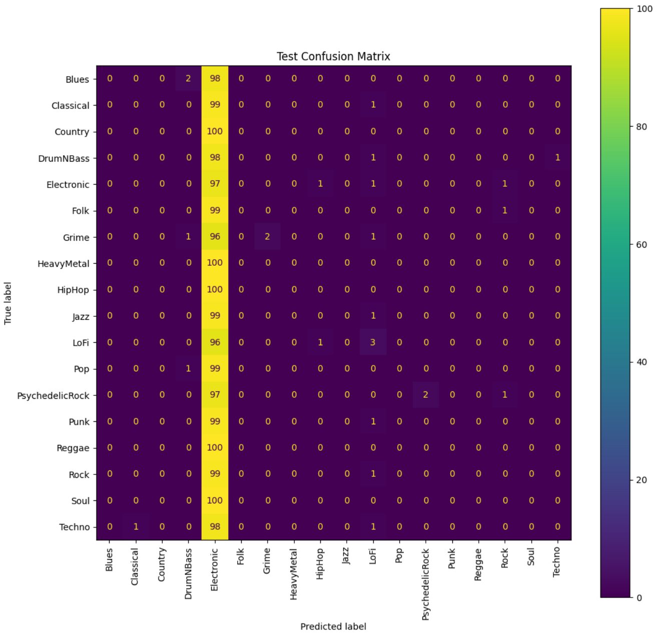


Figure 3. AlexNet Confusion Matrix

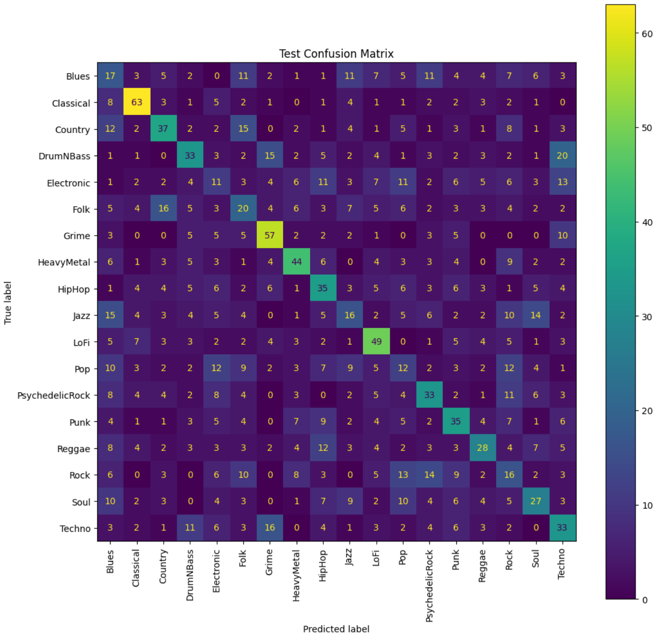
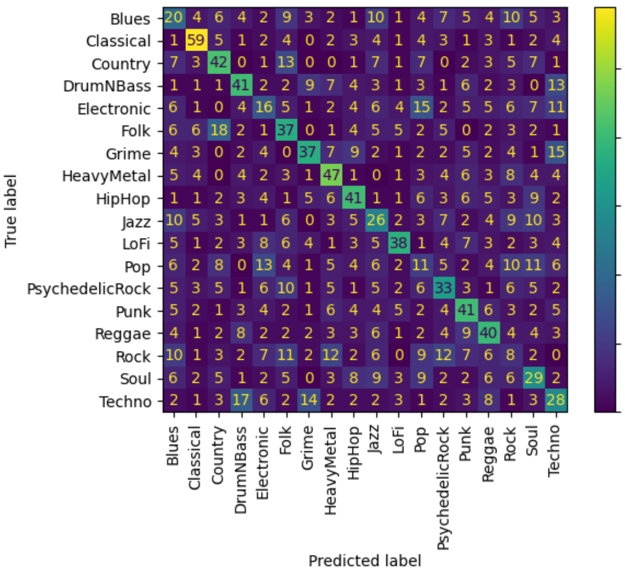


Figure 4. DenseNet Confusion Matrix

The results achieved by the DenseNet were more promising. The DenseNet achieved a top-1 accuracy of 31.44%, a top-3 accuracy of 54.72% and a top-5 accuracy of 67.00% on the test set. The confusion matrix is shown in figure 4. The genres that the model classified the best were classical, grime, and Lo-Fi. The genres the model had the worst performance with were Electronic, Rock, Jazz, and Pop. It is worth noting that every genre achieved a classification accuracy of at least 10%.

5.4. Vision Transformer

The model that achieved the best test accuracy was the ViT-B/16 model. To achieve this, we fine-tuned the model as well as tuned the hyper parameters. To allow for fine-tuning, the last two encoder layers of the model were unfrozen. An AdamW optimizer was used with a learning rate of 1e-4 betas = (0.9, 0.999), epsilon of 1e-8, and weight decay of 1e-2. We noticed that the model was outfitting the training data so we included data augmentation(cropping, horizontal flipping, and rotation) and dropout which was 0.7. We trained for total of 20 epochs. ViT-B/16 achieved a top-1 accuracy of 33.00%, a top-3 accuracy of 55.06%, and a top-5 accuracy of 66.89% on the test set. The confusion matrix is shown in Figure 5. The genres that the model classified the best were Classical, Heavy Metal, and Country. The genres the model had the worst performance with were Rock, Pop, Electronic. Of the 18 models classified, 15 of them achieved a classification accuracy of at least 20% and 12 achieved an accuracy of 30%.



	Blues	Classical	Country	DrumNBass	Electronic	Folk	Grime	HeavyMetal	HipHop	Jazz	LoFi	Pop	PsychedelicRock	Punk	Reggae	Rock	Soul	Techno
Blues	20	4	6	4	2	9	3	2	1	10	1	4	7	5	4	10	5	3
Classical	1	59	5	1	2	4	0	2	3	4	1	4	3	1	3	1	2	4
Country	7	3	42	0	1	13	0	0	1	7	1	7	0	2	3	5	7	1
DrumNBass	1	1	1	41	2	2	9	7	4	3	1	3	1	6	2	3	0	13
Electronic	6	1	0	4	16	5	1	2	4	6	4	15	2	5	5	6	7	11
Folk	6	6	18	2	1	37	0	1	4	5	5	2	5	0	2	3	2	1
Grime	4	3	0	2	4	0	37	7	9	2	1	2	2	5	2	4	1	15
HeavyMetal	5	4	0	4	2	3	1	47	1	0	1	3	4	6	3	8	4	4
HipHop	1	1	2	3	4	1	5	6	41	1	1	6	3	6	5	3	9	2
Jazz	10	5	3	1	1	6	0	3	5	26	2	3	7	2	4	9	10	3
LoFi	5	1	2	3	8	6	4	1	3	5	38	1	4	7	3	2	3	4
Pop	6	2	8	0	13	4	1	5	4	6	2	11	5	2	4	10	11	6
PsychedelicRock	5	3	5	1	6	10	1	5	1	5	2	6	33	3	1	6	5	2
Punk	5	2	1	3	4	2	1	6	4	4	5	2	4	41	6	3	2	5
Reggae	4	1	2	8	2	2	2	3	3	6	1	2	4	9	40	4	4	3
Rock	10	1	3	2	7	11	2	12	2	6	0	9	12	7	6	8	2	0
Soul	6	2	5	1	2	5	0	3	8	9	3	9	2	2	6	6	29	2
Techno	2	1	3	17	6	2	14	2	2	2	3	1	2	3	8	1	3	28

Figure 5. ViT Confusion Matrix

6. Discussion

6.1. Alexnet

The results for the baseline AlexNet were quite disappointing. The model did not do any true classification and instead almost always predicted one genre. This resulted in a test accuracy that was almost equal to random guessing. This statistic holds true for both Top-3 accuracy and Top-5 accuracy as well. It is interesting that the genre that the model over-fit to was Electronic. One potential reason this could occur is the wide range of styles and colors in Electronic album covers. Looking through this genre in the dataset revealed no particular album style, and instead showed albums that held characteristics that would generally be associated with other genres. Below are two album covers that are classified as Electronic, with the one on the left having characteristics of Country, Folk, or Blues, and the one on the right having characteristics of Psychedelic Rock, Techno, and Punk, among others.

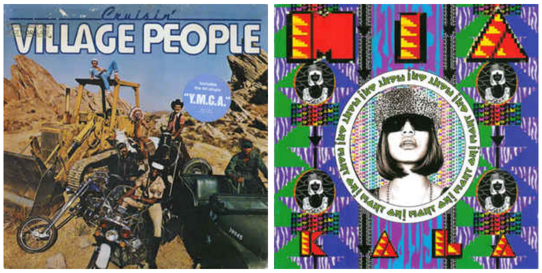


Figure 6. Two Electronic Album Covers

Due to such large diversity in the Electronic dataset, this could help to explain why the model over-fit to electronic in training and constantly predicted it. Furthermore, given how rudimentary this baseline was, a relatively poor performing was to be expected. Data augmentation or any type of regularization would have likely seen improvements in the test accuracy. Furthermore, a smaller learning rate would have likely performed better based on the experimental results of other models that were tested and used in the experimentation phases of this paper.

6.2. DenseNet

The DenseNet performed significantly better than the baseline. The strong diagonal in the confusion matrix shows that the model was classifying genres in a meaningful way. Classical being the genre with highest classification is not a surprising result, as the classical album covers have a lot of homogeneity as a genre, typically depicting instruments like pianos, woodwinds, and strings. The high performance of grime and Lo-Fi was slightly alluded to by the results from the AlexNet, as they were one of few genres that were correctly classified. Again, this is likely due to some sort of distinctive feature in their genre datasets; for grime, their is an relatively high frequency of album covers that have CDs on them. However for Lo-Fi, no real pattern stood out when analyzing the data. There are also some interesting patterns to see in the genres where the model struggled more. Pop was correctly classified just as many times as it was classified as Rock and Electronic. Electronic was classified more times as Techno than it was as pop. Patterns like these appear all over the confusion matrix where genres that are similar in their style to others are misclassified as others and genres that have a wide range of styles generally struggle to be classified accurately.

6.3. Vision Transformer

The ViT was able to achieve a slightly better performance than the DenseNet, outperforming it by 1.66% on the test set. Once again, classical was the genre with the highest classification accuracy. Although grime and Lo-Fi maintained relatively high performance, in contrast to the DenseNet, heavy metal and country were the two genres with the next highest accuracy. The high classification accuracy for heavy metal is expected for similar reasons to other genres that performed well, as it has a distinctive style that is not similar to other genres. However the high accuracy for country is more surprising, given that albums usually consist of landscapes, people, and some animals, which are seen frequently seen in genres like folk and occasionally blues and rock. The ViT still struggled with electronic, pop, and rock, confirming the generality of the style of these genres. Rock was especially difficult for the model, as 5 different genres were predicted for rock more frequently than rock itself.

7. Conclusion and Future Work

Album genre classification by album cover art has proven to be a challenge for researchers for years. The work done in this study showed significant progress towards being able to classify the genres of album covers based on the cover art. The DenseNet greatly improved on prior work done by VGG, ResNet, and AlexNet, showing the potential for more advanced and higher performing convolutional neural networks to solve this problem. Additionally, applying the transformer which is a more recently developed model proved to have slightly greater performance compared to the CNN approach. This speaks to the potential that transformers have in the field of image classification, especially as this relatively new style of architecture for image classification continues to improve.

For future work with more resources and time, continuing to improve the transformer model would be the path taken. Transformer models are particularly finicky to fine tune and take significantly longer to train. These constraints limited the amount of iteration of different hyperparamters, data augmentation, and fine-tuning that is possible within a class project. Additional time and resources toward fine-tuning the transformer could yield a set of higher performing hyper parameters. Furthermore, with additional storage capabilities, the SWIN model, from the paper ”Swin transformer: Hierarchical vision transformer using shifted windows” could allow for the detection of more complex features which could improve accuracy. [11] Based on the results of this paper, we conclude that the most promising approach to use when classifying music album cover art by genre is vision transformers.

8. Contributions & Acknowledgements

Ngorli and Natalie worked equally on all parts of the model including training the models and researching and writing the paper.

References

[1] S. Choudhury and J. Kim. Cs230 project report: Real-time emotion detection in music. Stanford University, 2020.

[2] T. Dammann and K. Haugh. Genre classification of spotify songs using lyrics, audio previews, and album artwork. Stanford CS229 Machine Learning, 2017, 2017.

[3] A. Dorochowicz and B. Kostek. Relationship between album cover design and music genres. In *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, G. Narutowicza 11/12, Gdańsk, Poland, September 2019. Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Audio Acoustics Laboratory.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2017.

[6] M. Kerr. 20k album covers within 20 genres. Kaggle, 2022.

[7] C. Koenig. Classifying album genres by album artwork. Stanford CS230 Deep Learning, Spring 2019, 2019.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates Inc., 2012.

[9] N. Lee and R. Baraldi. Predicting musical genre from album cover art. CSE546 Final Paper, University of Washington, 2019. Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA.

[10] J. Libeks and D. Turnbull. You can judge an artist by an album cover: Using images for music annotation. *IEEE*, 2015.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[12] S. Oramas, F. Barbieri, O. Nieto, and X. Serra. Multimodal deep learning for music genre classification. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 278–285, 2017.

[13] S. Oramas, L. Espinosa-Anke, M. Sordo, O. Nieto, and X. Serra. Mumu: A multimodal music dataset. `https://github.com/sergiooramas/mumu`, 2017. Accessed: 2024-06-03.

[14] S. Oramas, O. Nieto, F. Barbieri, and X. Serra. Multi-label music genre classification from audio, text, and images using deep features. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[15] A. Schindler and A. Rauber. An audio-visual approach to music genre classification through affective color features. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*. Springer, Cham, 2015.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[17] Spotify. Understanding recommendations. https://www.spotify.com/us/safetyandprivacy/understanding-recommendations.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[19] The Pudding. The massive genre breakdown, 2023. Accessed: 2024-05-15.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[21] Y. Xu, J. Du, L. Dai, and C.-H. Lee. Deep learning for acoustic event detection and classification. *IEEE Transactions on Multimedia*, 18(7):1305–1318, July 2016.