Predicting Musical Genre from Album Cover Art

Nathan Lee & Robert Baraldi, Applied Mathematics

1. Introduction

In this project, we attempt to classify the genre of album based on its cover art. While an important visual component, it is difficult to enunciate how style varies from image to image as well as quantify said difference. Hence, it would be interesting to see if we can train models to identify major components of certain styles of cover art and use these models to identify the corresponding musical genre.
Use the 10 most popular genres of music on bandcamp. These are:

- ambient (0)
- dubstep (1)
- folk (2)
- hiphop (3)
- jazz (4)
- metal (5)
- pop (6)
- punk (7)
- pop (6)
- rock (8)
- soul (9)

The training dataset has 8800 album covers and the testing dataset has 1000 randomly selected album covers. The number of albums per genre in the training set vary between 990 and 970. The testing set has an equal number of album covers per genre. A sample of these is displayed in Figure 1.

Figure 1: Sample album covers for ambient, metal, and punk respectively.

2. Color Analysis Part 1

We use `opencv` to scale images to 32x32 size as well as calculate and normalize each image's rgb histogram with 8 bins in each channel. We use these to find the similarity scores between 5 randomly selected albums in each genre, iteratively. A sample heatmap is given in Figure 2. Ranking the color histogram similarity shows that only metal and pop are most similar to themselves, while ambient and jazz are very different from themselves, indicating a wide variance for some genres. This shows color is a useful indicator for some genres, but certaintly some higher level features exist that are not quite being captured by this analysis.

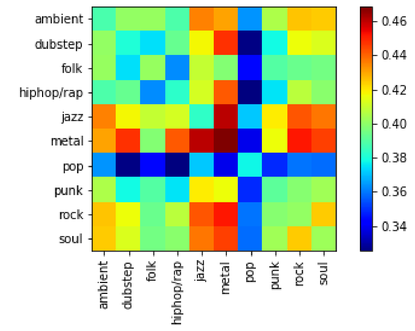

Figure 2: Sample heat map of averaged similarity scores between albums. Note that red (higher) values mean more related, while blue (lower) values are less related.

3. Color Analysis Part 2

Genre	Rank to itself
ambient	9
dubstep	7
folk	2
hiphop/rap	7
jazz	9
metal	1
pop	1
punk	6
rock	7
soul	7

Table 1: Ranked Distances in randomly chosen album covers.

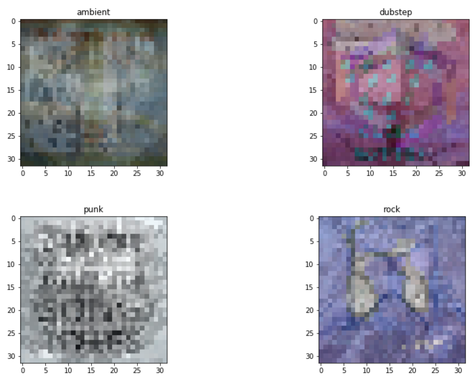

Figure 3: Sample averaged images.

4. Neural Network Analysis

We consider two approaches of classifying with neural networks: 1) training models on our training data and then using them to predict the test data, and also 2) implementing models that have previously done well in classifying Imagenet data and retraining the last layer on our data. This will allow us to determine if our dataset is rich enough to fully train a neural network as well as use some of the richness of the imagenet dataset in our favor. Our batch size is 10 and our learning rate is 0.01 with SGD and a momentum factor of 0.9. For all the networks used, we implement the softmax classifier with cross-entropy loss. We test several models: the first is a custom model consisting of 2 convolutions, a MaxPool, and 3 linear layers. This model and Alexnet were trained fully on our data, and the results are in Table 2. The custom model had an overall error of 17.6%, while Alexnet had an overall error of 15%.

Table 2: Fully Trained, % Accuracy.

Genre	Custom	Alexnet
ambient	24	49
dubstep	17	0
folk	24	40
hiphop/rap	6	3
jazz	10	0
metal	44	41
pop	14	0
punk	21	0
rock	3	1
soul	13	0

Table 3: Pre-Trained, % Accuracy.

Alexnet	Resnet18	Resnet34	Resnet152
19	19	22	19
15	17	13	13
24	12	19	11
11	19	6	7
11	14	7	11
38	35	47	39
16	11	9	13
19	15	15	16
7	12	6	9
13	17	12	10

Next, we used transfer learning on Alexnet, Resnet18, Resnet34, and Resnet152, all shown in Table 3. The best Alexnet model selected had an accuracy of 18% on the testing data. What was interesting in particular about this model in constrast to the others was that the training accuracy was usually about 49%. Resnet34 had an accuracy of 19% overall, while Resnet18 had an overall accuracy of 17.5%. Resnet 152 (which had the highest accuracy on Imagenet) only had an overall accuracy of 16%. Between these, it is interesting to see which genres are consistently higher. For instance, metal is consistently at least 30%, where as rock and jazz are typically low.

Conclusions

The color histograms showed that the color content alone of the album art was generally not useful for most genres, and many of the standard ML techniques from `sk-learn` more or less resulted in classification probabilities equal to guessing. While image classification with neural nets had some success in some genres, the overall result was poor. We are trying to classify *style* of an image rather than its content. Indeed, even the humans we tested this on did not perform well. Another concern we raised is that our dataset is not robust enough. We also believe that imposing strict genre classification is constraining and not indicative of reality.

References

[1] Yanir Serouss. Learning about deep learning through album cover classification. 2015.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
[3] Nitin Viswanathan. Artist identification with convolutional neural networks. 2017.
[4] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, arXive preprint: arXiv:1311.3715v3, 2014.
[5] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, Jun 2010.
[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.