CMSC 25040
Computer Vision
Professor: Michael Maire

Genre Style Translation for Album Artwork Creation

Isabel Garon
Neuroscience
The College, The University of Chicago
isabelgaron@uchicago.edu

Ruben Abbou
Computational and Applied Mathematics, Statistics
The College, The University of Chicago
ruben@abbou.com

Winter 2020

Code: GitHub

1 Introduction

A significant challenge in the field of computer vision is the expectation that in performing a visual task, the computer navigates the linguistic and often cultural facets of visual cognition. The wealth of data available on musical album covers provides an interesting tool to explore the capacity of a computer to meet these expectations. Within a single musical genre, although there is a high degree of variability in album covers, humans can guess the genre of an album cover with a high rate of accuracy. It also has been shown that some of the low-level statistical features of album covers can be associated with the genre of an album, and therefore encode valuable information that is useful for contextualizing music [1].

In the following study, we will investigate methods of translating the style from a given genre to albums of a different genres, using Cycle-Consistent Adversarial Networks [2]. We aim to perform collection-style transfers from entire genres to out-of-genre album covers. We anticipated this approach would be best suited to our dataset for two reasons: first, the unpaired nature of collection style transfer is feasible with this data pool – it would be impossible to generate pairs of the same album in multiple genres in a non-work-intensive manner. Second, we wanted to explore what generalizations would be made across an entire genre of music, instead of translating the style of a single artist to another. It is possible that the dataset is too variable for cycle-consistency to provide a sufficient constraint, in which case we will also evaluate the performance of conditional style transfer techniques to achieve artistic style transfer [3].

2 Dataset

The Multimodal music dataset used for the cycle GAN training is synthesized from the Amazon Reviews dataset and the Million Song dataset [4]. It is composed of 31,471 albums each tagged with their genres. Each album cover was resized to a consistent 256 by 256 pixels. Within this larger dataset, which provided varying levels of genre specificity (from over 23,000 "Rock" tags to less than 100 tags for "Ambient Pop"), we chose to work with only a few genres. It is important to note that a single album can be tagged with several different genres. - for instance, an album tagged as "Ambient Pop" can also be tagged as "Dream Pop," "Lo-fi," and "Rock." We focused our trials on genres of music that were most reliably correctly categorized in literature [1] based on several image distance metrics – classical was the best performing genre for album property based categorization, followed by metal and dance, and on genres on which we expected artwork to differ more naturally from others. Within the metal genre, we have in total 2,869 images. In classic, there are 2,807, in dance there are 2,255 images, and in country there are 1,524.

We observed a high degree of variability within each of these genres, which leads to difficulties in obtaining any kind of translation due to the cycle-consistency constraint. We therefore attempted different techniques of clustering, to group albums by similarities within each genre – these clustering strategies are outlined below. Additionally, we extracted each album with faces in the cleaning process, to remove artwork likely to include photography and focus on less realistic forms of art.

2.1 K-means

The first clustering method we attempted was k-means, one of the most general and basic clustering methods. With high dimensional data such as images, the use of Singular Value Decomposition prior to the clustering helps reducing our dataset to lower dimensional data while still retaining valuable information, in order to accelerate the clustering process [5]. The 10 singular vectors corresponding to the 10 highest singular values are kept. A value of $k = 10$ was sufficient to keep enough data for the training of the Cycle GAN while still retaining enough closeness between artwork within each cluster. The resulting training test sizes were in the range of 150-250 as the cluster of median size was picked for each genre. Sample clusters can be found in figures 5 and 6 of the appendices.

2.2 Feature Distance

Pairwise distance between each photograph in a genre and a target photo were calculated using feature detection functions available through OpenCV. The SURF feature detector was used to extract features from the individual album cover, both because it was efficient enough to make this method tractable, and because it uses Haar wavelets response distributions [6], which was the most descriptive single feature found in the original paper. L1 (Manhattan distance) distance was selected to match features by brute force, because it was reported to better captures human interpretation of image similarity [7]. Sample clusters can be found

in figure 7 of the appendices.

2.3 SSIM

Structural similarity index is a measurement originally intended grade image quality, by distinguishing between an original photo and slightly distorted alternatives [8]. Pairwise distance between each album in a genre was calculated using the SSIM function available through Sci-Kit Image. While not originally intended to be a clustering method, this produced some interesting results – albums within the broad category of metal that were most easily identified or stereotypically metal was clustered at the low end of SSIM distance. Classical albums most stereotypically appearing to be classical clustered at the high end of SSIM within their greater genre. In each case, the 25 albums with the most dramatic SSIM values of their respective ranges, and the 30 nearest neighbors to those 25 albums were split into training and testing datasets. Sample clusters can be found in figures 8 and 9 of the appendices.

3 Methods and Results

3.1 Cycle-Consistent Adversarial Networks

The Cycle-GAN was trained over the course of 100 epochs, with a learning rate of 0.0002 on the first 50 before decaying it for the last 50. We maintained $\lambda = 10$ as the relative importance of the two goals. The Networks were trained on the RCC using 3 GPUs with a batch size of 12.

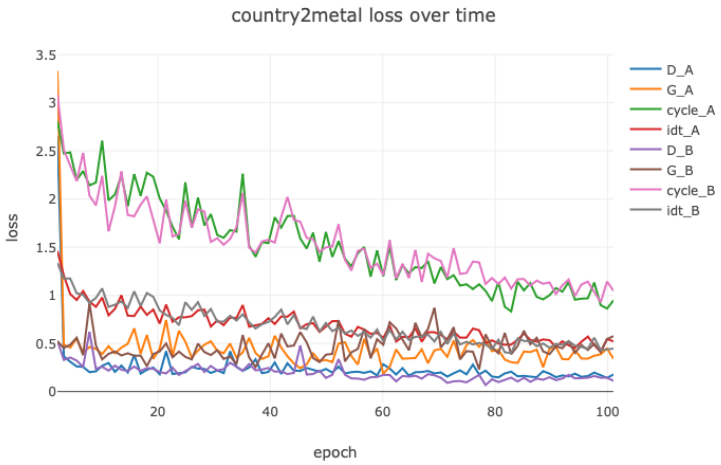

Figure 1: Training of the Cycle-GAN

After trying out each clustering method, the feature image clustering method [1] had mixed

success. The target image for the dataset is shown in the upper left corner. The resulting dataset seems to be a more or less random selection from the larger metal category, with no greater visual similarity than the data pool at large (Figure 7). The SSIM method was not intended to be a clustering tool, but originally was going to be applied as a method of comparing final and input images. It is intended to be used as a metric for evaluating very similar images, or the same image with a small degree of noise. SSIM as used here likely reflects the "busyness" of an image – the most prototypically metal albums are clustered at the low end of SSIM values because the albums have a significant amount of decoration and texture, and so are very structurally dissimilar to one another. Classical albums tend to be somewhat simpler, and so are detected as more similar by the SSIM metric (Figures 8 and 9). Finally, the use of k-means resulted in the strongest translations.

Figure 2: Translation of a Metal album in the Country genre

Figure 3: Translation of a Country album in the Metal genre

As we can see in figures 2 and 3, which were both obtained after training on data made with k-means clustering, the albums translated picked up some of the color and qualitative aspects of each genres: the metal album's dark colors were turned into warmer colors, while the country album's more joyful colors were turned into black.

However, even given these efforts at limiting the variety in our training sets, we found that the cycle-GAN was unsuccessful at producing any meaningful structural change in the image or texture, except for slight changes in color scheme, and slight blurring in the fonts.

3.2 Style transfer using Convolutional Neural Networks

We had much more success performing image-to-image translation using more general style transfer using Convolutional Neural Networks [3]. This method isolates image content from image style in two album covers, and then translates the style of one image to the content of another image.

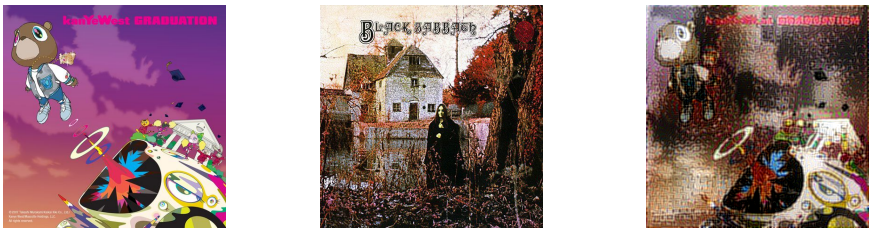

Figure 4: Kanye West's Graduation through Black Sabbath's style

While not capturing the broader trends of style within the music genre, this did make more significant structural changes to the images' style. There was a more drastic change in the rap album being translated through the Rock band's artwork. The album clearly picked up the qualitative features of Black Sabbath's artwork, and through different types of photo transfers, we observed various changes in the structure of the albums.

4 Conclusion

Regardless of the clustering method employed, the cycle-GAN did not produce noticeably different images. We predict that this issue stems from not only the variability of data within a genre, but also the variability between genres. Two types of losses were employed by the cycle-GAN: adversarial loss, and cycle consistency loss. Adversarial loss is the loss function structure typically employed by GANs, in which one network generates albums in the style of the target genre, and the other network distinguishes those generated albums, rejecting ones that do not sufficiently imitate the properties of the target genre. Cycle consistency loss is calculated by mapping the new album back to its original domain – the closer it can get to the original, the smaller the loss. We predict that an interaction between these loss functions caused the failure of the cycle-GAN to change the images. The variability between the two genres was so significant, that any album generated by the adversarial networks could not be mapped back to the original domain. Only very small changes in color or slight blurring of the images could sufficiently pass through both loss functions. Adjusting the λ value so that the relative importance of the generative adversarial network was greater than the cycle-GAN may have improved performance, but it is hard to predict how insignificant the cycle consistency loss could be made before the network was exclusively a GAN.

This paper further emphasizes the importance of dataset construction in the overall success of a computer vision project. A method of clustering the data both within and between genres may produce more significant changes in the final album cover, although by limiting the dataset by this constraint, the final albums may be harder to identify as the target genre.

In the future, it may be interesting to limit the dataset to more common patterns seen on many albums, i.e. translating the fonts used on the album covers from one genre to another, as the content of the images would be more similar. Neural style transfer was a much more successful method of accomplishing our goal – one possible future area of research would be to generalize the extracted style across several sample images, instead of only one.

References

[1] J. Libeks and D. Turnbull, "You can judge an artist by an album cover: Using images for music annotation," *IEEE MultiMedia*, vol. 18, no. 4, pp. 30–37, 2011.

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015.

[4] C. Koenig, "Classifying album genres by album artwork," *GitHub*, 2019.

[5] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Machine Learning*, vol. 56, no. 1, pp. 9–33, 2004.

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision –ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), (Berlin, Heidelberg), pp. 404–417, Springer Berlin Heidelberg, 2006.

[7] P. Sinha, "Face recognition by humans: Nineteen results all computer vision researchers should know about," in *Proceedings of the IEEE*, pp. 1948–1962, 2006.

[8] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

A Clustering samples using k-means

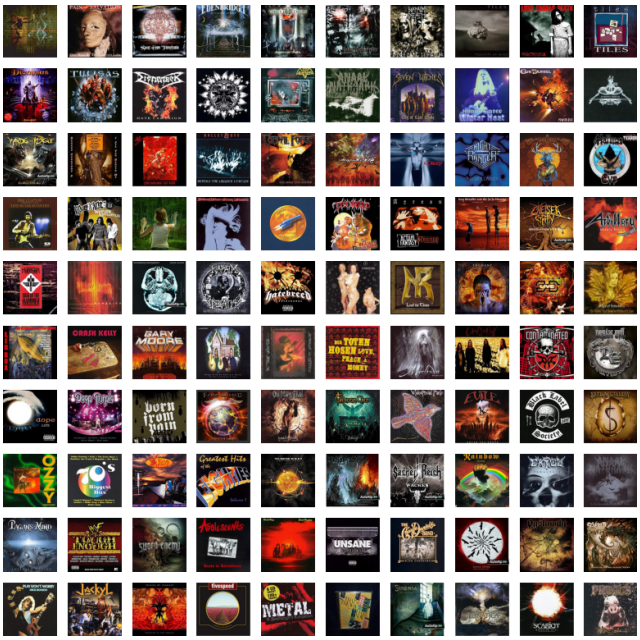

Figure 5: Cluster of Metal Album Artwork

Figure 6: Cluster of Country Album Artwork

B Clustering samples using feature distance method

Figure 7: Cluster of Metal Album Artwork

C Clustering samples using SSIM method

Figure 8: Cluster of Metal Album Artwork

Figure 9: Cluster of Classic Album Artwork