Art Cultural Classification with Deep Neural Networks

Zhenbang Chen
Massachusetts Institute of Technology
zhenbang@mit.edu

Zhenjia Chen
Massachusetts Institute of Technology
zhenjia@mit.edu

Abstract

Art classification has been a longstanding effort for historians and aestheticians alike. Categorizing art is generally done by carefully examining various elements of the artwork's medium, style, and form. These features naturally lend themselves to a computer vision problem, as stylistic elements of images are aptly identified by contemporary models. This paper investigates the use of deep neural networks for determining a painting's geographic and cultural origins.

1. Introduction

Art, specifically paintings, has undoubtedly been shaped by countless artistic movements, regional traditions, and preferences. Some art genres such as realism enjoy a long and illustrated history with many well-known pieces, while other genres such as minimalism are more modern. While art from different movements clearly have unique visual styles, the same can be said about art from different regions. European artworks are generally distinct from American works and even more so from Asian ones. This paper explores the use of neural nets in classifying images into their cultural origins.

2. Related Works

Leveraging the effectiveness of deep learning to investigate research questions in the artistic domain is hardly a novel concept. On a closely related topic, Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger's in-depth research publication *Recognizing Art Style Automatically in painting with deep learning* explores machine learning approaches to art style recognition [5]. They developed a model that labels artwork according to their art movement. Moreover, interest in style transfer and generative image modeling techniques continues to fuel new research at the intersection of creative art, vision, and computation [3][7].

3. Approach and Model

3.1. Data

To explore the research question and analyze the stylistic differences between works of art from distinct backgrounds using geographic location as a detailed example case, our project aims to develop a model that predicts a culture group given a specific piece of artwork. More specifically, while a typical curatorial analysis of an artwork may involve a detailed examination of its medium, known artists, and ownership history, this model will investigate the possibility of categorizing art

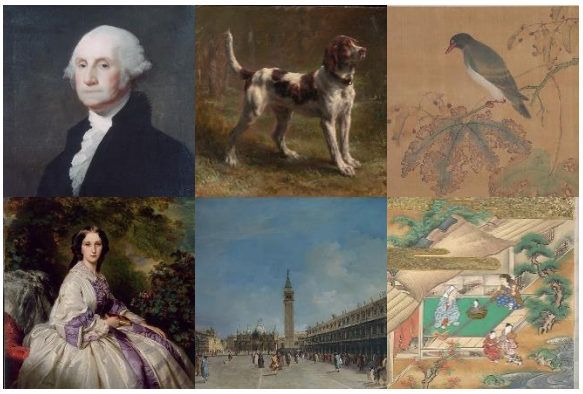

Figure 1. Paintings from the Metropolitan Museum of Art available under the Open Access policy [1]. Culture clockwise from top left: American, French, Chinese, Japanese, Italian, and German.

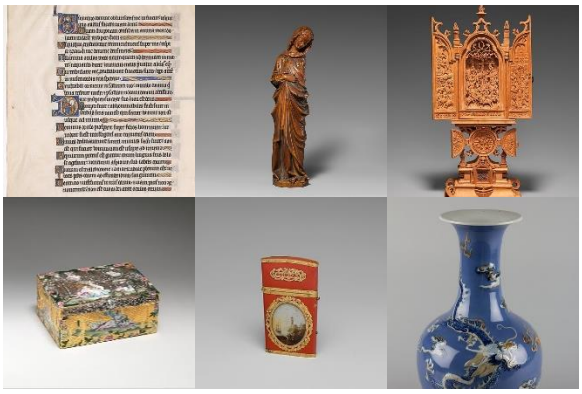

Figure 2. Some artworks in The Met collection were erroneously tagged as "Paintings" and were manually removed from the dataset.

only through a single simple image representation. Classification could involve inferring details including the time period, creators, authenticity, art movement, or origin. For this project, we confined the objective to determining the geographic origin or culture group associated with paintings.

To train our neural network, we decided to work with a dataset gathered from the Metropolitan Museum of Art (The Met). Since 2017, The Met has provided access to over 406,000 images of its public-domain artworks under its new Open Access policy [1]. In this extensive collection, we surprisingly only found 7,938 pieces of artwork that were tagged as "Paintings" as possible candidates for training.

The next step involved partitioning the image set into geographic/cultural categories. Fortunately, The Met collection already tags each piece as originating from any of 125 geographic locations. However, data sparsity for certain tags was a significant concern, and the majority of location tags were only associated with at most a few dozen unique artworks (e.g. only 62 paintings from Africa). Additionally, several tags were redundant and covered non-disjoint regions, such as "India", "Gujarat", and "Rajasthan" or "England" and "United Kingdom". To address these issues, we decided to exclusively consider geographic locations with more than 100 associated artworks and to take the union of collections sets from overlapping regions. This resulted in 11 distinct geographic locations or art cultures for our final categorization task: American, British, Chinese, Dutch, Flemish, French, German, Indian, Italian, Japanese, and Spanish.

The final hurdle in consolidating the dataset involved filtering erroneously tagged images in The Met collection. Despite being tagged as "Paintings", a small, albeit significant, portion of the collection were images of artwork of other mediums, including sculptures, books, manuscripts, decorative containers, jewelry, and even snuff boxes. While it would be an interesting challenge to train the model on a larger collection encompassing all art mediums, we decided to manually remove these image examples since they were relatively rare in the collection and occurred with nonuniform frequency across cultures. After this process, the final artwork dataset for our training task contained 6,663 images across 11 cultural regions.

Culture Group	Initial Images	Final Images
American	1592	1585
Japanese	1149	1084
Chinese	1097	1034
French	940	815
Italian	759	688
Dutch	429	412
Indian	559	372
British	300	280
German	259	172
Spanish	125	115
Flemish	109	106

Table 1. Size of the dataset categorized for each culture from The Met collection before and after removing images erroneously tagged as "Paintings."

3.2. Network Architecture

Upon collecting and processing the necessary training and testing data, we then had to decide on an appropriate neural network architecture. Thankfully, there are already several neural nets specifically designed to perform image classification. As this application is fundamentally an image recognition problem, we decided to iterate on existing neural networks using the machine learning library PyTorch.

Convolutional neural networks like ResNet, VGGNet, and AlexNet are already used to great effect in image recognition and classification [2][4][6]. These networks have many more outputs than needed for our culture classification task. As such, we replaced the last linear layer of these networks with three fully connected layers with ReLU nonlinearities with 11 outputs.

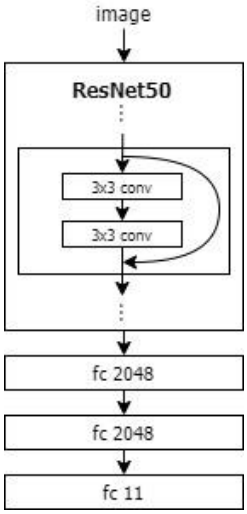

Figure 3. Architecture for transfer learning using ResNet50 as the base.

For each of the neural nets, we trained them on around 4000 examples. ResNet is composed of several residual blocks containing two weighted convolutional layers and a ReLU nonlinearity. These residual blocks are connected together to form a deep network. The specific variant of ResNet we trained with was ResNet50, which is built with 50 layers. This is a smaller version of the typical ResNet152, which makes it faster to train albeit with lower accuracy. We also experimented with AlexNet, a smaller convolutional network with only 8 layers for even faster training times and with VGG-16 (with batch normalization) which has significantly more training parameters despite being shallower than ResNet50.

3.3. Pretrained Models

After the initial training attempts, we experimented with fine-tuning pretrained models for our classification task. As their name suggests, these pretrained models have weights that have already been trained for general image recognition tasks. This means that we would not have to dedicate training

time for the model to learn an optimal image featurization.

In addition to fine tuning the pretrained models, we also implemented transfer learning by freezing the pretrained layers of the ResNet50, VGG-16, and AlexNet models and appending three fully connected layers with ReLU activations functions. While the training time per epoch improved significantly, the validation accuracy was substantially lower when compared to the fine-tuned model even after many epochs.

3.4. Training Process

In our training process, we followed a typical 80-10-10 randomized split of our dataset for training, validation, and testing. For each architecture, we tested multiple scheduler configurations, learning rates, and number of epochs during the training process. We employed a stochastic gradient descent optimizer for each iteration.

Out of the learning rate schedulers, we found that reduction in the learning rate based on a simple step function (StepLR) worked nearly as well as a more dynamic approach based on monitoring the running validation loss (ReduceLROnPlateau). More specifically, the StepLR approach involves reducing the learning rate by a factor γ after every k epochs. This allows for a transition from large updates early in the training process to finer adjustments later as the model approaches the minimum. The dynamic approach follows a similar paradigm. In this case, the learning rate is decreased by a factor γ after waiting for k epochs without a meaningful improvement in the validation loss.

For all the tested architectures (ResNet50, VGG-16, AlexNet) and for both fine-tuning and transfer learning approaches, the dynamic learning rate reduction method yielded marginally better results.

3.5. Visualizing Feature Maps

After training the model to categorize artworks according to their geographic origins, we also investigated ways to visualize what the network was seeing under the surface. This amounted to retrieving the feature maps generated after the convolution layers for the input images.

Working with the model that produced the best results (fine-tuned ResNet50), we started by removing the final few layers which are used for the classification task. This left the main convolutional layers of the base ResNet50 which produces a final (2048×7×7) tensor representation for each input. Since this tensor is then normally fed into the final linear layers for classification, it constitutes a latent representation of the image. Next, we collapsed this representation into a (7×7) tensor by taking a summation along the first dimension with normalization.

Upscaling and overlaying the feature map over the original image produced interesting results. The composite images highlight important parts of the inputs (e.g. people, objects, light/dark regions) that the convolutional layers picked out and passed along to the output. This approach gives some insight into what parts of the paintings influenced the final classification decision.

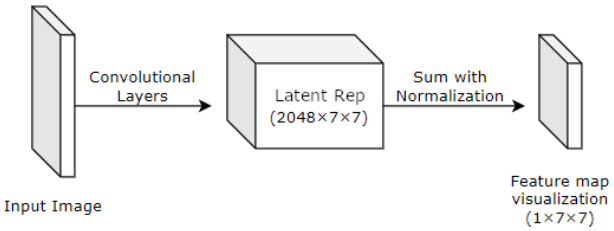

Figure 4. Feature map for each image obtained by condensing the latent representation of the input image after the convolutional layers.

Figure 5. Visualizing feature maps for the best performing ResNet50 model.

4. Evaluation and Results

4.1. Accuracy

For evaluating our proposed models, we tested each neural net on around 600 images and recording top-1 and top-3 accuracies. The testing examples were also sampled from The Met collection. In general, the different models performed with similar testing accuracies, with the exception of non-pretrained ResNet50 which yielded substantially worse results than the fine-tuned networks.

In terms of testing accuracy, ResNet50 with a ReducedLROnPlateau scheduler was the most performant of the proposed networks. With a top-1 and top-3 testing accuracy of 71.79% and 91.84% respectively, it performs slightly better than ResNet50 using a StepLR scheduler and noticeably better than the fine-tuned VGG-16 and AlexNet models. A top-3 accuracy roughly describes how well the model was able to differentiate between Western and Asian art styles.

All four of these fine-tuned networks did better than the transfer learning approach and much better than training the network from scratch. These methods resulted in accuracies of 60.88% and 48.04% respectively, appreciably lower than the fine-tuned neural nets with accuracies above 65%.

4.2. Culture Clustering

Theoretically, a well-trained model would map input images into a latent representation space such that the feature vectors of similar paintings are clustered together. This raises the possibility of computing the "distance" between the visual styles of artwork from different cultures.

Our approach to visualizing this cluster effect involved first featurizing every image in our dataset to a high dimensional vector using our best-performing model. For each culture, we then computed a single vector as the average of all representation vectors with the appropriate ground truth culture label. Using the pairwise L^2 distance between the averaged vectors, we obtained a "distance" that should positively correlate with the relative dissimilarity between the cultures' art styles.

After graphing the results, we found that the distribution of averaged feature vectors qualitatively matched the geographic distribution of the locations in the real world. Additionally, the resulting clusters corresponded to intuitive groups of Asian, European, and North American art. Interestingly, the model produced this geographic layout solely from analyzing visual differences between the regions' paintings. This indicates that the model is extracting relevant features from the images and correctly categorizing the paintings according to their cultural origins.

Architecture	Scheduler	Method	Validation Accuracy	Top-1 Accuracy	Top-3 Accuracy
ResNet50	ReduceOnPlateau	Fine-tuned	69.28	71.79	91.84
ResNet50	StepLR	Fine-tuned	68.28	70.09	91.69
VGG-16 BN	StepLR	Fine-tuned	67.37	68.13	90.63
AlexNet	StepLR	Fine-tuned	68.28	65.71	88.07
ResNet50	StepLR	Transfer	60.88	60.88	86.86
ResNet50	StepLR	Scratch	50.01	48.04	78.55

Table 2. Performances of models with different architectures, schedulers, and methods. All models were trained using an SGD optimizer. Training methods involved either fine tuning a pretrained model (*Fine-tuned*), freezing pretrained layers for transfer learning (*Transfer*), or training from scratch (*Scratch*).

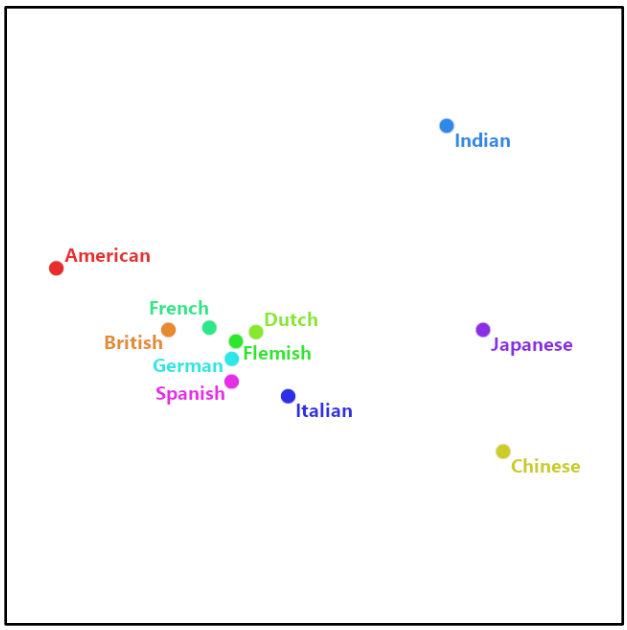

Figure 6. Averaged feature representation vectors from the best performing model for each culture's art plotted with random initialization using the force directed graph drawing package ForceAtlas2. There is a clear differentiation between Asian, European, and North American art styles.

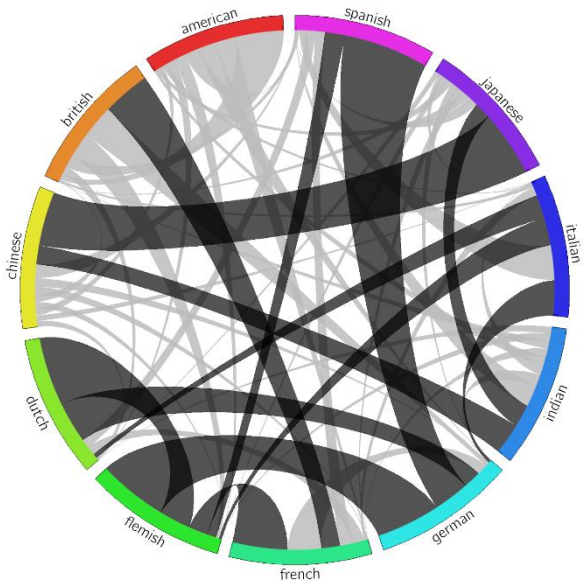

Figure 7. Circos chord diagram highlighting the top quartile of "normalized pairwise similarities" between cultures' art styles. We defined normalized similarity of culture A to culture B as $\frac{\text{distance}(A,B)^{-1}}{\sum_{C \neq A} \text{distance}(A,C)^{-1}}$. Intuitively, this is the ratio of the "gravitational force" between A and B to the total force that A experiences. This value is highest when A and B are relatively close to each other while being far from all other cultures (e.g. China and Japan).

4.3. Limitations

This model of art classification has a few limitations resulting from the lack of sufficient training examples for some of the cultures and the abundance of ambiguous artworks.

Of the approximately seven thousand artworks gathered from The Met collection, only around a hundred were paintings tagged as Flemish. This lack of training examples for these cultures manifested in reduced testing accuracy when presented with paintings from these regions. The model is much more likely to correctly classify an American painting than a Spanish one largely due to the disparity in training data. However, this limitation can be mitigated by simply increasing the amount of training examples. Sampling artwork from other collections not only expands the model's repertoire of paintings, but also exposes the model to more works from more culture groups.

Another limitation of the model also stems from the training data. Several of the pieces were inherently artistically ambiguous as to their origins. For example, in the process of filtering out incorrectly tagged artworks, we decided against removing pages from scriptures that contained a significant amount of nontext elements. This resulted in slightly more similarity between cultures that had considerable amounts of scripture illuminations. This problem is further exacerbated by the relatively small datasets of some cultures where scripture pieces comprised a larger percentage of the training examples.

5. Conclusion

There is great value in applying deep learning and machine vision techniques to questions in the artistic domain. Based on the results from our model, deep convolutional neural networks perform reasonably well at classifying paintings by their geographic and cultural origins.

6. References

[1] "Open Access at The Met." The Metropolitan Museum of Art. Accessed April 2020. https://www.metmuseum.org/about-the-met/policies-and-documents/open-access.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 770-778, 2016.

[3] X. Huang and S. Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. *IEEE International Conference on Computer Vision (ICCV)*, 2017: 1510-1519, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. Image Net Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

[5] A. Lecoutre, B. Negrevergne, F Yger. Recognizing Art Style Automatically in painting with deep learning. *Journal of Machine Learning Research*, 80: 1-17, 2017.

[6] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015: 730-734, 2015.

[7] Lu Sheng, Ziyi Lin, Jing Shao, Xiaogang Wang. Avatar-Net: Multi-Scale Zero-Shot Style Transfer by Feature Decoration. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018: 8242-8250, 2018.