Score Aggregation Techniques for Identifying Van Gogh Paintings with CNN and SVM

Chris Benson **Brendan Whitaker**

Eric Lowry
The Ohio State University
Columbus
{benson.433, whitaker.213, lowry.200}@osu.edu

Abstract

We give a high-level overview of the paper [1] by Folego, Gomes, and Rocha, which tackles the binary classification problem of painter attribution on Van Gogh paintings using the VGG-19 CNN image recognition model. They break each painting up into non-overlapping square patches on which they train the model and generate confidence scores. We propose several new methods for patch score aggregation: the process the authors use to generate a confidence score for a painting from the scores of the individual patches. We make predictions for which of these proposed methods are most likely to be effective using the authors data, and we train our own CNN model from scratch on the patches from their dataset to predict patch authorship.

1 Introduction

To identify the authorship of paintings, experts can narrow down the time period fairly quickly based on the knowledge of styles and the time periods they are from. Thereafter, even after cross-referencing the time period with an artists lifetimes, verified works of art they painted, and artistic familiarity, experts can only debate. To provide more evidence, authentication methods include optical microscopy, UV light, X-ray radiography, and infrared reflectography. These methods are potentially invasive or damaging to the artwork and even then may not succeed in authentication. A safer way to help with this process would be using machine learning. Using code written by Folegos team, we divided the pictures of paintings into patches of normalized size and pixel density. Our team wanted to see if evaluating the painting with different fusion techniques of patch SVM values would provide a more accurate classification.

Techniques we thought to apply were Geometric Mean, midrange, and Strong Mean – details of which will be in the Methods section of this report. We hypothesized that of these three, the Strong Mean would be the most accurate because it compared the means of only the patches most confident at predicting Van Gogh authorship, and disregarded patches with low confidence scores. However, since we were not able to access Caffe Zoo, evaluating our novel fusion techniques in the same way that was originally done was not possible. In resolution, our team created a novel CNN using Keras to create a binary classifier of the patches in order to determine authorship on a patch by patch basis. Using machines rented from Amazon Web Services (AWS) we were able to train the CNN in XXXX hours and achieve XX.XX% testing accuracy after training for X epochs.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Background

In order to address the painting authorship identification problem, researchers from the RECOD Lab at the University of Campinas, explored using CNN and SVM techniques in their paper From Impressionism to Expressionism: Automatically Identifying Van Goghs Paintings. Using publicly available data from Wikimedia Commons, researchers created a novel dataset of images of paintings by Van Gogh and his contemporaries (a Non-Van Gogh set.) Images in this dataset were then normalized to a pixel per inch (PPI) density of 196.3, which aligns with previously done studies [5]. This dataset was named the vgdb_2016 dataset, and consisted of 124 images authored by Van Gogh and 207 images authored by his contemporaries. After the novel vgdb_2016 database was normalized over PPI, the dataset was divided into a training and testing sets, with the training and testing sets consisting of 80% and 20% of the dataset respectively; this is summarized below in Figure 1:

	Training		***Test***	
Class	**Images**	**Patches**	**Images**	**Patches**
van Gogh	99	15,895	25	3,927
non-van Gogh	165	31,513	42	8,611
Total	264	47,408	67	12,568

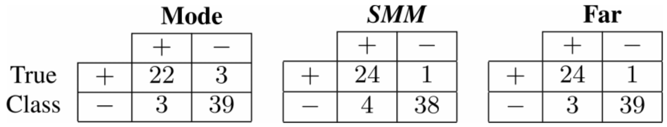
Figure 1: Summary of the `vgdb_2016` dataset divided into testing and training subsets.

Researchers then began the process of extracting features in order to build a classifier for images. Extracting features from an entire image at a time would be extremely computationally demanding, and lead to a representation model with potentially billions of parameters. In order to overcome this difficulty, researchers divided images into non overlapping patches sized 224 by 224 pixels. If an image did not evenly divide into patches of this size, then the borders of the image were discarded. This solution is favorable to resizing all of the images to the same size, as that approach would not overcome the computational burdens of extracting features of an entire image at once, along with disrupting the normalized image density of the dataset. Individual patches are much less computationally demanding to extract features from, however, fully training a CNN from scratch accurately requires more data than the dataset provides. This is overcome by using a pre trained CNN that excels at feature extraction from images. The selected CNN was the Visual Geometry Group, University of Oxford network (VGG19), which was trained on the ImageNet dataset that consists of over 1.3 million photos. Using this pre trained CNN, features were able to be extracted from the patches[6].

After extracting features from both the Van Gogh (VG) and Non-Van Gogh (NVG) patches, classification was done using a SVM with a linear kernel. SVM was the chosen classifier as SVM has been shown to excel with data with high dimension, however, any type of classifier could have been used at this point. For each patch, the SVM produced a distance from a separating hyperplane from -1 to 1, which can be interpreted as a confidence score with -1 representing complete confidence the patch is NVG, 1 complete confidence the patch is VG, and 0 complete lack of confidence either way.

From here, the scores from each patch had to be aggregated with other patches from the same originally image in order to create a holistic classification of the original image. Multiple fusion techniques were tested at this point. The Mode technique classified each image based off of whether the image had more patches with positive(VG) or negative(NVG) scores; this aggregation technique does not take into account the relative confidence expressed by each patch. The Sum technique simply added all of the patches scores together and evaluated the image based on this overall confidence sum. The Mean and Median techniques took the mean or median of all the positive patches in an image, and compared it with the mean or median, respectively, of all the negative patches in an image. Finally, the Far method simply compared the most positively scored patch with the most negatively scored patch to make a classification.

Of the fusion techniques described above, the Mode technique was the least accurate at 88%, the Sum, Median, and Mean (SMM) methods all were 90.6% accurate, and the Far technique was the



		Mode			*SMM*			Far	
		+	−		+	−		+	−
True	+	22	3	+	24	1	+	24	1
Class	−	3	39	−	4	38	−	3	39

Figure 2: Summary of the results of the original fusion techniques in [1].

most accurate at 92.3%. These percentages were calculated by testing the model over the 67 images in the testing set. These results are summarized below in Figure 2.

This paper aims to expand upon these original fusion techniques and determine if more targeted fusion techniques can more accurately determine painting authorship.

3 Methods

The three fusion techniques we wanted to implement were Geometric Mean, Midrange, and Strong Mean. In Folegos research, the Far, Mean, and Median fusion techniques involved separating the negative and positive scores and running operations on them separately and then comparing the two results. This is important because we did not want the number of positive and negative scores consistently affect the classification when it is possible that it is not a significant parameter. Our techniques followed this same structure.

The Geometric Mean technique calculated the geometric mean of the positive scores, the geometric mean of the negative scores, and classified based on the more significant of the two values. (reference)

The Midrange technique calculated the average between the minimum and maximum of the positive scores, the same for the negative scores, and classified based on the more significant of the two values. (reference)

The Strong Mean technique took the median of the positive and negative scores and calculated the average of the scores above and below the positive and negative medians respectively. The classification would be based on the more significant of the two values. (reference)

4 Experiments

The dataset provided by the authors of [1] consists of 331 `.png` files of paintings by either van Gogh or one of his contemporaries. They are split up into training and test datasets, and then patches are generated using the `patch_extraction.py` utility provided in [2]. The distribution is given above in Figure 1. We follow the method outlined by Folego in [2] to preprocess the given dataset.

5 Discussion

Although we were not able to confirm the accuracy of our novel fusion techniques, we hypothesis that the Strong Mean method would be the most likely to improve upon the accuracy of the Far method that was originally used. Of the four images that the Far method misclassified in the original testing, researchers observed that the magnitudes of the scores with the most confidence that an image was VG or NVG were similar. The scores of the four misclassified images are as follows in the format of (Max VG Score, Max NVG Score): (99.7%,98.3%), (99.95%,97.01%), (99.4%,96.9%), and (99.92%,99.97%). All of these are extremely confident scores, and the mean difference between the confidence of these scores from the misclassified images is 1.7225%. Our hypothesis is that by taking the average of the patches with confidence levels above the medians of the respective VG and NVG patches, we would be able to compute mean scores of only the most confident patches in the dataset. By disregarding patches that express low levels of confidence, we would be able to emulate the fact that the Far method only compared the most confident patches, however by taking the mean of multiple above- average confident patches, this method would be less susceptible to outlier patches, which may have been the cause of the Far methods misclassifications.

Regarding the CNN that we trained, this model has value due to how simple it was to create, along with the relatively small computational power that is required to train it. The CNN we created is a binary classifier that predicts the authorship of individual patches as either VG or NVG. Using this model, one could reimplement the Mode fusion technique from the original paper, by seeing if an image had more VG or NVG patches. While the Mode method was the least accurate fusion technique in the original paper, potentially due to the fact that it didn't incorporate individual patch confidence in its aggregation, it was still 88% accurate. By using our CNN, one could create a classifier of similar quality to that of the Mode technique. This is important as the CNN we trained was much less complex than the VGG19, therefore classifiers that are relatively accurate at predicting painting authorship can be created using simpler components.

6 Conclusions and Future Work

Clearly, modern machine learning and image processing techniques have the potential to contribute greatly as a non invasive solution to identifying authorship of controversial paintings. By using CNNs to extract features and creating classification models based off of SVM for authors with sufficiently large bodies of work, classifiers could be created for many famous authors. These classifiers could then be used to assist and supplant art experts when new paintings are discovered that may be historically significant.

For future work, the techniques described in this paper should be applied to more authors to see if similar accuracy results can be achieved. Additionally, different classifiers other than SVM could be tested to see if any result in better accuracy. Also, one could explore the effects of varying pixel density of normalized image datasets, along with adding horizontal and vertical flips to the datasets. Finally, future work could be done in creating Mode based classifiers from binary classifiers of patch authorship, in order to bring machine learning support to those who may not have the resources to create models as complex as those described in the original paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can go over 8 pages as long as the subsequent ones contain *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.