Detection of Hokusai's Gaifu Kaisei Replicas and Visualization of Regions of Interest

Natasha Ong
Department of Computer Science
Stanford University
natashao@stanford.edu

Abstract

Recently, computer vision techniques have been applied to image analysis of art to better understand artistic choices, context of the art, etc. In this paper, we explore applications of art forgery to woodblock prints – an art form quite unique to East Asia. Specifically, we use a CNN to detect original and replica prints for Hokusai's Gaifu Kaisei (ie Red Fuji). We found that the model is able to distinguish originals from replicas with a relatively high accuracy of 92%. Moreover, we attempt to visualize and better understand rules and features of prints used by the model in its classification (real/fake) of these prints.

Problem Statement and Contributions

As Deep Learning and Computer Vision rises, many have found new applications for the technology, including art analysis. Some researchers have gone to create models that can detect art forgery in paintings by examining brush strokes, etc. However, to the best of our knowledge, very few have attempted to detect forgery or replicas outside the general realm of paintings. Thus, we try to create a model – as a proof of concept– that replica detection can be applied to other artistic forms, specifically Japanese woodblock prints. Woodblock prints pose an interesting problem as there are many aspects to consider, including color of the print, patterns of the wood, natural wearing/deterioration of the wood from frequent use, etc. These are just some qualities differentiate this project from those done in the past on paintings.

In this project, we focus on Katsushika Hokusai's Gaifu Kaisei, also known as Red Fuji. Hokusai and Gaifu Kaisei were chosen for their popularity and beauty of the art. In other words, the more popular an artist or an artwork is, the more likely it is for its original prints to be housed in museums and its replicas to be created and distributed. These ability to obtain the digital images significantly contribute to the potential success of the model as neural networks require lots of examples for training.

In addition to creating a detector, we wish to create a visualization of regions or aspects a model finds important in determining whether a print is real or fake. In other words, we seek to open up black-box models and come up with explanations for classifications. Doing so will interest and benefit art historians and connoisseurs, who will have a better understanding of what to look for in distinguishing the real and fake. This work can also lead to other questions and analyses regarding the regions identified.

Moreover, this project and its work can be used by the general population, specifically people who wish to purchase prints of Gaifu Kaisei. In particular, the detector will be integral in ensuring that people do not get tricked into paying more than what their art/print is actually worth.

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

fake / replica real

Figure 1: Sample Prints with their Respective Labels

Related Work

Art Forgery

The problem of detecting art forgery has been explored by many others in recent years, though there are none to our awareness that tackle the specific application of woodblock prints. In fact, most research in this area seem to focus on paintings and drawings. Polatkan et al., for example, approached art authentication through a supervised machine learning lens, where they applied such techniques on features derived from Hidden Markov-Tree-modeling [8].

Meanwhile, others have taken different approaches, zooming in to specific aspects of an artwork. In other words, instead of training a model on the entire artwork, these researchers focus on local features or descriptors [3][4]. Researchers detecting forgery through brustroke analysis, for instance, have created algorithms capable of segmenting individual strokes and quantifying stroke characteristics [5].

Model Explainers and Visualizers

As machine learning models increase in prevalence and complexity, many researchers have attempted to better understand features a model uses in making a decision, especially in classification tasks. In the case of CNN models (which include art analysis along with general image analysis), researchers have explored several different methods of explaining a model. These include separating layers of a model and backtracing to produce a decision tree [9][7], using feature importance [7], using various gradients and propagations like Kullback-Leibler divergence gradient [1] or Layer-wise Relevance Propagation (LRP) [2][6][9].

Many of these methods are complicated and non-intuitive (as models generally are), so they involve training an entirely separate model that is used to open the black box of the model of interest.

Dataset

No dataset currently exists of Hokusai originals and replicas. The dataset used in this project was constructed by scraping the web. More specifically, Google Arts and Culture and several museum websites – both in Japan and abroad – were used in building the "original/real" dataset. For the "replica/fake" dataset, digital images of cheap duplicates which are sold online were used. These images come from specific sellers such as Adachi or Takamizawa in Japan or more general e-commerce websites like Etsy, Amazon, or Ebay.

This dataset consists of 68 total images, split into 36 originals and 32 replicas. Each image is of different sizes, and pre-processing was applied to remove extra borders that surrounded the prints. Moreover, we ensured that the relative dimensions of images were identical to the original print (ie the dimensions are factors of one another). Then, these images were scaled down to 400px by 400px and fed to the model.

For the original/real dataset, the 36 images were split as follows: 25 training, 7 validation. For the replica dataset, the 32 images were split as follows: 29 training, 7 validation.

Methods

CNN Architecture

We use a Convolutional Neural Network (CNN) architecture to learn features of original and replica images that allow for the distinction between the two. As the problem is to distinguish real from fake, our ultimate result is a binary classifier.

The model summary is as followsS:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_293 (Conv2D)          (None, 398, 398, 16)      448
_________________________________________________________________
max_pooling2d_293 (MaxPoolin (None, 199, 199, 16)      0
_________________________________________________________________
conv2d_294 (Conv2D)          (None, 197, 197, 32)      4640
_________________________________________________________________
max_pooling2d_294 (MaxPoolin (None, 98, 98, 32)        0
_________________________________________________________________
conv2d_295 (Conv2D)          (None, 96, 96, 64)        18496
_________________________________________________________________
max_pooling2d_295 (MaxPoolin (None, 48, 48, 64)        0
_________________________________________________________________
flatten_63 (Flatten)         (None, 147456)            0
_________________________________________________________________
dense_126 (Dense)            (None, 512)               75497984
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_127 (Dense)            (None, 1)                 513
=================================================================
Total params: 75,522,081
Trainable params: 75,522,081
Non-trainable params: 0
_________________________________________________________________
```

We train the model using RMSProp optimization with $lr = 1e-3, \rho = 0.9, \beta = 0.0, \epsilon = 1e-07$, metric as accuracy, and loss function as binary cross entropy (since there are only two label classes). Due to the small dataset, the model was trained for 10 epochs, with 5 steps per epoch, such that the following equation was satisfied:

$$\textbf{total training examples} = \textbf{batch size} \cdot \textbf{steps-per-epoch}$$

To prevent overfitting, which is a promiment issue in situations with limited training data, a number of measures were taken.

First, the number of layers of the model was reduced. Initially, the model was too complex model– the ratio of hidden neurons compared to the number of samples available in training was too high. Moreover, dropout was added to reduce overfitting or over-reliance.

Hyperparameter Tuning

Most of the hyper-parameter tuning had to do with ways to reduce overfitting and increasing the accuracy. Factors such as number of epochs, steps-per-epoch, input image size, dropout, number of layers, etc. were explored.

CNN Hyperparemeters	Value
Number of Epochs	10
Steps-per-epoch	5
Batch Size	10
Learning Rate	0.001
Optimizer	RMSProp
Input Image Dimensions	(400, 400)
Dropout	0.5

Approximating Explanations of CNN Classification

In addition to creating and training a classifier, we sought to understand what the model relies on in distinguishing original Gaifu Kaisei prints from replicas. For this project, we do so using SHAP – Shapley Additive Explanations. SHAP uses a game theoretic approach to explain the output of any machine learning model, in our case a CNN.

Shapley equations attempt to linearize components such as max, softmax, products, divisions, etc to better understand aspects the model focuses on. In particular, for our core explainer, we use a Deep Explainer, which is designed to approximate SHAP values for deep learning models. It works by approximating conditional expectations of SHAP values using a selection of background samples. Since our dataset is small, we use all of the training examples that we have to help SHAP approximate these values.

With the Deep Explainer, SHAP values are approximated such that they sum up to the difference between the expected model output on the passed background samples and the current model output: $(f(x) - E[f(x)])$. We then test it on the validation set to produce colored visualizations (red and blue) that are overlayed on the original print. The locations of the overlayed colors indicate regions of interest. The visualizations can be understood as follows: Red pixels increase the model's output while blue pixels decrease the output. The less transparent the color the more strong an association it creates.

As such, by looking for any overlayed colouring and their respective colours, we can better identify what which regions the model relies on to decide what label to assign.

Results

Replica Detection

The model produced satisfactory results, with an accuracy of approximately 92.9%. See below for figures of loss and accuracy. We see that the loss and accuracy graphs look as we expect it to. Losses for both training and validation set decrease while the accuracy of both increases (or plateaus). From the shape of the curve, we can conclude that the model is indeed learning (vs randomly outputting labels) and is not significantly overfitting. Our training accuracy is slightly higher than our validation accuracy, which may point to slight overfitting or different distributions of images in the validation set. These are both possible due to the limited dataset we had.

CNN Explanation: Visualizing Regions of Interest

Given our model works well in detecting replicas, we proceeded to create visualizations of regions of interest. (We can trivially see why doing this second task when the model is inaccurate is quite pointless...). After using SHAP and training it on the entire training set, we separate results of the visualization based by labels. Figure 3 provides the visualization of a subset of correctly labelled originals. Figure 4 provides a visualization of a subset of correctly labelled replicas.

Based on the clustering and opacity of the colors, we can conclude that the model heavily relies on the peak of Mount Fuji to make its decisions. From the images, we can see that in the original images, the peaks of the mountains are marked blue while in the replicas, they are marked red. We can actually spot the outlier – the one image the model predicted wrong – on Figure 5.

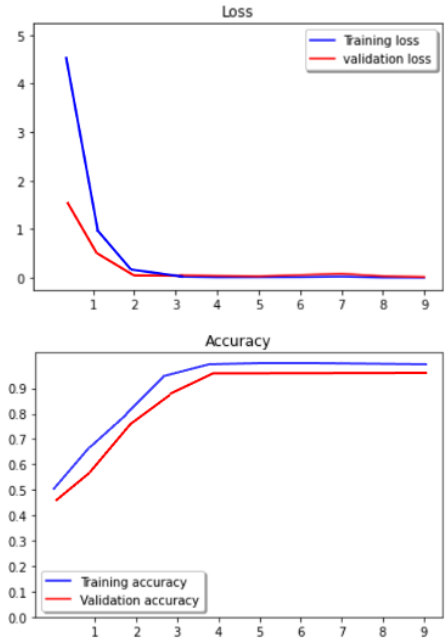

Figure 2: Loss and Accuracy of the CNN

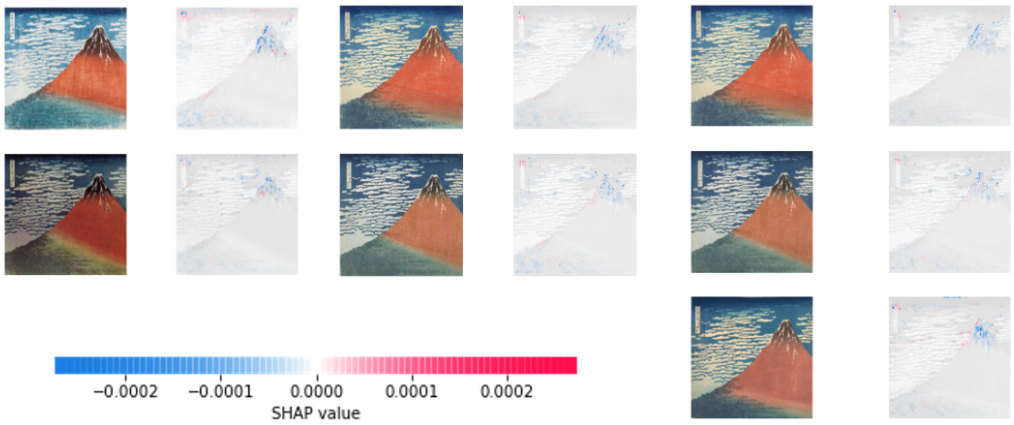

Figure 3: Visualization of Originals

Moreover, to a lesser extent, the coloring informs us that the model relies on the surrounding clouds in the "background" and also the forests in the bottom left corner of the prints, especially for the replicas.

Limitations and Future Work

To summarize, the CNN performs very well in distinguishing original and fake Hokusai's Gaifu Kaisei woodblock prints, achieving accuracies of over 90%. From visualizations of the model and its "rules", we notice that the peak of Mount Fuji is most used to distinguish originals from replicas. Less frequently or extensively, the model relies on the background skies/clouds and the forrest locations (bottom left half) of the print.

One of the biggest limitations of this project stems from the small dataset, which caused the validation sample to be limited in both number and scope. In the future, should more images and data of

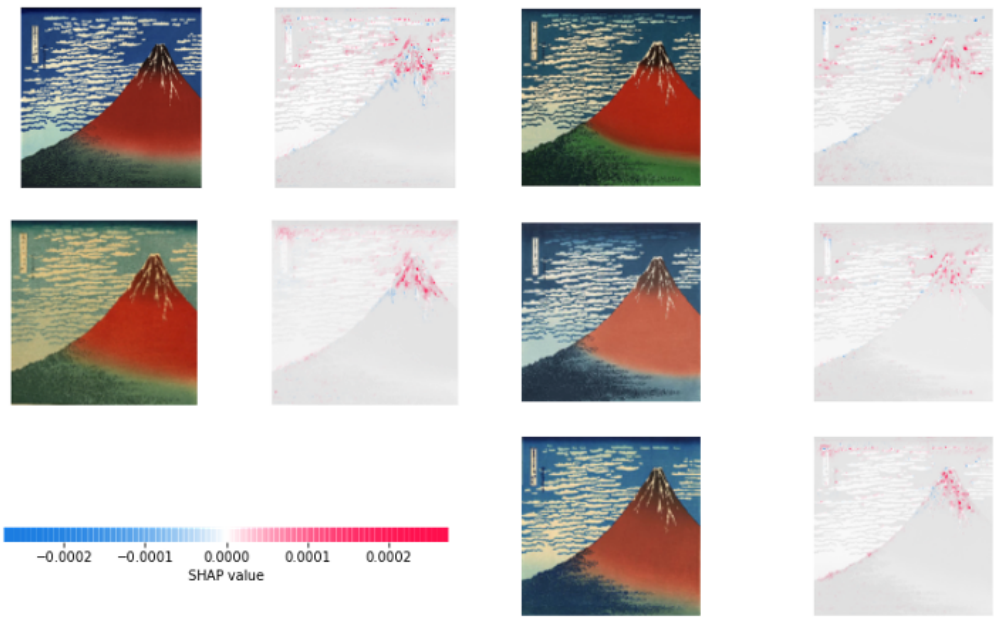

Figure 4: Visualization of Replicas

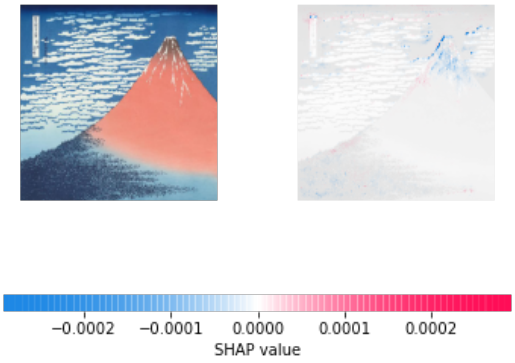

Figure 5: Replica Mislabeled as Original

Hokusai's Gaifu Kaisei be available (both originals and replicas), one might be able to create a more robust model.

Another limitation is related to the visualizations of regions of interest. We use the Shapley Additive Explanations as it is a high-speed approximation algorithm. Other visualizations, like those mentioned under the Related Works section was not explored, mainly due to time restraints in completing this project. Nevertheless, attempts at understanding or visualizing relevant areas contributing to the final classification is very new, so as newer research is done and more literature is produced, they are guaranteed to shape the future potential of this work and project.

Other potential future work involve scaling up this project. In this project, we deal with one artwork by one artist in one time-period. We can try to generalize this technique to various other kinds of forms or art or even generalize it to an artist more generally.

References

[1] Babiker, H. K. B., & Goebel, R. (2017). *Using KL-divergence to focus deep visual explanation.* arXiv preprint arXiv:1711.06431.

[2] Bologna, Guido. "A Simple Convolutional Neural Network with Rule Extraction." Applied Sciences 9.12 (2019): 2411.

[3] D. Cozzolino, D. Gragnaniello and L. Verdoliva, "Image forgery detection through residual-based local descriptors and block-matching," 2014 IEEE International Conference on Image Processing (ICIP), Paris, 2014, pp. 5297-5301, doi: 10.1109/ICIP.2014.7026072.

[4] Cozzolino, Davide, Diego Gragnaniello, and Luisa Verdoliva. "Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques." 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014.

[5] Elgammal, A., Kang, Y., & Leeuw, M. D. (2017). Picasso, matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication. arXiv preprint arXiv:1711.03536.

[6] Lapuschkin, S.; Binder, A.; Montavon, G.; Muller, K.R.; Samek, W. Analyzing classifiers: Fisher vectors and deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2912–2920.

[7] Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; Giannotti, F. Local rule-based explanations of black box decision systems. arXiv 2018, arXiv:1805.10820.

[8] Polatkan, G., Jafarpour, S., Brasoveanu, A., Hughes, S., & Daubechies, I. (2009, November). Detection of forgery in paintings using supervised learning. In 2009 16th IEEE Internation

[9] Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering, 19(1), 27-39.