Using Convolutional Neural Networks to Predict Completion Year of Fine Art Paintings

Blake Howell
Stanford University
450 Serra Mall, Stanford, CA 94305
bhowell@stanford.edu

Abstract

This paper investigates the feasibility of using a convolutional neural network to predict the year of completion of a fine art painting. We confirmed the feasibility of this problem by training a network that achieves a 48% accuracy classifying a test set into 5 different 25 year periods between the years of 1875 and 2000. The approach taken to achieve this was to take a pre trained model designed to classify ImageNet images, and reset both fully connected layers and several of the deepest convolutional layers with the goal of learning high level feature representations that are more useful to the task of art classification than ImageNet classification.

1. Introduction

In this paper, we investigate the feasibility of using a convolutional neural network (CNN) to predict the year of completion of a fine art painting. The network will accept an RGB image of varying dimension sizes and will output a prediction for the year it was completed. The prediction will take the form of a categorization of completion year into pre-chosen groupings.

This network could be used for dating newly discovered works of art, and for discovering trends in the way artworks have evolved over time. Hopefully, we can use back propagation to generate paintings that are exemplary of a given time period. Furthermore, we would attempt to analyze the weights of the network to discern intelligible trends in the evolution of features of fine art paintings over different time periods.

2. Related Work

Given that no work has been found that has attempted to categorize art by completion date before, there is very little in the way of directly related work. This paper attempts to demonstrate that this task is feasible.

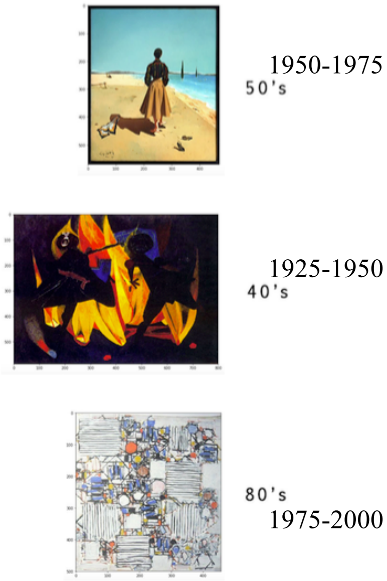


Figure 1: Input to output examples.

That being said, Babak and Saleh [1] demonstrate respectable performance in the categorization of fine art paintings into different styles. They do not, however, take a deep learning approach. Also, they do not aim to predict the completion year of each work.

Dumoulin and Shlens [2] use deep learning to capture artistic style across different paintings by interpreting style as visual texture that can be recognized by existing networks.

3. Architecture

We intend to use a CNN to predict painting completion year. A CNN is a form of neural networks containing convolutional layers, which slide a filter over all regions of the image and output an activation for each region.

After getting poor performance training a CNN from scratch, we chose to tackle this problem by applying the concept of transfer learning to a pre-trained VGG11 [3] architecture. As the shallowest of the VGG networks, we chose this one as it should be easier to train with a small to moderate sized dataset.

3.1 VGG Architecture

The VGG network passes an image through a stack of convolutional layers, which use filters with a 3x3 pixel receptive field. The convolution stride is fixed to 1 pixel. The spatial padding of convolutional layer input is 1 pixel, which ensures that the spatial resolution is preserved after convolution. Spatial pooling is carried out by five max-pooling layers interspersed throughout the network. Max-pooling is performed over a 2 × 2 pixel window, with stride 2. Following the convolutional layers, there are three Fully-Connected layers: the first two have 4096 channels each, the third contains 5 channels (one for each class). The final layer is the soft-max layer. All hidden layers are equipped with the ReLU non-linearity.

After getting poor performance by retuning only the fully connected layers, we went further by resetting the fully connected layers and the final three convolutional layers to the initialisation used in the original training of the VGG network, with weights sampled from a normal distribution with the zero mean and 1e−2 variance. The biases were initialised with zero.

The motivation behind this was that all of the low-level feature detection happens in the earlier layers. After the third layer, the features are heavily tuned to detect objects from the initial ImageNet dataset. Resetting these layers provides an initialisation which promotes the harnessing of existing low level features to develop of new higher level features useful to the task of art classification rather than preserving the existing feature representations.

4. Training

The network was trained using the same hyperparameters as the original VGG network.

The loss function was chosen to be the cross entropy function.

The weights were optimised using mini batch stochastic gradient descent with a momentum of 0.9.

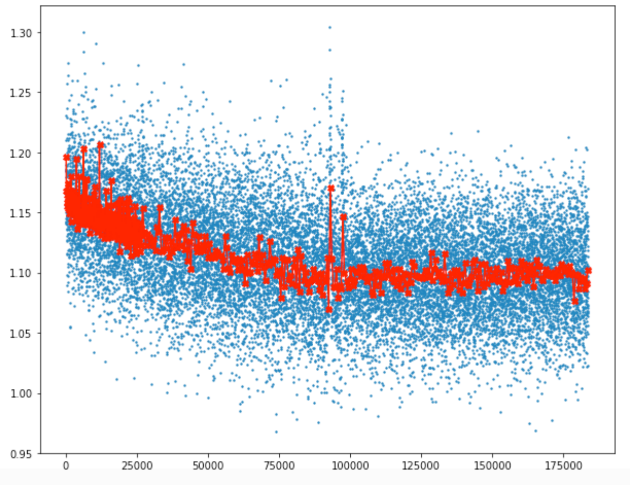
The batch size during training was set to 64. While this was the largest batch size we could fit while training on only one GPU, it was still smaller than the VGG's original training batch size of 256, which likely made for more stochastic convergence behaviour.

Training was regularised by weight decay, the L2 penalty multiplier was set to 5e-4.

The learning rate was initially set to 1e-2. The original paper opted to decreased the learning rate by a factor of 10 each time the validation accuracy stopped improving. This, however, had virtually no beneficial effect on the convergence of our network.

These hyperparameter choices made for very smooth, monotonically-decreasing training loss curves with no virtually no overfitting.

Figure 1: Training loss curve convergence.



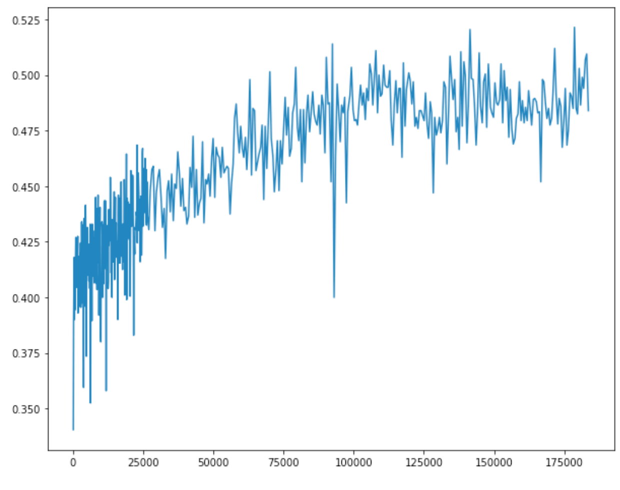


Figure 2: Training accuracy convergence.

We trained the network for a total of 183,120 iterations over 168 epochs. This is significantly less than the 370K iterations (74 epochs) that the original VGG network was trained over. This is likely because the lower level features had already been trained in our network.

5. Dataset

The dataset we are using to train the CNN is the 'WikiArt' collection [4]. This collection is publicly available dataset of digitised fine art pieces spanning fifteen centuries. Below is a histogram showing the dates of all items in the collection.

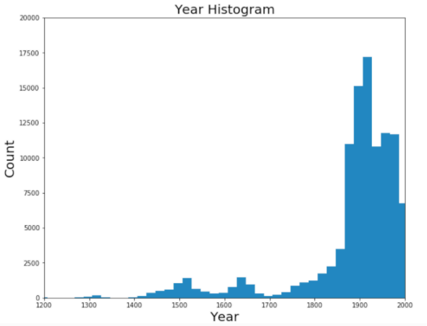


Figure 3: Data Set Year Histogram

Given the disproportionate amount of data located between the years 1875 and 2000, we chose to bound the buckets between these years. After this bounding, the total number of samples was just under 80,000.

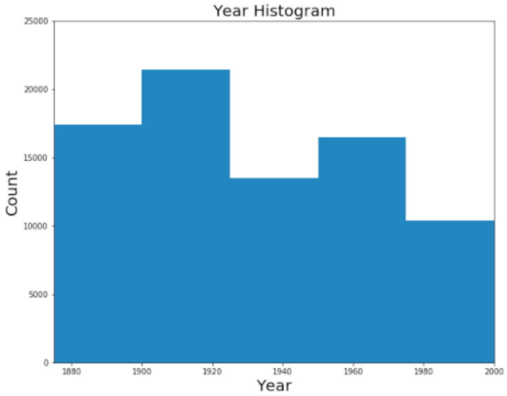


Figure: Distribution of training data in chosen buckets.

The above shows the distribution of the buckets chosen for this model. Previous experimentation suggested that the network had little luck identifying the bucket for any buckets whose counts were smaller than 5000. This bucketisation was chosen as our final choice as it ensured all buckets were comfortably larger than 5000 in number. Similarly, this choice of discretisation offered one of most uniformly distributed bucket sizes.

In spite of this, training networks without correcting for even the more modest class imbalance present in the chosen bucket configuration resulted in poor performance, with the network biasing heavily towards predictions of the two most frequent categories in the training set. To correct this, we experimented with two forms of correction: weighted sampling and weighted loss.

The below figure shows the results when only weighted sampling was used. We used weighted sampling to oversample from classes that were under reperesented in the data set.

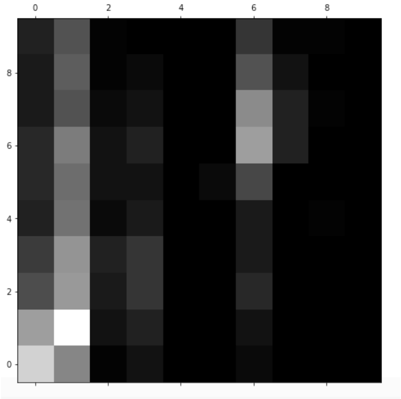


Figure 4: weighted sampling confusion matrix

The below figure shows the results when both weighted sampling and weighted loss were used. We used weighted loss to penalize wrong predictions of the less represented classes.

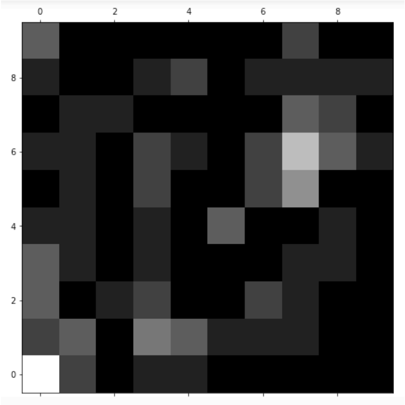


Figure 5: weighted sampling and weighted loss confusion matrix

From these results, we hypothesized that weighted loss leads to a better spread among predictions, though the network still is unable to learn much without sufficiently large class buckets.

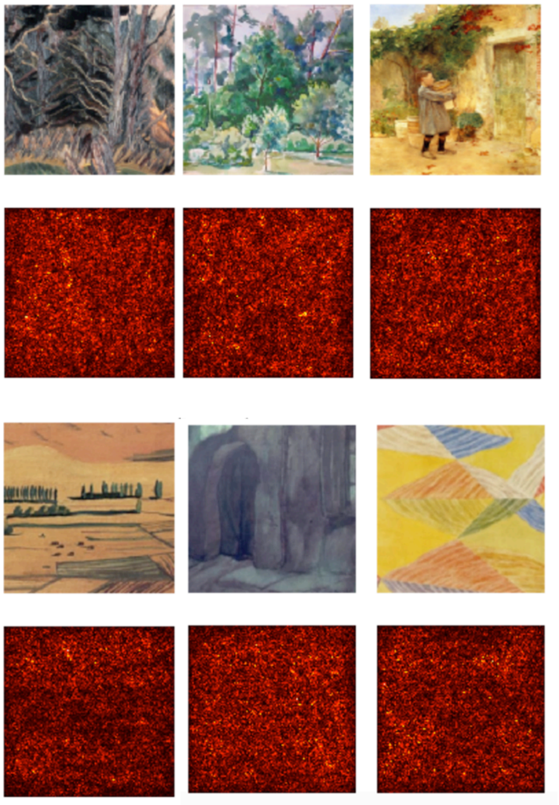
5.1 Image preprocessing

Each piece is of varying dimension sizes. Thus, before feeding each painting into the network, work needs to be done to standardize the dimension sizes. The primary ways to achieve this are cropping and padding. Cropping subsamples a fixed rectangle from each image.

For the sake of preserving maximum resolution within the 224x224 window, we scaled each art work to take the maximum sized square crop, discarding any information that fell outside of the square region. As with the original VGG network, the mean was then subtracted and the variance was standardized.

5.2 Data augmentation

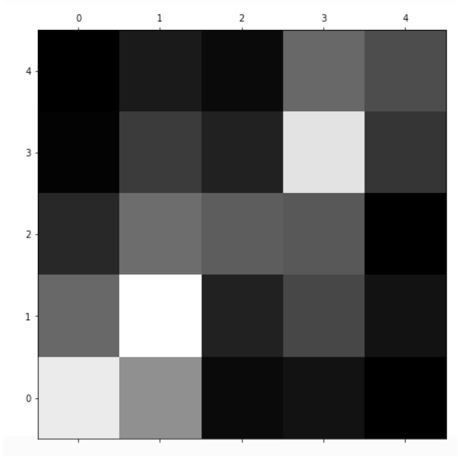
Our training process utilised very little data augmentation. As can be shown from saliency maps produced by our final network, categorisation of artwork by year requires an "all-over" examination of each image. Data augmentation techniques that would be suitable for categorisation of images based on the presence of objects within a certain region of an image (such as random cropping and rotation) would do much to corrupt the semantics of a piece of art.



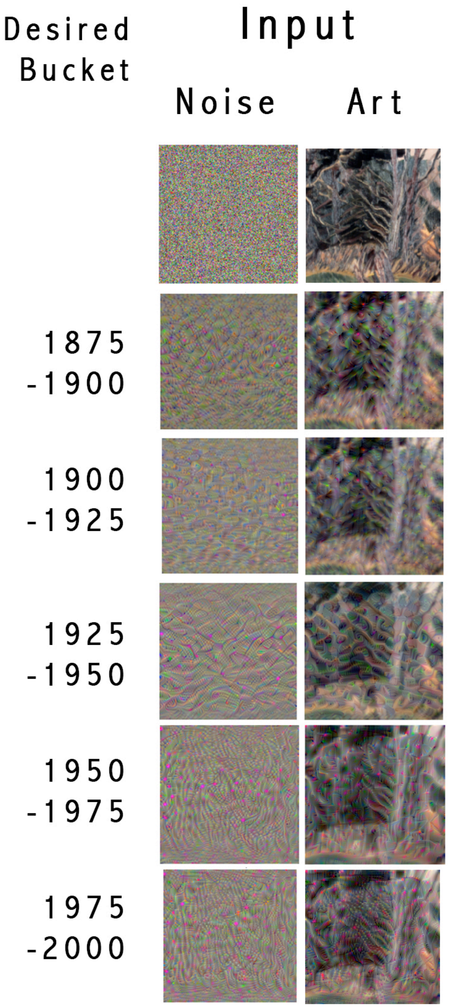
The one data augmentation technique we did utilise was horizontal flipping, as this preserves framing and symmetry semantics, which we deemed to be essential for successful categorisation.

6. Results

Our network achieved an accuracy of 48.63% on a test set of 5000 samples. A confusion matrix is shown below.



The current performance of our network suggests that our hypothesis that weighted sampling from buckets that are sufficiently large enough would correct for the biases of class imbalance was incorrect. For future work, we would retrain our network using both weighted sampling and weighted loss.



The above figure demonstrates class visualization techniques on each of the completion year buckets. The left column shows the results when the input image is randomly sampled noise. The right column shows the results when a piece from the training set is randomly sampled. The results for each year category show semantically distinct textures and patterns when viewed up close. These changes are not particularly decipherable however.

7. Conclusion

This paper has achieved the goal set out to demonstrate the feasibility of predicting completion year of pieces of fine art. With a test accuracy of 48.63%, we have demonstrated performance way above the 20% accuracy that would be achieved with random guessing. Given the difficulty of the problem present at its very core by virtue of the variance present in art, the network has demonstrated that it has captured a significant level of understanding of art by achieving these results.

7.1 Future Work

There is still a lot of work than can be done to explore the structure generated by the model we have trained. The first area for future research we would recommend is using Deep Dream to visualize the features that this network has been trained to recognize. Deep Dream takes the feature activations at each layer and backpropagates them to update the input image. By varying the layer from which activations are being backpropagated, it is possible to accentuate the features that the network thinks is present in certain regions of the input. Given that we have retrained not only the fully connected layers, but also the final three convolutional layers, it is very likely that the network has leveraged the existing low level VGG features to create higher level feature representations that are more relevant to art than to ImageNet classification. Using Deep Dream would be a way of exploring what these higher level features look like. A similar way of doing this would be to analyse activation maps at each layer and visualise the regions that activate each filter the most strongly.

Another area of research to be carried out for analysis of our current model would be feature space projection. One could use dimensionality reduction techniques to visualize training examples that are proximate in feature space. This could give an effective means of visualizing the semantic similarities picked up by the network's highest level features. Similarly, performing this kind of visualization in tandem with visually representing each sample's ground truth completion date could give an indication of just how hard the problem of classifying art based on date is. For example, if training examples from similar time periods demonstrate little clustering in feature space, that would give an indication that the level of variance present in the problem makes the problem less tractable.

Next, there are many ways to investigate ways of improving performance. Our first recommendation would be to extend the depth of the network as was done in the original VGG network. This would result in a network with a richer feature representation without the difficulties of training a deep network from scratch.

Similarly, one could experiment with clearing different layers of the original VGG network as a means of improving performance. It is unclear that clearing the final three layers was the optimal choice for training. Different choices may lead to very different performance.

Our final suggestion for leveraging the concept of transfer learning to improve performance would be to iteratively retrain the network to produce progressively more and more fine grained categorization. We would hypothesize that by initially training the network to perform binary classification, the network would converge on the most salient points of distinction and would have very large bucket sizes with which to distinguish features. Ultimately the goal would be to achieve a regression model with reasonable performance. We hypothesize that sigmoid regression would be most effective. In this regime, the network would output a value between 0 and 1 which would be scaled and shifted to output a year within a specified range.

There are also a great deal of hyperparameters to choose from with regards to image preprocessing.

Taking the max crop from each image risks discarding important features. Instead, we could explore padding input images. Each image would be scaled to fit inside a square, with the left over regions colored in. The color used for the background may be a hyperparameter to be tuned.

The choice of input dimension is also a hyper parameter that could be tuned in future experiments. 224 x 224 was chosen to match the input dimensions of the original ImageNet input to the VGG network. Increasing the resolution of input images has the potential to greatly increase network performance.

Another possible area of research to attempt to improve performance would be the use of ensemble models. This would be very likely to improve accuracy, though it is not possible to analyse the features of ensemble methods in the same way as individual networks, which would reduce the possible insights we could gain about the evolution of the presence of certain features across different historical periods.

Finally, it may be possible to improve accuracy by narrowing the problem. This could be achieved by training networks to classify pieces from a particular group based on year. This group could be a particular school of art or art from a particular region of the world. With a narrower population to sample from, it may be easier to filter out the salient details with less variance in the training data.

References

[1] Babak, Saleh: “Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature”, 2015; arXiv:1505.00855.

[2] Vincent Dumoulin, Jonathon Shlens: “A Learned Representation For Artistic Style”, 2016; arXiv:1610.07629.

[3] Karen Simonyan: “Very Deep Convolutional Networks for Large-Scale Image Recognition”, 2014; arXiv:1409.1556.

[4] https://www.wikiart.org/

[5] Code from CS 231N A2 was used.

[6] Code from CS 231N A3 was used.

[7] Code from Tutorial for data loading and fine-tuning used.