The Effect of Art Identification with Convolutional Neural Networks According to The Difference in The Dataset

Or Ben-Gal
Tel Aviv University orbengal@mail.tau.ac.il

Abstract

Style identification of known works of art is a challenging task even for experts and skilled curators. The vast art style and artworks make it difficult mission for humans, however for a computer it's a different case. My paper takes the work done in the article "Artist Identification with Convolutional Neural Networks"[1] of Nitin Viswanathan, Stanford University. In the paper they are classifying artist of a given artwork. The main problem in the paper is the amount of training data, 300 per artist. As a result, the difference between shallow CNN and deeper CNN are not noticeable. In this paper I am identifying art style from the same data set, which have more data per class. I train the same Convolutional Neural Networks on different data set with a goal of seeing the effect on the accuracy of the CNNs, and to measure the effects from different types of datasets. My dataset consists of at list 1,600 paintings per style from 27 art style and total of 100,000 approximately. I train the models as in "Artist Identification with Convolutional Neural Networks" [1], a simple CNN designed from scratch and ResNet-18 network. My best results demonstrate a bigger difference between the shallower and the deeper CNNs. And has higher classification accuracy than the control dataset.
In the process I preform experiments on a few data set to learn and understand the important properties of a good dataset.
My results demonstrate that deep CNNs are a powerful tool for art style classification and there is yet a lot to explore in that path.

1. Introduction

Art Style identification is the task of identifying the style of a painting given with no other information about it. This is an important requirement for cataloguing art, especially as art is increasingly digitized. One of the most vast and diverse datasets, WikiArt, has around 250,000 artworks from over 200 different art styles by 3,000 artists [7]. There are many websites whose purpose is to digitize works of art and make them accessible to the general public. As these collections grow, it becomes demanded to be able to efficiently label and identify newly digitized art pieces, and maybe even identify unknown work of arts. With the advancement of CNN classifying and identifying we also will be able to identify forgeries, A problem that exists to this day, and no effective solution has yet been found. However, CNNs have already transcended humans in such assignments.

Art identification traditionally performed by art historians and curators who have expertise and familiarity with different artists and styles of art. The complexity of this task and the reason it's a difficult challenge comes from the facts that specific artist with verity of painting techniques can belong to more than one style of art, furthermore style of painting consist of dozens of Subdivisions of styles and fashions, so that each class has no clear characteristics that belong to all members of the class.

Figure 1. Two images of the same style. Illustrate the difference that can exist within the class itself.

The previous work has attempted to identify artists by a given work of art, in this experiment I would like to emphasize the roll of the dataset in identifying

works of art, which are detailed and in many cases complex type of images.

Art style is the most diverse data to identify in the realm of art because all I said above.

In my experiment I set a control dataset to be used as reference, because in my research I would like to see the effect of changes in the dataset, therefore the results of [1], in compare to mine cannot isolate the dataset factor alone.

the reference dataset has the same properties as the dataset in [1].

My hypothesis is the fundamental principal of machine learning of using a vast amount of data to train the models, will be shown in this experiment, and we will see that to some degree expanding amount of data will improve the network accuracy. The question that remain open is, do we expand in all cost, when it comes to art identification?

I believe that original data have advantage over data that was made from data augmentation.

My results show that increasing indeed better the accuracy, and it is preferable to use original data, however If there is no access to original information it is recommended to use data augmentation.

2. Related Work

As I mentioned previously, art identification has executed in the paper *"Artist Identification with Convolutional Neural Networks"*. In this work, the identification was limited to 17,000 images size dataset, for the reason it identifies artists. I changed the identification goal so that I will have more useful data from the same data source.

Additionally [9] prior work has also explored using features that are more specific to art, such as identifying distinct brushstrokes, this work is primarily done in the context of style identification in oppose to [1].

In the past decade, CNNs have made great progress on many image recognitions tasks, for example achieving a top-5 classification error of 3.6% on the ImageNet dataset [3]. This accuracy is higher than human performance on the same dataset [4].

When using neural network on art, many applications achieved for example creating fake images using GAN that have the same properties of the dataset it was trained on to such a level that no difference can be discerned. When using CNN on art to decompose a painting into style and content components and to transfer style from one painting to another, implying that a neural network can capture the style of paintings. However, this work does not explicitly identify or label the style or artist.

3. Dataset

3.1. Overview

To train CNN to identify art style, I needed a large dataset divided to art styles. First, I obtain a large dataset of art compiled by Kaggle [10] that is based on the WikiArt dataset [7]. This dataset contains roughly 100,000 paintings and 19 useable art styles which consist more than 1,000 artworks. According to the missing artworks, I have obtained and completed data set of 27 styles from WikiArt with 100,00 usable images of fine art.

Figure 2. Sample images from this dataset

The smallest class has 1,592 images of artworks and the largest class has 13,060. In total I have 99,167 images of artworks from 27 different styles, with many resolutions by 3,000 artists spanning a variety of time periods. Figures 1 and 2 have some sample images from this dataset.

In this experiment I use four datasets built from the dataset I mentioned above. Every dataset has separated .csv file, and every image is labeled with its style. I split every dataset into training, validation and test sets in

ratio 80-10-10 respectively. This dataset is significantly larger than those used in prior works.

The first will be called "control" dataset will be the reference dataset. Because [1] use artist classification the results of this paper will not serve me well for finding the difference between the other datasets in compare to it. This dataset has the same properties ad in [1], 300 and little more (still balanced) images per class.
The second dataset, we will refer to him as the "small" dataset, is the smallest except to the control, the data balanced contain 1,600-1,700 per class, total of 45,509 images.
The third dataset, we will refer to him as the "large" dataset, is the largest, the dataset is unbalanced I use all the images that I have the smallest class has 1,592 images and the largest class has 13,060 with total of 99,167 images.
The fourth dataset, we will refer to him as the "synthetic" dataset, I took the small dataset and increase the number of images per class with data augmentation to 3,200-3,400 per class with total of 91,018.

The purpose of creating four data set is to determine if there is an significant improvement for a bigger dataset, and if so is it preferable to use unbalanced dataset to increase it significantly or is it better to use data augmentation to expand the dataset in balanced way. In general, I increase the dataset twice. The first time to the "small" dataset from the control dataset. And in the second time I increase it to the large or the synthetic which have similar number of images and different properties.

The control dataset emphasizes the use of less data just like in [1]. The need for this dataset arose when I try repeatedly to implement the results of [1], and realize this is not the same task. The "control" dataset has 300 images per class and total of 8,100. In [1] where also approximately 300 images per class but were 57 class so he had 17,000 in total.

3.2. Preprocessing and Data Augmentation

Because the art in the dataset comes in a variety of shapes and sizes, I modify the images before passing them into our CNNs. First, I zero-center all the images. Next, I center-crop them to 448x448, and then I take a 224x224 random-crop of each input image and normalize them. I always take a 224x224. center crop of the image to ensure stable and reproducible results, and then to add variety and uncertainty I use the random crop on smaller area.

I have experiment with resize and crop transforms, and the latter performed better and gave more stable results. I hypothesize that to determine style, it is important to preserve the minute details that might be lost with rescaling and resizing.

The datasets are big and diverse. I didn't use data additional data augmentation on the datasets, except from the synthetic data set, because its creation involves data augmentation. For the creation of the "synthetic" dataset I use Torchvision [6].

I used random horizontal flip, random vertical flip, random grayscale and random rotation. All these actions were done with probability of 0.5, so on the one hand we will have double size dataset which is balanced and on the other hand we will add an uncertainty to the dataset.

4. Methods

I built and train two different CNN architectures to identify art styles. Every network takes as input a 3x224x224 RGB image and outputs the scores for each of the 27 art styles present in my datasets.

In [1] there are three networks. The additional one is ResNet-18 pre-trained on ImageNet. This model does not fulfill the goal of my experiment, because it already had trained weights that trained on a significantly larger dataset, ImageNet [4]. Therefore, I decided not to use this model.

For all the networks, I use a SoftMax classifier with cross-entropy loss:

$$L_i = -\log(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}})$$

Where L_i is the loss for *i-th* example in the training minibatch, f is the score for a particular class calculated by the network, *j* is one of the possible classes (of the 27 art styles), and y_i is the correct class for example *i*. This loss function ensures that the network is constantly trying to maximize the score of the correct artists of its

training examples relative to other artists during training.

4.1. Baseline CNN

I train a simple CNN from scratch for art style identification. I took the architecture from [1], and you can see it in Table 1. As the name implies, this network serves as a baseline for comparison with the other model. The network is relatively shallow compared to the other network. The network is obviously don't feet to identify complex images, and it's all purpose is to be a reference point to the deeper model. Every layer in the network down-samples the image by a small factor to reduce computational complexity. The downside of this model is that it will not analyze the image small details.

Input size	Layer
3x224x224	3x3 CONV, stride 2, padding 1
32x112x112	2x2 Maxpool
32x56x56	3x3 CONV, stride 2, padding 1
32x28x28	2x2 Maxpool
1x6272	Fully connected
1x228	Fully connected

Table 1. Baseline CNN architecture (ReLU and batch normalization layers omitted for readability) as used in [1].

4.2. ResNet-18

The next network is based on the ResNet-18 architecture from [6]. ResNet is a residual neural network that use shortcut connections, shortcut connections are those skipping one or more layers that can be seen on the ResNet-18 architecture diagram on figure 3. The ResNet I am using has 18 layers and set with a new fully connected layer to allow for right number of classes predictions, just as in [1]. ResNets use residual blocks to ensure that upstream gradients are propagated to lower network layers, aiding in optimization convergence in deep networks [2]. This model trained from scratch to allow the network to learn features solely for the purpose of style identification. The architecture can be seen in Figure 4. The 18-layer ResNet version, which it's the shallowest version in the Torchvision lobrary, allow to train ResNet with reduced the memory requirements.

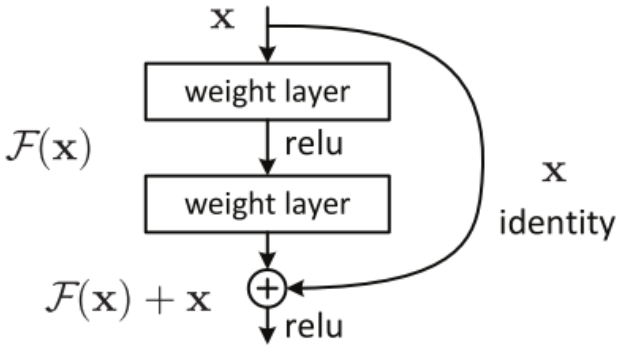

Figure 3. Residual learning: a building block [*].

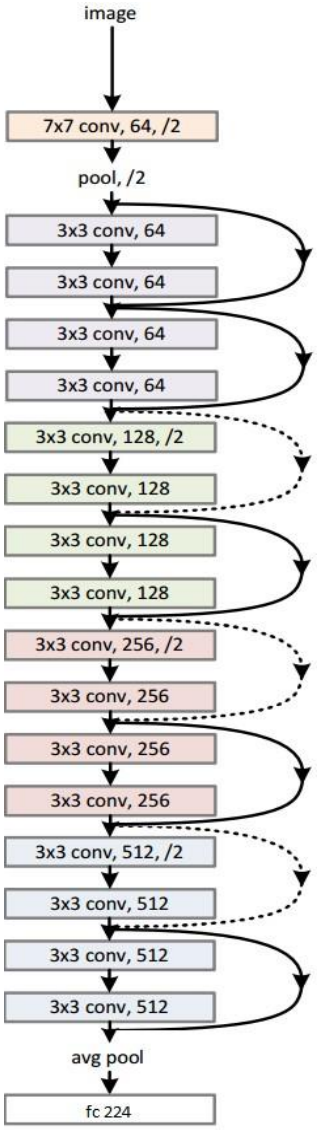

Figure 4. Resnet-18 with a fully connected layer for identification.

5. Experiments

5.1. Setup

All the models and experiments are implemented in PyTorch [5]. I used [1] models as part of my experiment and the architecture if the Baseline CNN. I used Torchvision [6] to set up the ResNet-18 architecture and added a fully connected layer. I built from scratch the Baseline CNN according to the specification in [1], and I programmed the ResNet-18 models with fully connected layer to match the number of class in this experiment, in oppose to 1000 out channel that come with the model. The projects files are available on [9]. All experiments were performed on computer using NVIDIA RTX 2080 GPU.

5.2. Implementation Details

I trained all my models using an Adam update rule. For the two networks trained from scratch, I started with the default Adam parameters of learning rate = 10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. I observed the accuracy and loss for the training and the accuracy for validation datasets over the training epochs and decreased the learning rate by a factor of 10 if improvement slows down significantly.

I experienced with Stochastic Gradient Descent (SGD) with varying parameters on all the models, and the Adam Optimizer performed better, so I performed this step using the same default Adam parameters described previously.

5.3. Evaluation Metrics

Using the scores generated by the networks, I report top-1 classification accuracy, precision, recall, and F1 scores and also top-3, which considers a painting correctly classified if the correct style is in the top 3 highest scores generated by a network., Just as used in [1].

my goal is to evaluate the effect of increasing and changing the data. I compare the Datasets against each other, over all models to see the effect it has on a shallow CNN and deep CNN. Because I use unbalanced dataset as well as balanced dataset, I use the precision, recall and F1 scores.

Precision and recall are defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F1 score, a weighted average of precision and recall, is defined as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The precision score is all the cases per class which correctly identify in relation to all class identification. the recall score is the relation between correct identification of class and all the times the item from the group did appear the F1 score is the harmonic mean of precision and recall, therefore it mainly affected by the small values. due to use of unbalanced dataset, these score methods allow me to better determine the quality of performance of all the dataset.

6. Results and Discussion

In the experiment I had executed eight different combinations between networks and datasets. The two networks, Baseline CNN and ResNet-18, trained on each of the four datasets, "control", "small", "large" and "synthetic".

The results of the experiment divided to table per dataset, so we can the effects better against the compared dataset. Table 2 is the result for the control and the three table after are for the small, large and synthetic respectively.

Worth mentioning the fact that the accuracy I have obtained from the networks are less then [1], yet it's a different assignment and we need to remember that the goal of this experiment is to find the relative effects on the accuracy of the network when given a different dataset in its properties.

Control Dataset						
Model	Top-1					Top-3
	Train Acc	Test Acc	F1	Precision	Recall	
Baseline CNN	26.726	26.049	0.244	0.228	0.262	48.642
ResNet-18	29.773	30.741	0.278	0.251	0.312	51.605

Table 2. Art style identification with control dataset results summary.

Small Dataset						
Model	Top-1					Top-3
	Train Acc	Test Acc	F1	Precision	Recall	
Baseline CNN	26.023	25.797	0.25	0.239	0.261	50.868
ResNet-18	31.033	30.15	0.274	0.251	0.303	55.307

Table 3. Art style identification with small dataset results summary.

Large Dataset						
Model	Top-1					Top-3
	Train Acc	Test Acc	F1	Precision	Recall	
Baseline CNN	30.844	32.016	0.228	0.196	0.272	56.196
ResNet-18	37.703	34.066	0.248	0.232	0.266	59.53

Table 4. Art style identification with large dataset results summary.

Synthetic Dataset						
Model	Top-1					Top-3
	Train Acc	Test Acc	F1	Precision	Recall	
Baseline CNN	28.487	28.356	0.278	0.273	0.283	54.768
ResNet-18	35.314	31.74	0.28	0.249	0.32	51.011

Table 5. Art style identification with synthetic dataset results summary.

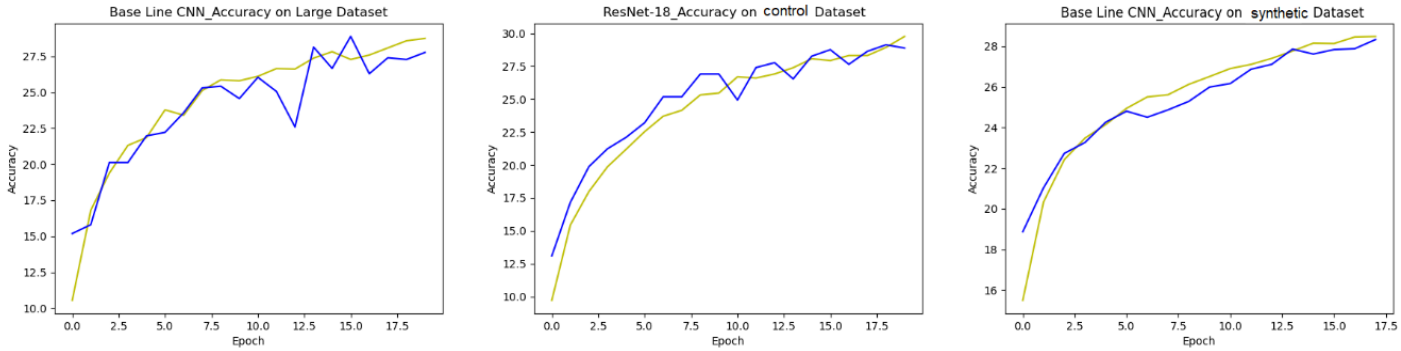

Figure 5. Selected accuracy of training & validation by epoch (train- yellow, validation-blue).

6.1. Quantitative Analysis

We can see in tables below the results of every dataset compare the accuracy of the different CNNs across the key metrics. When we compare between the control dataset results and the small dataset results of the Baseline network, we can notice that there is no increase of top-1 accuracy, however there is a little increase in precision score and therefore in F1 score and top and increase in 5% in the top-3 accuracy, meaning that the Baseline CNN that trained on the small dataset is indeed little more accurate.

With the same datasets on the ResNet-18, we can notice increase in top-1 and top-3 accuracy, but F1, precision and recall score are the same.

Next let's look at the comparison between the small dataset results and the large dataset results. For the Baseline CNN there is an increase in top-1 accuracy of more than 15% from the small and the control datasets, indicating correlation between increasing amount of data and accuracy of CNN. Still we can observe on a decrease in the F1, precision and recall scores, expected from an unbalanced dataset. The ResNet-18 kept increasing its accuracy, top-1 close to 20%, top-3 by 7% from the small dataset and by 15% from the control dataset. Although decreased in the F1, precision and recall, again as expected for unbalanced dataset.

When comparing the small dataset results and the synthetic dataset results of the Baseline network we can see small increase in top-1 accuracy, and more significant increase in top-3, it tells us that the network still improving and indeed getting better, despite its simplicity. And the less expected change is change in F1, precision and recall increase in 12% from the small dataset score.

When applying the ResNet-18 on the synthetic dataset we observe an increase of 14% in top-1 accuracy from the small dataset and about the same F1, precision and recall score. However, we can notice a decrease in top-3 accuracy.

The difference in the score indicates that training on more data proves the accuracy, but the quality of the accuracy reduced when the dataset is not balanced.

In all datasets the ResNet-18 network performed better than the Baseline network.

From the results we can see that ResNet-18 network needs large amount of data to be efficient for image classification.

All the networks applied on dataset for 20 epochs as in [1]. After executing this experiment and witness the results it may be needed to increase the number of epochs to get to a better classifier with the ResNet-18 network, which is not my goal in this experiment.

6.2. Qualitative Analysis

For both network we can see the top-1 accuracy correlate with the amount of data the network trained on as well as the top-3 accuracy, except for the ResNet-18 network trained on synthetic dataset. Such a phenomenon can result from the complexity of the ResNet-18. The Baseline network has 2 convolution layers, while the ResNet-18 has 17 convolution layer and three times deeper. The ResNet-18 can analyze small details in these complex images and the data augmentation didn't play the roll of adding diversity to the synthetic data set.

When we look at the F1 score of the ResNet-18 network we can notice it stays the same when the data is balanced, for the control, small and synthetic datasets, and decrease when the dataset is unbalanced, for the large dataset.

It seems that the Baseline does not improve as the ResNet-18 when expanding the dataset. Only after increasing to the large dataset, more than 10 times the size of the control dataset, and more than twice the size of the small dataset, we see significant increase in accuracy.

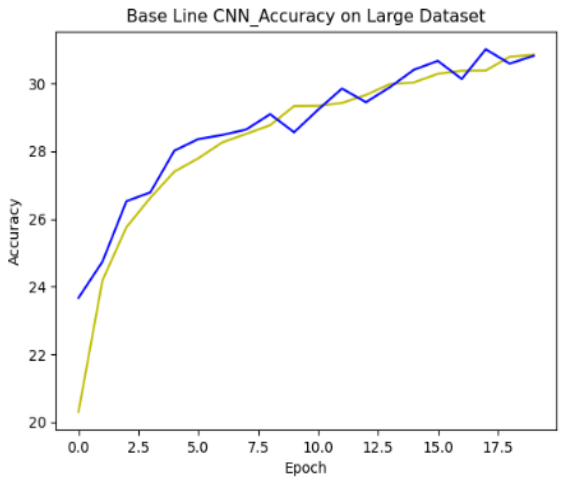

Figure 6. Accuracy by epoch Baseline network on large dataset

We can see in figure * that on the large dataset the Baseline network validation accuracy had unstable growth, as well as the ResNet-18 on the control dataset. Which can occur from big learning rate, or insufficient data. for the ResnNet-18 the validation portion of the control dataset was too small, and therefore the growth in accuracy is not monotonic.

We can see that all the train accuracy graphs on all dataset are growing monotonically.

Furthermore, the most monotonic growth of the Baseline network is on the synthetic dataset. Maybe due to the size of the dataset.

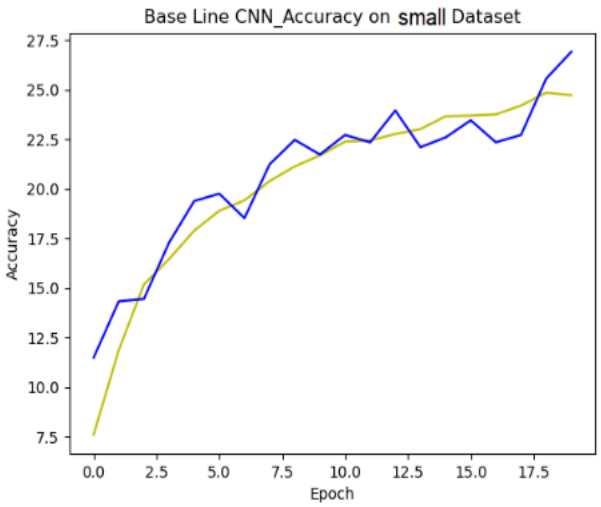

Figure 7. Accuracy by epoch Baseline network on small dataset

We can see from the top-1 results, that there is no overfitting apart of the Resnet-18 on the large dataset and on the synthetic dataset, when the synthetic difference in the test set and the train set is larger deviation relatively, a 13% on the synthetic dataset, and 9% on the large dataset.

When looking at top-3 classification accuracy, we see that all networks perform better than their top-1 accuracies, this is pretty trivial, but that might indicate that for more epochs there will be a significant growth in top-1 accuracy.

7. Conclusion

I introduce the problem of art identification, and art style identification as function of dataset type in particular and applied training of a shallow and a relatively deep CNN architectures to demonstrate the effects of the type of dataset on art classification and on image classification in general, which has not been done in prior work.

My best result was when I applied the ResNet-18 model on the large dataset and it was significantly higher for top-1 and top-3 accuracy.

From the fact that the Resnet-18 performed better on the large dataset than on the synthetic dataset we can deduced that the level of diversity of data augmentation is less than we hypothesize. A picture that create by data augmentation does not amplify training as new original photo. We can witness it also from the highest overfitting of the synthetic dataset. The large dataset is highly unbalanced and still the synthetic dataset overfitted by 13% and the large dataset overfitted by 9%, when applied on them the ResMet-18.

The accuracy results were pretty low, and I believe that training the model on more epochs will get a significantly better the accuracy results.

Some of the pictures in the dataset are very detailed with unique curves and lines.

Figure 8. An example of a painting with un usual lines.

In the experiment I showed a positive correlation between the depth of the network and the accuracy scores of the network.

Deeper networks would do a better job classifying these complex images with higher accuracy than the ResNet-18.

The models I have trained were limited by relatively shallow architectures that might not fit to classify complex images as some of the painting.

For future works, I would like to explore the world of deeper neural networks and see the difference between effects datasets have on deeper ResNet, VGG etc. I also would like to run them on a large number of epochs trying to find the top accuracy. And after that maybe explore with aiding data such gradient layer etc.

The success of analyze painting with CNN will may open a new realm of understanding in analyzing complex images and maybe to understanding art better.

References

[1] "Artist Identification with Convolutional Neural Networks" of Nitin Viswanathan, Stanford University http://cs231n.stanford.edu/reports/2017/pdfs/406.pdf

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint*, abs/1512.03385, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint*, arXiv:1502.01852, 2015.

[4] http://www.image-net.org/

[5] PyTorch. https://github.com/pytorch.

[6] Torchvision. https://github.com/pytorch/vision.

[7] WikiArt. https://www.wikiart.org/en/about.

[8] E. H. J. Li, L. Yao and J. Z. Wang. Rhythmic brushstrokes distinguish van gogh from his contemporaries: Findings via automated brushstroke extractions. IEEE Trans. Pattern Anal. Mach. Intell., 2012.

[9] https://github.com/Bengal1/The-Effect-of-Art-Identification-with-Convolutional-Neural-Networks-According-to-The-Difference-in-T

[10] https://www.kaggle.com