Classic art classifier

Chen Yu-zhen
The Hong Kong University of Science and Technology
ychenhq@connect.ust.hk

December 14, 2024

Abstract

As the digitization of art continues to expand, the categorization of oil paintings is becoming increasingly important. The diverse styles of artworks present a challenge in capturing both their overarching characteristics and fine details simultaneously. To address this issue, this project presents a Classic Art Classifier designed to automatically identify and categorize artwork styles and predict artists from given paintings using Convolutional Neural Networks (CNNs) and transformers. This project is composed of mainly 2 models, a Xception+DenseNet model to classify artists, and a vision transformer model that classifies styles. The classic art classifier is able to recognize top-k artists' paintings and its styles on a relatively small dataset, containing around 6,700 images and 20 artists. Ultimately, the model is able to achieve a 80% accuracy in predicting styles of a painting and a 30% accuracy in predicting the corresponding artist.

1 Introduction

Alongside the focus on material well-being, there is a growing interest in nurturing the inner spiritual world. Art, particularly classical paintings, is being appreciated and collected by a growing audience. However, a key challenge remains in accurately classifying the authors of these artworks. In this paper, multiple CNN-based models were reconstructed with additional dense layers following the convolutional layers, fine-tuned, and tested with the objective of identifying artists and their corresponding styles. These models were designed to learn and recognize the nuanced characteristics across diverse artistic styles, enabling them to accurately predict both the author of an anonymous painting and its associated style.

2 Related Work

Several studies and projects have explored the dataset titled *Best Artworks of All Time*, which is a collection of paintings from the 50 most influential artists. However, most of these works primarily focus on tasks such as style transfer rather than artist classification.

One notable online notebook attempted to classify artists using DenseNet and other pretrained models. However, the dataset was improperly split, with significant overlap between the training and validation sets, leading to biased results. Despite reporting a 90% accuracy for the top 10 artists, the reliability of the model is questionable due to this flaw in data partitioning. Another project hosted on GitHub approached the problem of style classification using Convolutional Neural Networks (CNNs). While it achieved moderate success, reaching an accuracy of around 50%, it was limited to classifying only the top 5 artistic styles. This highlights the challenge of accurately distinguishing between artistic styles, which often exhibit overlapping characteristics.

However, one paper demonstrated that Vision Transformers (ViT), when trained from scratch on the WikiArt dataset, achieved over 39% accuracy

across 21 classes. [5] This result establishes a critical baseline for accuracy in art classification tasks, serving as a benchmark for future research and development in this domain.

Furthermore, another paper on Classification of Oil Paintings [11] presents a Bilinear Vision Transformer Neural Network (BViTNN) that combines the Vision Transformer for extracting fine-grained features with a Convolutional Neural Network (CNN) for capturing low-level semantic features. The accuracy achieved an impressive 86% on classifying the custom dataset. However, oil paintings represent only a small subset of classical art and were not the primary focus of this project.

Hence, this project aims to address these limitations by adopting a more robust data preprocessing and partitioning strategy, coupled with state-of-the-art Vision Transformers (ViT) and hybrid models like Xception+DenseNet.

3 Data

The dataset used in this project is sourced from Kaggle, titled *Best Artworks of All Time – Collection of Paintings of the 50 Most Influential Artists of All Time*. The dataset is publicly available and can be accessed via the following link: Best Artworks of All Time Dataset.

The dataset is relatively small, although comprising thousands of high-resolution images, with varying numbers of samples per artist. The imbalance existing also presents a challenge for accurate classification, for instance, the artist Vincent van Gogh had 877 paintings, while Jackson Pollock had only 45.

3.1 Data Preprocessing

To ensure reliable model performance, the following preprocessing steps were applied:

- **Data Splitting**: The dataset was carefully partitioned into training, validation, and test sets to avoid overlap between samples, ensuring unbiased evaluation.
- **Image Resizing**: All images were resized to 224 × 224 pixels to conform to the input requirements of the Vision Transformer (ViT) and other deep learning models.
- **Normalization**: Pixel values were normalized to the range [0, 1] to improve model convergence during training.
- **Top N Artists**: Selecting the Top N artists based on the number of images available and the data was trained and evaluated your based on this subset.

This project leverages these steps to preprocess dataset, in the view of building robust models for artist and style classification, and aiming to achieve higher accuracy and generalizability compared to existing and previous works.

4 Methods

In previous work, Convolutional Neural Networks (CNNs) were predominantly used for both artist classification and style recognition. Yet, the accuracies were low. Further to to improve the models, CNNs were targeted for artists classification, while transformers and attention mechanisms were anchored for style recognition. Hence, this paper explored and reconstructed two models for artist classification: VGG19 and a hybrid Xception+DenseNet. For style recognition, Inception and Vision Transformer (ViT) models were implemented. Below is the detail the methodologies and architectures employed.

4.1 Loss Computation

In all models, the loss function used is the categorical cross-entropy loss, defined as:

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

where L denotes the total cross-entropy loss, N is the number of samples, C represents the number of classes, $y_{i,c}$ is a binary indicator (0 or 1) indicating

if class label c is the correct classification for sample i, and $\hat{y}_{i,c}$ is the predicted probability of sample i belonging to class c. The objective is to minimize this loss and thereby maximize classification accuracy.

4.2 Classifying Artists

4.2.1 VGG19 Model

VGG19 employs a series of sequential 3x3 convolutional layers stacked deeply to learn complex visual features [7]. For this project, the VGG19 model was adapted to classify artists. The architecture of the modified model is shown in Table 1.

While VGG19's deep structure enables it to capture intricate details in images, it also leads to longer training times and higher computational demands, especially when dealing with high-resolution artwork.

Layer	Output Shape	Param #
vgg19	(N, N, N, 512)	20,024,384
global_avg_pooling2d	(N, 512)	0
max_pooling2d	(N, N, N, 64)	0
dense1	(N, 1024)	525,312
dense2	(N, 1024)	1,049,600
dense3	(N, 512)	524,800
dense4	(N, 49)	25,137

Table 1: Modified VGG19 architecture for artist classification.

4.2.2 Xception with DenseNet121

The **Xception** model enhances traditional CNN architectures by using depthwise separable convolutions, which separate spatial and channel-wise convolutions to reduce computational complexity while improving accuracy [10].

In contrast, **DenseNet121** introduces dense connectivity between layers, where each layer is connected to every other layer in a feed-forward manner [4]. This dense connectivity mitigates the vanishing gradient problem and allows for more efficient feature reuse.

Inspired by the effectiveness of multi-level feature utilization in CNNs [3, 6], we combined Xception and DenseNet architectures, as shown in Figures 1 and 2,

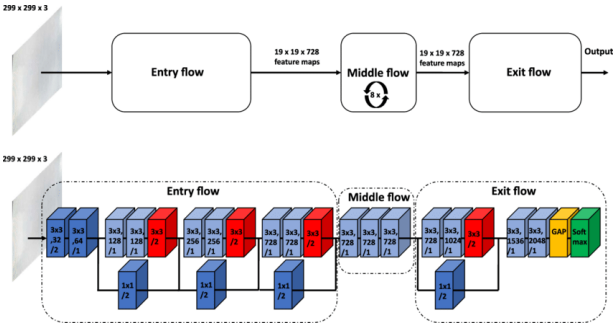

Figure 1: Xception model with the last 30 layers unfrozen, resulting in 3,322,758 trainable parameters.

with the last 30 layers unfrozen to allow fine-tuning on our dataset.

4.3 Classifying Styles

4.3.1 Inception Model

The Inception model was initially considered for style recognition due to its efficient architecture that balances depth and computational cost [8]. The model applies 1x1, 3x3, and 5x5 convolutions in parallel within its inception modules, capturing multi-scale features.

The architecture of our modified Inception model is shown in Table 2. Despite its efficiency, the model achieved only moderate accuracy, indicating the need for a more sophisticated approach.

4.3.2 Vision Transformers (ViT)

Inspired by the paper "Attention is all you need" [9], a major shift of the project was utilizing transformers to classify styles. Vision Transformers (ViT) represents a paradigm shift from traditional CNN-based methods to a transformer-based architecture designed for image classification tasks [1]. ViT divides an image into fixed-size patches, flattens them, and feeds them into a standard transformer encoder. This allows the model to capture long-range dependencies and complex patterns that CNNs might overlook.

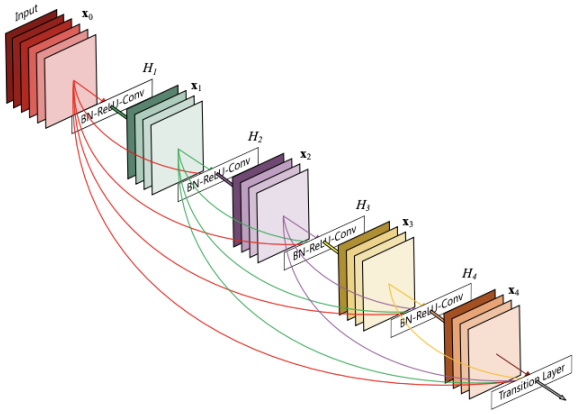

Figure 2: DenseNet121 with the last 30 layers unfrozen, resulting in 4,322,758 trainable parameters.

Layer	Output Shape	Param #
conv2d	(N, N, N, 64)	1,792
conv2d_1	(N, N, N, 64)	36,928
max_pooling2d	(N, N, N, 64)	0
conv2d_2	(N, N, N, 128)	73,856
conv2d_3	(N, N, N, 128)	147,584
max_pooling2d_1	(N, N, N, 128)	0
conv2d_4	(N, N, N, 256)	295,168
global_avg_pooling2d	(N, 512)	0
dense1	(N, 1024)	525,312
dense2	(N, 1024)	1,049,600
dense3	(N, 512)	524,800
dense4	(N, 49)	25,137

Table 2: Modified Inception model with a total of 1,468,959 parameters.

In this project, ViT outperformed CNN-based models in style classification, achieving significantly higher accuracy. The architecture can be seen at the table 3 and the model can be broken down into:

- **Patch Embedding**: The input image is divided into non-overlapping patches of size 16x16.
- **Transformer Encoder**: Each patch is passed through a multi-head self-attention mechanism, enabling the model to learn intricate relationships between different parts of the image.
- **Classification Head**: A fully connected layer outputs the final class probabilities.

By leveraging the global receptive field and attention-based feature extraction, ViT demonstrated superior performance in recognizing artistic styles, particularly when distinguishing between closely related styles.

Layer	Output Shape	Param #
vit (TFViTMainLayer)	multiple	85,798,656
classifier (Dense)	multiple	11,535
Total Param		85,810,191
Trainable Param		85,810,191
Non-trainable Param		0

Table 3: ViT Model Layer Structure with a total of 85,810,191 parameters.

5 Experiments

Traditional image classification methods, particularly those based on CNNs, tend to struggle with variability, as they rely heavily on feature extraction within fixed receptive fields, which is less effective when stylistic boundaries are ambiguous. Hence, the experiments have been divided into 2 parts, training on the entire dataset, and training on Top N Artists/Styles.

5.1 Training on entire dataset

5.1.1 Classifying Artists

Training across the entire dataset, the accuracy rates were exceptionally low. The VGG model achieved only 2% accuracy, which is below the baseline of random guessing, indicating minimal learning effectiveness.

The reason is that **VGG19** is extremely deep with a high parameter count, leading to substantial memory and computational demands. Its reliance on sequential convolutional layers limits its ability to capture complex, non-linear features across varying artistic styles. Although the Xception with DenseNet model achieved 12.5%, it is still not satisfactory, as

it tends to misclassify "Surrealism" and "Impressionism". This is actually expected, as quoted by a museum article "Impressionism dealt with realistic everyday scenes painted in a stylized way, while surrealism depicted unnatural scenes painted in a realistic style". [2] Hence, the nature of the styles is chosen upon realisticism, which poses a huge challenge for traditional CNNs to correctly distinguish between them.

5.1.2 Classifying Styles

Inception, designed with wider layers that use various filter sizes in parallel, performs better in capturing intricate features but still struggles with high intraclass variability present in classical art. In the end, having to low accuracy of 1.5%.

On the other hand, **ViTs** work exceptionally well with pretrained models. Achieving 80.5% accuracy even without selecting the top 10 prominent styles.

5.2 Training on Top K labels

As reducing the number of training classes is expected to increase model accuracy, data was augmented for both the art classifier model and the vision transformer model.

5.2.1 Top-k Artists Selection

For the selection of artists, we consider only those whose number of paintings exceeds a threshold that is defined as 1/5 of the artist with the most paintings. Let the total number of paintings of an artist i be denoted as p_i, and the artist with the maximum number of paintings is denoted as $p_{\max}$, where:

$$p_{\max} = \max(p_1, p_2, \ldots, p_N)$$

The threshold t is defined as:

$$t = \frac{1}{5} p_{\max}$$

Thus, the set of selected artists A_{selected} is:

$$A_{\text{selected}} = \{\text{artist}_i \mid p_i \geq t\}$$

5.2.2 Top-k Genre/Style Selection

For the selection of styles, we define the genre groups and choose the top k most frequent styles. Let s_i represent the style of a painting, and S_{group} represent the redefined genre group. The frequency of occurrence of a style s_j is denoted as $f(s_j)$. The top k selected styles are then:

$$S_{\text{selected}} = \text{top-}k\left(\{f(s_1), f(s_2), \ldots, f(s_m)\}\right)$$

Where m is the total number of distinct styles in the dataset.

5.3 Classification of Artist and Style

The final classification task predicts both the artist and style for a given image I:

$$\hat{A}(I) \in A_{\text{selected}} \quad \text{and} \quad \hat{S}(I) \in S_{\text{selected}}$$

Where $\hat{A}(I)$ is the predicted artist and $\hat{S}(I)$ is the predicted style.

Hence, the artist classifier focused on the top 16 most prolific artists, avoiding the artists that have too small dataset. On the other hand, the vision transformer focus on the top 8 styles, reducing styles that overlap and causes noise in the model.

5.4 Results

All models were evaluated for both artist and style classification tasks. The Vision Transformer (ViT) achieved the highest accuracy of 81.50% for style classification after regrouping genres, while Xception + DenseNet reached 32.35% for artist classification when applied to the top 16 artists. In comparison, the VGG19 and Inception models showed lower accuracy, with VGG19 achieving only 2% for artist classification and Inception reaching 1.5% for style classification. These results highlight the effectiveness of advanced models, particularly ViT, in handling art classification tasks. The results can be seen in Table 4.

Model	Target Classification	Pretrained Weights	Accuracy (%)
VGG19	Artist	No	2.00
Xception + DenseNet	Artist	Yes	12.50
Xception + DenseNet (top 16 artists)	Artist	Yes	32.35
Inception	Styles	No	1.50
Inception (genre regrouping)	Styles	No	9.85
ViT	Styles	Yes	80.50
ViT (top 8 styles)	Styles	Yes	81.50

Table 4: Model Accuracy Results for Classic Art Classifier

Figure 3: ViT Model Results for Style Classification

6 Conclusion

In this project, various deep learning architectures for classifying artists and artistic styles were explored. A range of models, including VGG19, Xception + DenseNet, Inception, and Vision Transformers (ViT), were used to classify artworks, who the artist was, and what the style was. By leveraging pretrained weights and fine-tuning the models, we saw improvement in classification accuracy despite challenges posed by data imbalances.

Our experiments demonstrated the significant potential of ViT models, which achieved the highest accuracy for style classification (81.50%) before or after grouping related genres. While traditional CNN architectures like VGG19 and Inception were effective to some extent, their performance was limited, no matter in artist or style classification. Xception + DenseNet, with its improved design, showed better performance for artist classification, particularly when we focused on the top 16 artists.

The results confirmed that leveraging models with pretrained weights, especially ViT, offered considerable improvements. Additionally, the genre regrouping technique led to better handling of similar styles, contributing to enhanced accuracy.

7 Future Work

The findings suggest that incorporating advanced architectures like Vision Transformers and refining datasets through techniques such as expanding the paintings for an artist can significantly enhance model performance for art classification tasks. Future work could involve expanding style recognition by integrating additional datasets to improve accuracy and generalizability across diverse art styles.

Additionally, an interactive application could be developed, allowing users to upload a photo of a painting and receive predictions on both the artist and the style. This would not only demonstrate the practical utility of the model but also provide a valuable tool for art enthusiasts and researchers alike. Further efforts could focus on fine-tuning models, scaling them to larger datasets of over 20,000 images,

and leveraging attention mechanisms to enhance both style and artist classification.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[2] Eyewire. Eyewire museum: Impressionism vs surrealism, Apr. 2019. 5

[3] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization, 2015. 3

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 3

[5] Lazaros Alexios Iliadis, Spyridon Nikolaidis, Panagiotis Sarigiannidis, Shaohua Wan, and Sotirios K. Goudos. Artwork style recognition using vision transformers and mlp mixer. *Technologies*, 10(1), 2022. 2

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015. 3

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 3

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. 3

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[10] Erik Westphal. A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks. *Additive Manufacturing*, 41:101965, 03 2021. 3

[11] Tianrui Wu, Jiejie Chen, Haiming Zhao, Zhuzhu Zhang, Mingyuan Qin, and Xiaohan Huang. Classification of oil paintings based on improved vision transformer. In *2023 International Conference on Neuromorphic Computing (ICNC)*, pages 263–268, 2023. 2