Automatic Classification of Portraits: Application of Transformer and CNN Based Models for an Art Historic Dataset

Sebastian Diem[1], Thomas Mandl[1]

[1]University of Hildesheim, Germany

Abstract

This research compares the performance of a Visual Transformer to a ResNet50 on a small art historical dataset. The ResNet is a widely used model based on Convolutional Neural Networks (CNNs) and has achieved good performance in a variety of computer vision experiments. Our experiments show how the relatively novel Visual Transformer performs compared to ResNet50 for a dataset from the Digital Humanities. We are using a large collection of portraits from the 15th to the 19th century and select the 10 most frequent artists for a classification task. Portraits reveal social values and artistic styles over the centuries. Like many other collections in the Humanities, they lack annotations and require automatic methods for generating metadata. We observe that the Visual Transformer achieves a top-1 accuracy of 87.09 % in contrast to the ResNet's 46.13 % accuracy. Analysing features like the printing technique and active period of the artist in question shows, that these features could be important to explain the model's inference process. Other features like the portrait type seem to have less impact. To further analyze the performance of the models, we applied Centered Kernel Alignment method, Gradient-weighted Class Activation Maps (GradCAMs) and Attention Map visualizations. On the one hand, the importance of the printing technique can be further emphasized when visualizing the models' hidden layers, where both models seem to attend to the portrait backgrounds, as these parts could be the easiest to distinguish the distinctive printing patterns. On the other hand, tends the Visual Transformer to focus on the portrayed person as they seem to be important for the artist classification.

Keywords

Digital Humanities, Portraits, Image Processing, Deep Learning, CNN, Computer Vision

1. Introduction

Since the Iconic Turn, research involving pictures and visual media has been established in the Digital Humanities. Images are very important in the spread of knowledge. For example, the invention of lithography in the 19th century resulted in declining manufacturing costs for printed images, giving an increasing number of people access to a wealth of visual information [1]. In the Digital Humanities the potential of image processing capabilities also gained significance. As libraries and museums further digitalize their collections, more researches gain accessibility to art historical data to conduct experiments [2]. Diachronic developments within image collections are particularly fascinating, because research might show trends in stylistic and

LWDA'23: Lernen, Wissen, Daten, Analysen. October 09–11, 2023, Marburg, Germany

diem@uni-hildesheim.de (S. Diem); mandl.uni-hildesheim.de (T. Mandl)
0009-0000-9196-4814 (S. Diem); 0000-0002-8398-969 (T. Mandl)

aesthetic representation.

The creation of suitable tools and techniques for distant viewing, or the automatic analysis of massive volumes of visual data using computer vision technologies, is crucial for the Digital Humanities. Often, the tasks and collections within Digital Humanities are not well suited for generating annotations. However, future search and analysis systems need to provide many more options than current tools. As a consequence, the automatic generation of metadata for large datasets is one solution to improve the research opportunites within image collections. This is also necessary to overcome critical positions within the Humanities toward digital methods (e.g. [3]). Such automatically generated metadata can be used in retrieval tools for more flexibel access. However, there is still much doubt about the quality of such data.

In this research, the performance of a Transformer based model is compared to the results of an established CNN model for an art historical dataset. Previously, Convolutional Neural Networks have been the state of the art for a variety of Computer Vision tasks. The introduction of the Transformer architecture in the field of Computer Vision might represent an alternative to CNNs, as models like the Visual Transformer achieve comparable results to modern CNNs[1]. In order to observe how the new Transformer based models perform on an art historical dataset this work compares the Visual Transformer with the often used ResNet50 which is based on CNNs. Additionally, this research uses a small custom art historic dataset to compare top artists who created printed portraits.

We utilize different methods to explain which features are important for the individual models in their classification process. It can be observed that the used printing technique influences the prediction quality as different printing techniques have significantly different detection rates. This can be proven when observing the visualization of different hidden layers, where lower layers attend to these local features. Furthermore, the epoch when the artists were active seems to be an important factor, as artists active at the same point in time are harder to distinguish than others. In the related works we introduce the medium of printed portraits before looking at related art historical applications of computer vision models. Lastly, we show the current situation in computer vision with the introduction of the Transformer architecture. The experiment setup explains how the dataset was created, how the models were trained and how we evaluate the results, before presenting them and discussing the findings.

2. Related Work

2.1. Printed Portraits

The medium of printed portraits only started to get acknowledged for scientific studies in the last few decades as it was previously regarded as copies of popular artworks. For Europe in the early modern period printed portraits were a widespread medium (1450 until the end of the 18th century). Over time the profession of printers was established [4]. Printers were mainly regarded as illustrators and produced commissional work without being necessarily painters themselves. This led to the increased production of printed portraits as more social groups could afford portraits of themselves (e.g. aristocrats, scholars, craftsmen and wealthy citizens).

[1] https://paperswithcode.com/sota/image-classification-on-imagenet

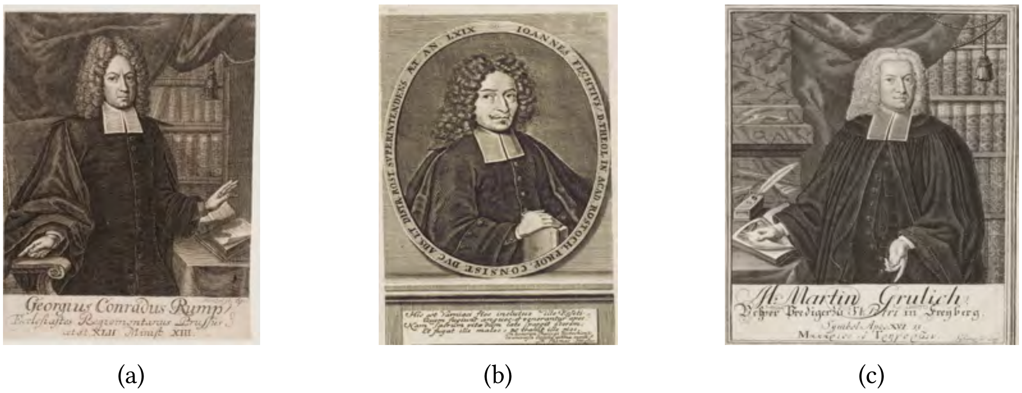


Figure 1: Three examples from the top artists depicting similarly dressed scholars
(a) from Johann Georg Mentzel
(b) from Johann Georg Mentzel
(c) from Johann Christoph Sysang
Artworks part of the printed portrait collection of the Herzog August Library
Source: http://portraits.hab.de/

This popularized the trend of collecting and trading portraits for individuals and even led to portraits being cut out of books to expand private collections, thus removing them from their original historical context [4].

With this acknowledgement, more research has been conducted to analyse printed portraits. One of the most popular forms of analysing portraits is iconography. Iconography analyses the content and style of an image and interprets it to gain historical and art historical insights. One interesting observation is, how popular motives change over the centuries as the zeitgeist changes conventions of depiction [5]. This can be best seen in the reinterpretation of popular motives where the clothing or gestures stay mostly the same and other elements of the artwork are changed to suit the likings of the current epoch [6]. In this form of analysis, the visual representation of a motif could only be described. This extraction of information limits the comparability of visual elements as they might possess more nuanced information that get lost in a description.

For visual interpretation an attempt is made to contextualize a portrait based on reoccurring features or other conventions of representation typical for a period. Different epochs usually possess comparable representations like the occurrence of certain objects or clothes [4, 6]. These and other typical elements can give insights into the social status of the depicted person or used stylistic conventions of the time. In Figure 1 this can be seen for different depictions of scholars. All of them have similar clothes and objects which show their status in society. Often, further insights can be gained about the origins of a portrait based on the reoccurrence of objects and other elements. Printer often reused elements like the portrait frame or common objects like books to reduce production time and cost. Additionally, different printing techniques have been used over the years from wood and copper engravings to advanced techniques like lithography, which are differentiable by their individual properties (e.g. wood grain).

2.2. Computer Vision

Computer Vision has advanced significantly in recent years, particularly due to progress with deep learning methods and representation learning. These data-driven methods have been successful for a variety of tasks and frequently perform better than conventional image processing techniques focusing, for instance, on color and form analysis. Contemporary Deep Learning techniques identify pertinent features from images and learn their own representation schemes [7]. Multiple-layer Convolutional Neural Networks (CNNs) in particular have demonstrated to be quite successful [8].

The ResNet architecture, especially ResNet50 is currently one of the most relevant models in Computer Vision. It was developed in 2015 and has been used as a baseline for a variety of research papers since then [9]. It introduced the skip connection and used the ReLU activation function for its hidden layers to achieve state of the art results with 76.1 % top-1 accuracy on the ImageNet dataset. In 2021 a research team revisited the original ResNet50 architecture and used novel optimization and data augmentation techniques without changing the architecture to achieve a top-1 accuracy of 80.4% on the ImageNet dataset, which emphasizes its relevance even today [10].

The Visual Transformer (ViT) was presented in late 2020 and introduced the successful Transformer architecture to Computer Vision. With its introduction in the field of Natural Language Processing the Transformer architecture with known models like BERT or GPT1 became the new standard for a variety of tasks. As the Transformer was build to handle sequences of text the architecture of the ViT differs a lot from a Convolutional Neural Network (CNN). The ViT splits an image into 16x16 pixel patches and realigns them into a sequence of patches. Spatial information are retained by the position embedding of each patch. It uses the self-attention function to focus parts of the image and has multiple attention heads per layer. The Vit-H (Huge) variant achieved 88.55 % top-1 accuracy on the ImageNet Dataset [11].

Methods for a deeper understanding of how Computer Vision models work are being developed with different approaches. Visualization techniques like the Class Activation Maps allow to look into deep learning model's inner representations and utilize the activation function to see which part of an image is attended to the most [12]. Other visual approaches use cluster algorithms to differentiate classes into potential clusters. While processing an image, a CNN creates an embedding vector to capture the information about the image before using this embedding for their prediction. With dimensionality reduction algorithms like t-SNE or UMAP these embedding vectors can be reduced to a visualizable number of dimensions. These cluster visualizations help to distinguish which classes are easier or harder to differentiate [13, 14].

In 2021, research compared how ViT and ResNet utilize image information for classification tasks. They used Centered Kernel Alignment to save the hidden states of all layers and calculate a similarity score between all possible combinations of layers [15]. They observed, that ViT have a more uniform representation across all model layers whereas the similarity between lower and higher model layers of a ResNet is weaker [16]. Comparing every layer of the ResNet with every ViT layer shows, that the first 30 ViT layer have the most similar representations with the first 60 ResNet layers. The higher the representations of the models are, the lower is the similarity between them. This implies that local information aggregation, which is mostly captured in early layers, is import for both architectures and later more abstract representations are used for the final classification.

2.3. Applications of Digital Humanities for Art History

In recent years, multiple tools emerged for the automatic analysis of art [17, 18]. They often utilized Computer Vision models and focused on very different aspects of art [19]. The predominant dataset for these studies is the WikiArt dataset. The WikiArt dataset consists of around 250,000 artworks by over 3000 artists and provides metadata regarding style, date, artist, genre and more for the individual pieces. Other commonly available datasets that have been used are the Web Gallery of Art (WGA)[2] and the TICC Printmaking dataset[3]. The WGA dataset consists of 52,867 pieces and like the WikiArt dataset includes artworks from many epochs and a variety of different media. The TICC dataset with 58,630 images and 210 artists is a more specific dataset in comparison. It focuses on printed artworks from the Netherlands State Museum (Rijksmuseum) and excludes other media.

The observed studies mainly focus on classification tasks. They include the differentiation of aspects like art style, genre, artist and the painting style [20, 21, 22, 23, 24]. They use Support Vector Machines and different iterations of CNNs like the CaffeNet, ResNet18, ResNet50 or the All Convolutional Net. The results for the WikiArt dataset range from 33.62 % to 79.1 % for artist classification top-1 accuracy [22, 20]. For the WGA dataset, artist classification reached a score of 69.6 % (top-1 accuracy) and on the TICC dataset 76.2 % top-1 accuracy and 82.12 % mean class accuracy [20, 24]. In another study, a maximum accuracy of 80 % was obtained [25]. Many other experiments were published, however, there they were applied to diverse datasets.

In regards of printed media from the early modern period only a few examples have been found. The TICC dataset includes printed portraits in their artist classification but do not specify findings or challenges regarding this medium like the influence of the used printing technique. Different printing techniques, like woodcuts, wood engravings and copper engravings are shown to have varying detection rates for CNN based models [26]. These properties also seem to influence the quality of applications on printed media datasets like the detection of objects in early modern children and youth books. Beyond classification experiments, similarity has often been considered as an important concept in art history [27, 28].

Utilizing visualization techniques like the Class Activation Maps reveal, that algorithms might not consider the content parts of an image and rather focus on other parts with more distinctive patterns like the frame of an image [29].

3. Experiment Setup

For this research a comparison between two deep learning models is conducted. First, the dataset for this experiment is introduced. Afterwards, the used models and the training process are briefly described. The last part describes the evaluation process to examine the classification results.

[2] https://www.wga.hu/index_database.html

[3] https://auburn.uvt.nl/

3.1. Dataset

The dataset used in this experiment is part of an art historical collection of printed portraits from the Herzog August Library in Wolfenbüttel[4]. The collection consists of nearly 32,000 portraits of which roughly 28,000 have been digitalized over the last decade heterogeneously. Based on the metadata of the collection the ten most occurring printing artists have been selected for this classification experiment. In total 2834 images can be associated with these ten artists. As the distribution between the artists is highly diverse, with 631 portraits for the most prevalent artist and 156 examples for the least prevalent, the training dataset was limited to 140 randomly selected images per artist. Other studies used between 96 to 500 artworks per artist, which indicates that the number of examples is sufficient [20, 21, 22]. The training dataset was split into 80 % training and 20 % test data per artist.

3.2. Models and Training Process

For this research the performance of a ResNet50 and a Visual Transformer will be compared based on the classification dataset. The ResNet is, as previously described, one of the most renown architectures and has been featured in a variety of different comparison studies. The other model is a large Visual Transformer model with 16x16 pixel patch size (ViT-L/16). Both models are trained with different hyperparameter configurations. The images have been resized in a preprocess step to fit the models' expected input size. The best models based on validation top-1 accuracy and validation loss are selected. Previous art historic works described that the ResNet is prone to overfit easily [20, 23]. To counteract this tendency early stopping is implemented. As the ViT-L is a much bigger model than the ResNet50 with 307 million parameters to compared 26 million parameters its possible overfitting tendencies are also monitored. For the ResNet 12 different hyperparameter configurations are tested. For the ViT the ViT-B/16 variant is also tested to observe possible differences in the training process. All models have been trained from scratch and in a resolution of 224x224 pixels or 384x384 pixels.

3.3. Evaluation Process

To evaluate the performance of the models beyond the top-1 accuracy additional data sources and analysis tools are used. The metadata of the printed portrait collection also includes further information regarding the context and content of the portrait. These information include the printing technique that was used in creation of the portrait and also which section of the person was depicted (e.g. half-length portrait or chest up portrait). Additionally, the metadata includes the year of origin for 1656 portraits. One of the artists has been excluded in the productive periods comparison as only three portraits possess date information (Fennitzer, Georg). Afterwards the median per artist was calculated and used as a reference value for wrongly classified images. This way it can be observed if a wrongly assigned portrait was created in the same period as the artists' productive period. Before analysing the individual features a chi squared significance test has been made. The results here showed that there is a statistical dependency in the data, as the null-hypothesis was rejected.

[4] Herzog August Bibliothek Wolfenbüttel, Germany, (www.hab.de)

For insights into the model's inner processes the previously mentioned Centered Kernel Alignment method, Gradient-weighted Class Activation Maps (GradCAMs) and Attention map visualizations are used. As only the core principle of CKA is implemented in the demonstration of previous works, the implementation in this work might differ. Other than in the related works proposed comparison of ResNet50 and ViT-L only the outputs of the individual blocks have been used as the calculation per layer is computationally demanding. This results in 50 outputs from the ResNet50 and 24 outputs from the ViT-L. To get standardized results the CKA is calculated on 25 examples per artist with 250 examples in total. The example representations are averaged layer wise before calculating the CKA. For the GradCAMs the last hidden layer of every ResNet block is visualized and manually compared to results from the Attention maps. The attention maps of the ViT are created by taking the sequence length x number of patches and averaging them layerwise over all attention heads [5].

4. Results

This chapter summarizes the main findings of the research. The final results for the test dataset with 1274 datapoints reached a top-1 accuracy of 46.13 % for the ResNet and 87.09 % for the Visual Transformer. Compared to all 2834 examples the ResNet50 achieved a top-1 accuracy of 51.9 % with 1471 hits and the ViT-L 92.27 % with 2615 hits. The distribution of error is comparable for both results. To have more datapoints for the following comparisons the full dataset of 2834 images will be used. The accuracy per artist is summarized in Table 1. For the ResNet the best prediction is for the artist Tobias Stimmer with 92.95 % and the lowest accuracy for Johann Martin Bernigeroth with only 17.39 %. The ViT achieved the highest accuracy also for Tobias Stimmer with 99.36 % and the lowest for Martin Bernigeroth with 84.79 %.

Evaluating the detection rate in comparison to the printing technique shows that ResNet had the highest accuracy for the wood engraving technique with 93.55 % when excluding mezzotint/ etching which only occurs two times in the whole dataset. The ViT also has the highest detection rate with 100% for wood engraving. Both models also have the lowest accuracy for the combination of etching / copper engraving with 37.34 % for the ResNet and 89.21 % for the ViT (see Table 2).

The mean difference between the artists productive time and wrongly assigned portraits is at 34.6 years for the ResNet and 20.4 years for the ViT. Lastly, for 75 % of the false predictions are within a time difference of 51.0 years for the ResNet model and 31.3 years for the ViT model. Transferring the detection rate to the portait type the highest detection rate for the ResNet was 73.17 % and 97.56 % for the ViT. Both of these results are for the headpiece portrait type. The lowest detection rate for the ResNet was 36.36 % and 81.81 % for the ViT. Both results are for the chest up portaits as seen in Table 3.

5. Discussion

The results of the presented experiments show the performance of two very different deep learning architectures on a rather small dataset. They also show an extreme difference in the

[5] Code from: https://github.com/jeonsworld/ViT-pytorch/blob/main/visualize_attention_map.ipynb

Table 1
Top-1 accuracy for the ten artists

Artist	Top-1 accuracy in %		Quantity
	ResNet	ViT	
Johann Martin Bernigeroth	17,39	87,96	299
Martin Bernigeroth	31,70	84,79	631
Georg Fennitzer	84,49	97,61	419
Johann Franck	89,38	97,50	160
Wolfgang Philipp Kilian	38,77	91,63	227
Johann Friedrich Leonart	53,68	95,24	231
Johann Georg Mentzel	54,65	94,19	172
Matthias van Somer	83,47	96,61	236
Tobias Stimmer	92,95	99,36	156
Johann Christoph Sysang	24,42	92,08	303
Total	51,91	92,27	2834

Table 2
Detection rate per printing technique combinations

Printing technique	Detection rate in %		Quantity
	ResNet	ViT	
Copper Engraving	42,78	90,22	1503
Etching	55,33	97,54	244
Etching/ Copper Engraving	37,34	89,21	482
Mezzotint	81,66	96,89	447
Mezzotint/ Etching	100,00	100,00	2
Wood engraving	93,55	100,00	155

Table 3
Detection rate per portrait type

Portrait type	Detection rate in %		Quantity
	ResNet	ViT	
Bust	36,36	81,82	11
Full length	40,74	74,07	27
Half length	40,57	89,69	456
Head piece	73,17	97,56	41
Kit-cat	44,86	91,83	1090
one-quarter length	63,57	94,25	1131
three-quarter length	41,54	87,69	65

accuracy of the ResNet50 model and the ViT model.

5.1. Classification of Artists

Previous works achieved different results utilizing ResNet variants from 49.4 % top-1 accuracy for style detection to 80.0 % for artist classification which indicates that there might be room for improvement [23, 25]. As previously mentioned the ResNet's training process was prone to overfit. A bigger dataset with more classes or more examples per class might reduce the risk of overfitting. Contrary to that are the results of the ViT-L with 92.27 %. The developers of the Vision Transformer claim, that the ViT models are more prone to overfitting and perform worse on small datasets in comparison to ResNets [11]. The final tests are conducted over the whole dataset of 2834 examples of which only 1120 were used for the actual training. This indicates that the ViT did not overfit in our experiments and extracted valid internal representations of the dataset.

Comparing the experiments' results with further information from the metadata additional trends can be observed. The first observation indicates different detection rates between artists. For the ResNet this trend is very strong with the best detection rate for Tobias Stimmer with 92.95 % accuracy and only 17.39 % for Martin Bernigeroth. This clearly shows that different artists possess differentiable features. This trend can also be observed for the ViT model, although the gap is smaller between 84.78 % for Martin Bernigeroth and 99.5 % for Tobias Stimmer. This supports the thesis that the portraits of Tobias Stimmer are easier to differentiate.

5.2. Analysis of Metadata

In Figure 2, the period of the productive period can be seen for the artists with available date information. Here it can be seen that Tobias Stimmer is the only of the nine artists activate in the late 16th century, whereas Martin Beringeroth is active in the first half of the 18th century together with four other artists. Overall, the performance of the ResNet is the worst in this period with the accuracy between 17.39 % and 54.65 %. The ViT again shares this trend with smaller extremes between 84.78 % and 94.19 %. For both models Johann Georg Mentzel is the easiest artist to identify for the models in this period. Another good indicator for the differentiability between the artists seems to be the printing technique. The ResNet assigned 93.5 % of the 155 Wood Engravings to the correct artist and the ViT achieved 100% accuracy for this printing technique. As it is one of the older printing techniques the only artist that used it is Tobias Stimmer. Another example is the Mezzotint technique which was nearly exclusively used by Georg Fennitzer. For this technique, ResNet assigned 81.65 % of 447 Mezzotint portraits correctly. The ViT classified 96.87 % correctly. This could indicate that exclusively used printing techniques are a good indicator for high classification accuracy. Contrary to this, the ViT achieved its second highest prediction accuracy with 97.54 % for the etching technique which was used by three different artists to a certain extend. This shows, that good accuracy can also be achieved without relying solely on the printing technique. This can be further emphasized for Copper Engravings which is the most used printing technique with 1503 portraits and frequently used by 8 of the 10 artists. Both models were able to achieve high accuracy for Matthias van Somer with 83.47 % for the ResNet and 96.61 % for the ViT, who used Copper Engravings in 212 of his 236 portraits in this dataset. This could indicate that he used a different style or other common elements in his portraits, which made it easier to identify his works.

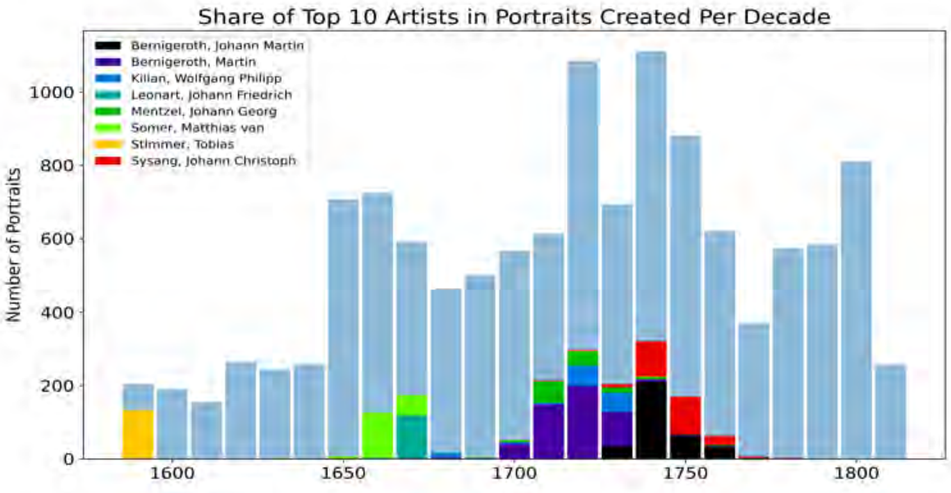


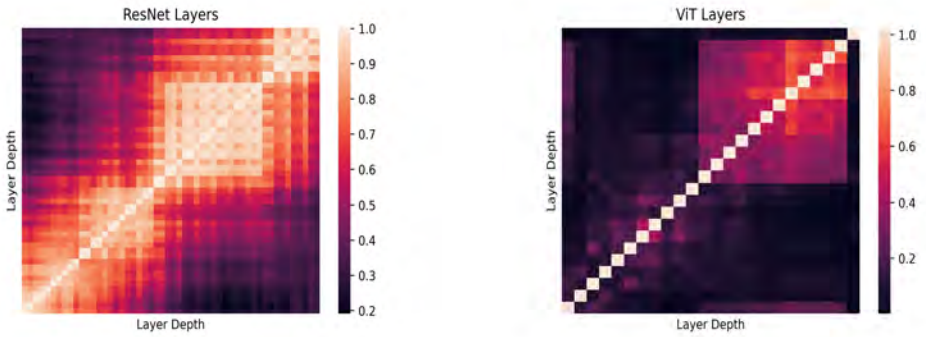
Figure 2: This graph shows the share of the top 10 artists over all portraits from the Herzog August Library's collection of printed portraits per decade. Fennitzer, Georg has been removed as previously mentioned as no dates were available

All in all, as both the ViT and the ResNet acquire a large amount of low-level information in the first few layers, the printing technique might be a good indicator for classification accuracy when comparing portraits from multiple centuries [15, 16].

The last used metadata feature, the portrait type, seems to have less influence on the accuracy although fluctuations can be observed here too. This could mainly be due to the fact that 2221 portraits are part of the two main portrait styles (chest up and half figure portraits). Different to previous comparisons the ResNet's detection rate had less divergent extremes with a difference of only 37 % (see Table 3). For the ViT, the difference accumulated to 23 % between the best and the worst detection rate. This indicates that this is a harder feature to focus on for the model. This could very well be due to the unbalanced distribution of portait styles in the used dataset.

5.3. Model Analysis

Figure 3 shows the Centered Kernel Alignment representations for the ResNet on the left and ViT on the right. As previously mentioned the CKA method determines how similar the representations between the individual layers are. This is measured by the similarity score ranging from 0 to 1, where 1 is the highest similarity. As both axis of the graphs display the layers of the model the diagonal line always has a similarity of 1 as the layer is compared to itself. For the ResNet, it is clear to see, that layers in closer proximity share more similar representations than layers further apart due to the nature of its architecture. This is in line with the findings of previous work [15]. The grid pattern, that can be observed if Figure 3a arises from the architecture of the model [15]. In previous studies this specific pattern in a Transformer model



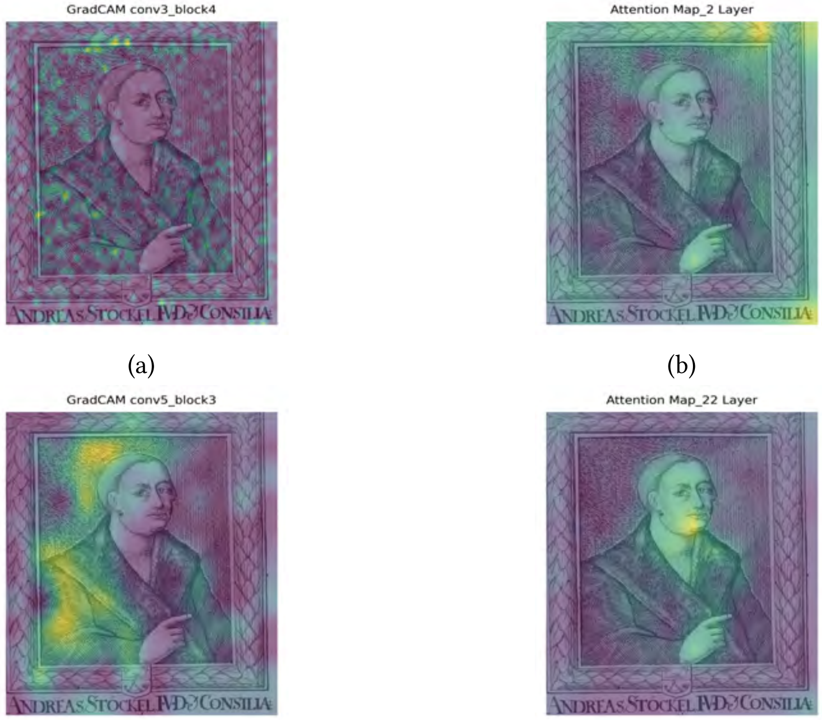
(a)

(b)

Figure 3: Centered Kernel Alignment visualization
(a) Similarity between the ResNet layers for all 50 layers
(b) Similarity between the ViT Layers for all 24 layers

was attributed to the skip connection [16]. ResNet's architecture also includes a skip (shortcut) connection [9]. This could imply that this connection is also visible in the the ResNet's CKA. Contrary to previous studies Figure 3 (b) shows the CKA of the ViT model has the highest similarity in the last third of the layers and a few corresponding layers in the middle part of the model. A possible explanation was given by Raghu et al.. If the ViT is not supplied with enough data it cannot learn local representations in early layers. This could result in lower similarity as displayed in Figure 3. Visual Transformer have transitional phases where the representation between the layers shifts from lower layers to higher layers. Lower layers attend to local and also global information whereas higher layers attend only to global features [16]. This could be supported by visualizing the GradCAMs and Attentions.

In Figure 4 the visualizations show which parts of the image are most important for the classification of the image. The overlay color determines how important a region is. Image regions that are yellow have a big impact on the classification, green a slight influence and blue regions (the base color) are irrelevant. In the early layers it is possible to see, that the ResNet focusses on the lower level details as it highlights part of the clothes, background and portrait frame, possibly to determine the structures of the printing technique, as previous art historical works observed similar behaviour (Figure 4 (a)) [29]. This can also be observed for the ViT, although in lower detail, as it seems to observe all parts of the portrait around the portrait person first (Figure 4 (b)). With higher layers the ResNet keeps its focus and propagates the detected local features to global features for a final classification (Figure 4 (c)). For the ViT the attention shifts completely and nearly exclusively attends to the portrayed person for its classification (Figure 4 (d)). This supports that there is no similarity between the lower and higher layers in the CKA analysis (the dark areas of Figure 3 (b)). These differences can be observed for multiple examples. As the ViT focusses on the portrayed person means, that this part of the artwork must possess important information. This could be due to a multitude of reasons. One explanation could be, that the artists must have distinctive styles. It might also be



(a) (b) (c) (d)

Figure 4: GradCAM and Attention visualization
(a) Middle layer of ResNet
(b) Second layer of ViT
(c) Last layer of ResNet
(b) Third to last layer of ViT
Artworks part of the printed portrait collection of the Herzog August Library
Source: http://portraits.hab.de/

due to the social classes of the customers being portrayed by an artist. Lastly, also the depiction conventions of the different time periods could be of significance.

6. Conclusion and Future Work

This research shows the applicability of modern deep learning models for an art historical dataset. On the one hand it demonstrates how newer state of the art Transformer based models perform in comparison to established CNN based models. This is especially significant in regards of the high top-1 accuracy which the Visual Transformer achieved with over 87 % and a very small dataset size. This performance shows that this architecture might be useful for niche or other art historical classification problems to potentially outperform older models and thus support the work of art historians more reliably. It might also be interesting how the ViT would perform with a bigger dataset and more artists. For this the dataset could be expanded to include more artists or an eleventh “other” class, which could be useful for metadata generation. It is to be said, that the performance of a different CNN model like the EfficientNet could achieve superior results [30]. This needs to be analyzed in future work.

On the other hand, this research utilized a variety of different tools to analyse both the prediction results and model representations. For this artist classification it can be seen that other features of the portraits and even historical aspects are significant indicators. Features like the printing technique have a noticeable impact for the prediction quality, especially for the ResNet model. The year of a portraits creation also seems to be important, as depiction trends could influence how persons are portrayed. The usage of visualization techniques showed that both models seem to focus on the background on lower layers as they might be attending to the small distinctive features of the printing technique. The ViT often focused on the displayed person to determine the portrait’s artist. This indicates, that the portrayed person might present valuable information for the classification process. Further models for analysing deep networks could be used in future work [31].

References

[1] N. van Noord, A survey of computational methods for iconic image analysis, Digital Scholarship in the Humanities 37 (2022) 1316–1338. doi:10.1093/llc/fqac003.

[2] G. Mercuriali, Digital art history and the computational imagination, International Journal for Digital Art History (2018) 141. doi:10.11588/dah.2018.3.47287.

[3] A. Bentkowska-Kafel, Debating digital art history, International Journal for Digital Art History (2015). doi:10.11588/dah.2015.1.21634.

[4] P. Poch, Porträtgalerien auf Papier. Sammeln und Ordnen von druckgrafischen Porträts am Beispiel Kaiser Franz‘ I. von Österreich und anderer fürstlicher Sammler., Böhlau Verlag, 2018. doi:10.7767/9783205208556.

[5] S. Skowronek, Autorenbilder: Wort und Bild in den Porträtkupferstichen von Dichtern und Schriftstellern des Barock, Würzburger Beiträge zur deutschen Philologie, Königshausen & Neumann, 2000. URL: https://books.google.de/books?id=_Jpx8FkObicC.

[6] N. Niedermeier, Visuelle Ähnlichkeit als relationaler Formbegriff: Automatische Bilderkennung von Reproduktionen frühneuzeitlicher Porträtgrafik, 2022. URL: https://kunstgeschichte-kongress.de/programm/programm-2022/, Deutscher Kunsthistorikertag.

[7] T. Mandl, S. Diem, C 5 Bild- und Video-Retrieval, in: R. Kuhlen, D. Lewandowski, W. Semar, C. Womser-Hacker (Eds.), Grundlagen der Informationswissenschaft, De Gruyter Saur, Berlin, Boston, 2023, pp. 413–422. doi:doi:10.1515/9783110769043-035.

[8] C. C. Aggarwal, Neural Networks and Deep Learning A Textbook, Springer, Cham, 2018. doi:10.1007/978-3-319-94463-0.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385.

[10] R. Wightman, H. Touvron, H. Jégou, Resnet strikes back: An improved training procedure in timm, CoRR abs/2110.00476 (2021). URL: https://arxiv.org/abs/2110.00476. arXiv:2110.00476.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, CoRR abs/2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929.

[12] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization, CoRR abs/1610.02391 (2016). URL: http://arxiv.org/abs/1610.02391.

[13] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[14] L. McInnes, J. Healy, UMAP: uniform manifold approximation and projection for dimension reduction, CoRR abs/1802.03426 (2018). URL: http://arxiv.org/abs/1802.03426. arXiv:1802.03426.

[15] S. Kornblith, M. Norouzi, H. Lee, G. E. Hinton, Similarity of neural network representations revisited, CoRR abs/1905.00414 (2019). URL: http://arxiv.org/abs/1905.00414. arXiv:1905.00414.

[16] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, CoRR abs/2108.08810 (2021). URL: https://arxiv.org/abs/2108.08810. arXiv:2108.08810.

[17] C. Hastik, P. Hegel, Bilddaten in den Digitalen Geisteswissenschaften, Berlin, 2020. doi:10.17169/refubium-30108.

[18] T. Arnold, L. Tilton, Distant viewing: analyzing large visual corpora, Digital Scholarship in the Humanities 34 (2019) i3–i16. doi:10.1093/llc/fqz013.

[19] M. Wevers, T. Smits, The visual digital turn: Using neural networks to study historical images, Digital Scholarship in the Humanities 35 (2020) 194–207. doi:10.1093/llc/fqy085.

[20] E. Cetinic, T. Lipic, S. Grgic, Fine-tuning convolutional neural networks for fine art classification, Expert Syst. Appl. 114 (2018) 107–118. doi:10.1016/j.eswa.2018.07.026.

[21] N. Viswanathan, Artist Identification with Convolutional Neural Networks, 2017. URL: http://cs231n.stanford.edu/reports/2017/pdfs/406.pdf.

[22] B. Saleh, A. M. Elgammal, Large-scale classification of fine-art paintings: Learning the right metric on the right feature, CoRR abs/1505.00855 (2015). URL: http://arxiv.org/abs/1505.00855. arXiv:1505.00855.

[23] A. Lecoutre, B. Négrevergne, F. Yger, Recognizing art style automatically in painting with deep learning, in: Proceedings Asian Conference on Machine Learning, ACML 2017, Seoul, Korea, Nov. 15-17, PMLR, 2017, pp. 327–342. URL: http://proceedings.mlr.press/v77/lecoutre17a.html.

[24] N. van Noord, E. O. Postma, Learning scale-variant and scale-invariant features for deep image classification, CoRR abs/1602.01255 (2016). URL: http://arxiv.org/abs/1602.01255. arXiv:1602.01255.

[25] C. Cömert, A. M. Ozbayoglu, C. Kasnakoğlu, Painter prediction from artworks with transfer learning, 7th Intl. Conference on Mechatronics and Robotics Engineering (ICMRE) (2021) 204–208. doi:10.1109/ICMRE51691.2021.9384828.

[26] C. Im, Y. Kim, T. Mandl, Deep learning for historical books: classification of printing

technology for digitized images, Multimedia Tools and Applications 81 (2022) 5867–5888. doi:10.1007/s11042-021-11754-7.

[27] W. Helm, S. Schmideler, C. Im, T. Mandl, S. Kollmann, Müller, Wie sich die Bilder ähneln, in: Fabrikation von Erkenntnis: Experimente in den Digital Humanities, 2022. doi:10.26298/melusina.8f8w-y749-wsdb.

[28] S. Lang, B. Ommer, Attesting similarity: Supporting the organization and study of art image collections with computer vision, Digital Scholarship in the Humanities 33 (2018) 845–856. doi:10.1093/llc/fqy006.

[29] Y. Kim, T. Mandl, C. Im, S. Schmideler, W. Helm, Applying computer vision systems to historical book illustrations: Challenges and first results, in: Post-Proceedings 5th Conference Digital Humanities in the Nordic Countries (DHN 2020), Riga, Latvia, Oct. 21-23, volume 2865 of *CEUR Workshop Proceedings*, 2020, pp. 255–260. URL: https://ceur-ws.org/Vol-2865/poster7.pdf.

[30] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, CoRR abs/1905.11946 (2019). URL: http://arxiv.org/abs/1905.11946. arXiv:1905.11946.

[31] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. M. Karimi, S. Nandanwar, S. Bhattacharyya, S. Rahimi, An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives, Electronics 12 (2023). doi:10.3390/electronics12051092.