Vision Transformers and their applications as discriminators - Training models to detect AI manipulated Art

1st Evangelos Papamitsos
National Technical University of Athens
Athens, Greece
vpapamitsos@gmail.com

***Abstract*—In this paper we continue our research on the applications of Vision Transformer models on detecting AI generated content but in a more practical scope, experimenting specifically on detecting AI generated art produced by generative models. With these experiments we aim to evaluate the model's performance in the task and gather useful insights. We will also experiment and compare with the more recent ConvNeXt, which became the state-of-the-art in many Vision tasks, dethroning the Transformer architecture.**

***Index Terms*—Vision transformers, ConvNeXt, AI manipulated Art**

I. The Dataset

The images we will use are taken from [1], (***kaggle link***) Specifically, we will use the Inpaintings and Style Transfer categories. Inpaintings are images that a part of them has been masked and has been re-generated by the generative model ***Stable Diffusion 2***. The Dataset contains 6063 pairs of original and generated images (inpaintings) as shown in Fig. 3.

Similarly, the Style transfer category contains 3074 pairs of original and generated images. In this case, the images are generated by the ControlNet model [8] by keeping the "edges" of the original painting but altering the style of the painting. The original images are selected from [7] and contains various painting genres such as Abstract Expessionism, Analytical Cubism, Minimalism and more. An original and genreated pair can be seen in Fig. 4.

II. The Models

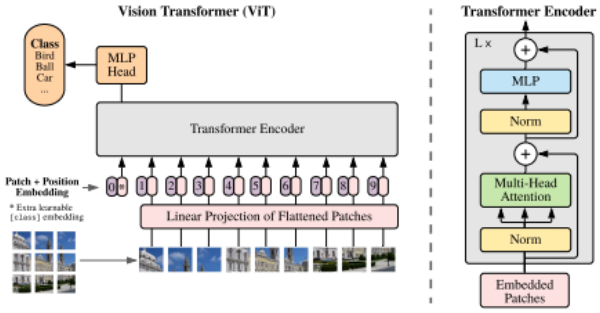

Fig. 1. Vision Transformer Architecture

The Vision Transformer model that we will fine tune is ***ViT-Base*** introduced in the original Vision Transformer Paper [2]. The model follows the standard Vision Transformer architecture we have already discussed [6] and is pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224. It has 86M parameters, 12 layers (a layer is shown in Fig.1.) and is the smallest compared to Vit-Large and ViT-Huge.

The ConvNext model that we will fine tune is ***ConvNeXt-Tiny*** introduced in [5]. ConvNeXT marks the return of Convolutional Networks as it is a pure convolutional model (ConvNet), inspired by Vision Transformers, and seemingly outperforming them. Starting from a ResNet [3] and improving the design getting inspiration by the Swin Transformer [4] ,we can see a picture of the architecture below.

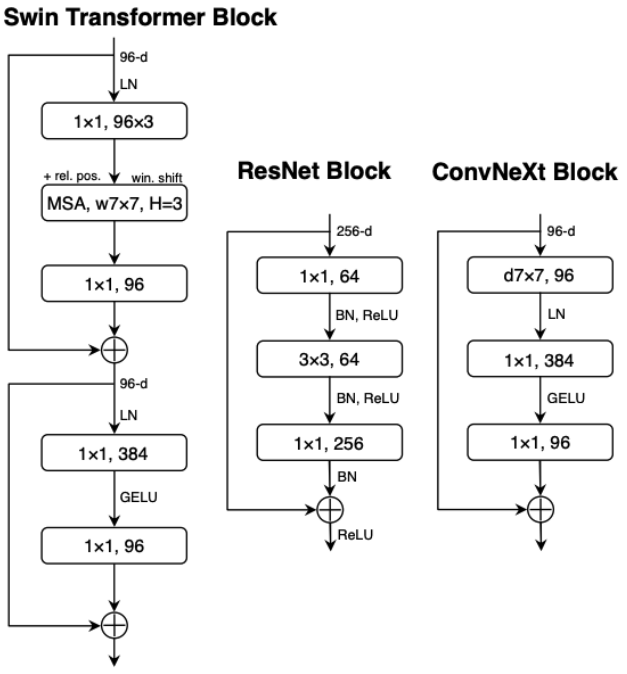

Fig. 2. Enter Caption

The Tiny model has 29M parameters, a lot less than VitBase.

III. The experiments

For the experiments we fine tune each model (***ViT-Base*** and ***ConvNeXt-Tiny***) on each of the sub-datasets (Inpainting and Style Transfer) for 4 epochs. The models are loaded from Hugging Face and the Training and preprocessing code is available in ***my github repo***.

A. Training results

Let's view the results in the tables below:

TABLE I
INPAINTING DATASET RESULTS ON TEST

Model	Metrics			
	Accuracy	Precision	Recall	F1 Score
ViT-Base	0.8713	0.9026	0.8323	0.866
ConvNeXt	0.841	0.7616	0.9927	0.8619

As we can see the models perform fairly well with 0.8713 and 0.841 on Test for ViT-Base and ConvNeXt respectively. One interesting takeaway is that we the ViT model seems to have greater accuracy than ConvNeXt but also much greater ability to accurately detect the original photos (Precision is the ratio of correctly predicted positive observations to the total predicted positives. In this case positive corresponds to Original images). On the other hand, ConvNext is successfully identifying almost all of original images but doing so it compromises in accuracy (False positives).

Let us now compare the models on the Style Transfer category:

TABLE II
STYLE TRANSFER DATASET RESULTS ON TEST

Model	Metrics			
	Accuracy	Precision	Recall	F1 Score
ViT-Base	0.9846	0.9881	0.9798	0.9839
ConvNeXt	0.9895	0.9882	0.9899	0.9890

As we can see the models perform almost perfectly in this task with much greater accuracy across the board for both models.

B. Introducing Data Augmentation

Due to the lower accuracy in the Inpainting Dataset in contrast to the Style Transfer(but also in contrast to the Training accuracy which was close 0.96 and 0.99 respectively) we introduce Data Augmentation aiming to help the models generalise better. The Data augmentation technique we use is Horizontal Flip and slight Color Jitter (slight random changes in contrast, hue, saturation and brightness).

TABLE III
INPAINTING DATASET RESULTS ON TEST WITH DATA AUGMENTATION

Model	Metrics			
	Accuracy	Precision	Recall	F1 Score
ViT-Base	0.8543	0.8878	0.811	0.8477
ConvNeXt	0.9318	0.9470	0.9148	0.9306

As we can see the Data Augmentation had very positive results in the ConvNeXt increasing the accuracy on test by almost 0.1 . Surprisingly, these techniques had negative impact on the ViT model dropping the accuracy by almost 0.02 .

The Fine-Tuned models as well as the dataset are available for direct use in ***my hugging face repo*** and the code and extra material (presentation, poster and testing with new data) in ***github***.

IV. Conclusion and further goals

To conclude our research, we have seen that ViT and ConvNeXt models can be used effectively for AI manipulated image detection of specific type (namely Inpainting and Style Transfer). There is certainly room for improvement especially in the ViT model in the Inpainting category possibly with the use of different and more complex augmentation techniques. We can also experiment by Training a single model aiming to detect AI manipulated data in a more general scope, detecting images that have been generated from various generative models and different techniques. In the context of comparing the two models, we can say that ConvNeXt-T has the edge over ViT-Base due to the improved performance in the Inpainting Dataset with Data Augmentation but also due to the smaller size.

References

[1] Hossein Aboutalebi, Daniel Mao, Carol Xu, and Alexander Wong. Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection. *arXiv preprint arXiv:2306.01272*, 2023.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[6] Vaggelis Papamitsos. Vision transformers and their applications as discriminators. 2022.

[7] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

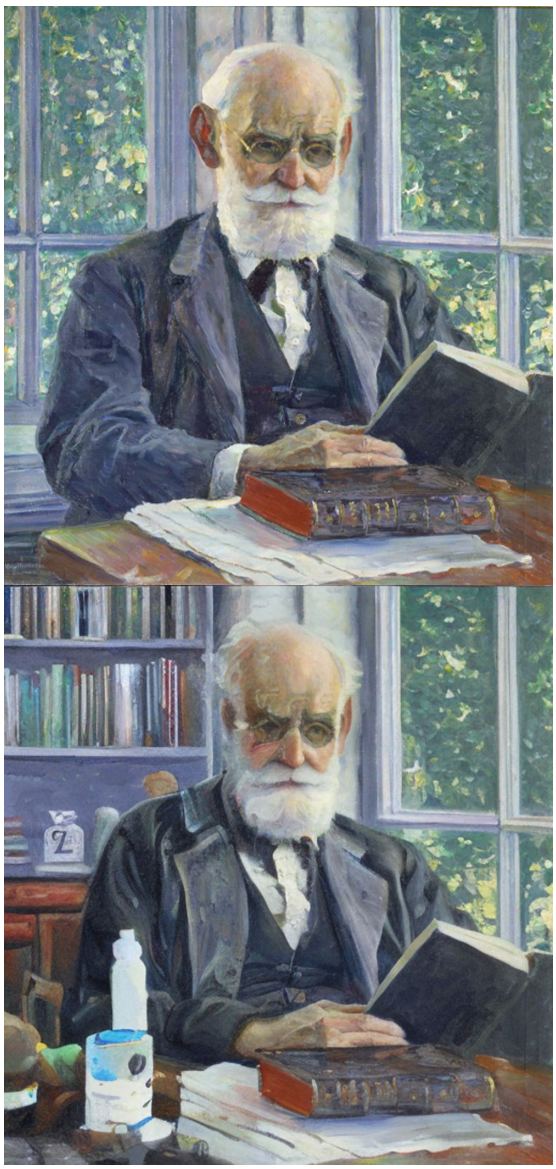

Fig. 3. Original and Inpainting image from the Dataset

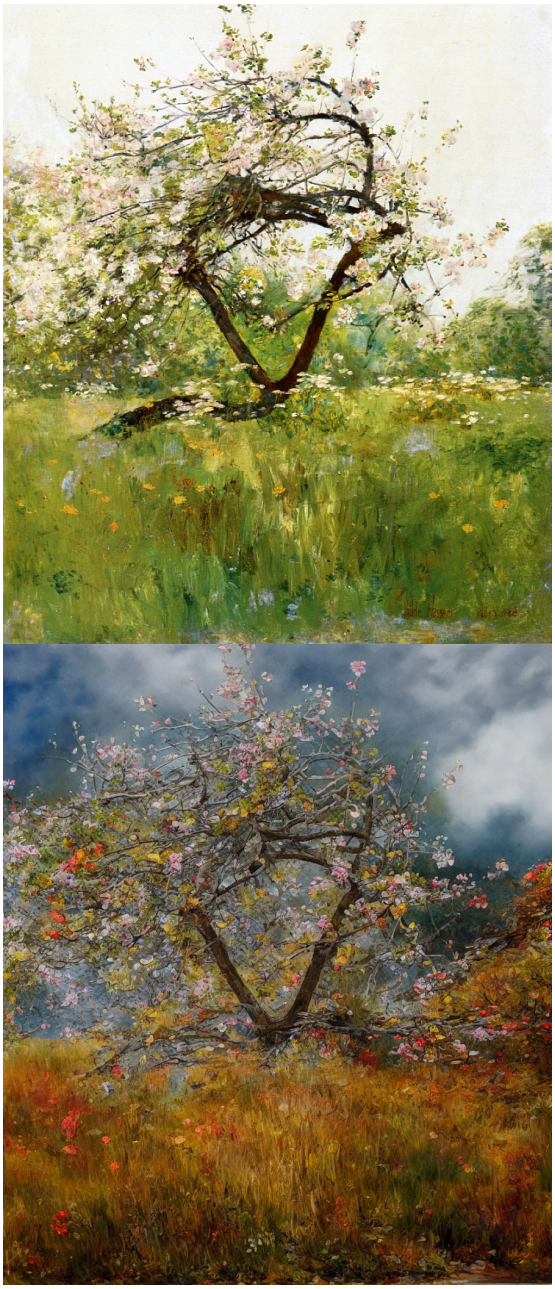

Fig. 4. Original and generated image from the Dataset