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

Vaggelis Papamitsos

National Technical University of Athens

Overview

As Generative models ,such as Stable Diffusion, DALLE, Control Net and more are gaining popularity, it is also important to have models with the ability to distinguish between original and AI generated or manipulated images. Our goal is to specifically experiment with the performance of Vision Transformer models in such tasks, focusing on images of Art paintings and also compare the performance with the more recent ConvNeXt model which has dethroned ViTs as the state-of-the-art in many Vision Tasks.

A Glimpse of the Data

The images we will use are taken from the DeepfakeArt Challenge Dataset and specifically 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 Figure 1.

Similarly, the Style transfer category contains 3074 pairs of original and generated images. In this case, the images are generated by the ControlNet model by keeping the "edges" of the original painting but altering the style of the painting. The original images are selected from WikiArt and contains various painting genres such as Abstract Expessionism, Analytical Cubism, Minimalism and more. An original and genreated pair can be seen in Figure 2.

Figure 1. Original and Inpainting image from the Dataset

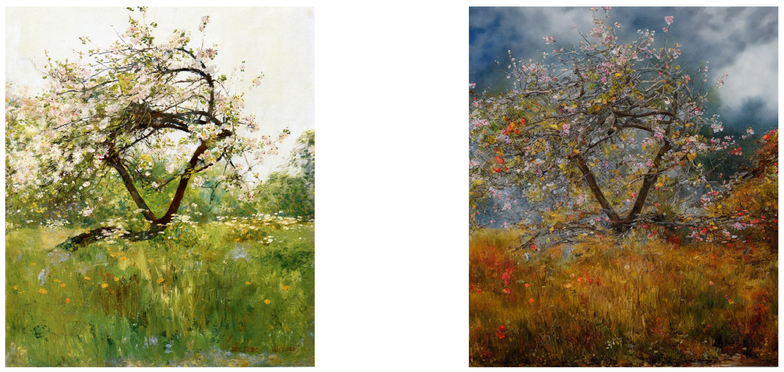

Figure 2. Original and generated image from the Dataset

The Models

The Vision Transformer model that we will fine tune is ViT-Base introduced in the original Vision Transformer Paper. The model follows the standard Vision Transformer architecture we can see in Figure 3. 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 and is the smallest compared to Vit-Large and ViT-Huge.

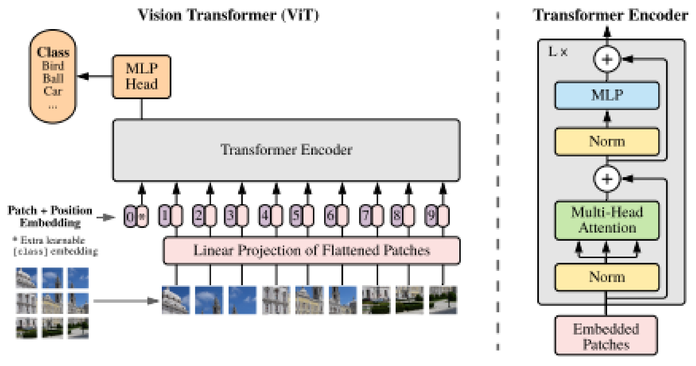

Figure 3. Vision Transformer Architecture

The ConvNext model that we will fine tune is ConvNeXt-T. 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 and improving the design getting inspiration by the Swin Transformer ,we can see a picture of the architecture in Figure 4.

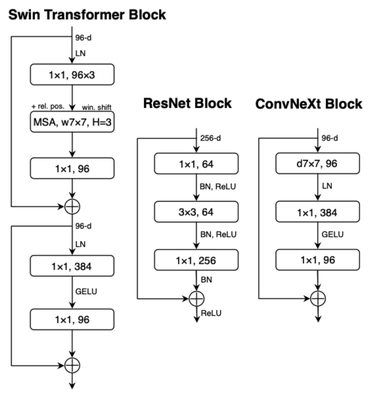

Figure 4. ConvNeXt Architecture

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

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 form **https://huggingface.co** and my Training and preprocessing code is available in **https://github.com/VaggelisPap/Vision-Transformers-as-Discriminators**.

Let's view the results in the tables below:

Table 1. 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 2. 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.

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 3. 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 **https://huggingface.co/VaggP** and the code and extra material including testing with new data in **https://github.com/VaggelisPap/Vision-Transformers-as-Discriminators**.

vpapamitsos@gmail.com