Introduction

Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. While the act of creating fake content is not new, deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content that can more easily deceive. The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders or generative adversarial networks (GANs). Deepfakes are a media of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information. Uses advanced artificial intelligence techniques to manipulate or generate media and can be used to manipulate public opinion around the world.

The Problem

Deepfake technology can lead to privacy issues, generating fake news and spreading misleading information. It can easily be used to diverse public opinion and can pose serious threats. It's easy to generate due to wide variety of applications and many open-source projects. The number of deepfake videos online has been increasing at an estimated annual rate of about 900% which can be a problem later to identify the real or fake media. It is important to have a system to detect deepfake media as it can help to stop fake news, misinformation and can help prevent cyber bullying, harassment etc.

Dataset

The dataset was taken from Kaggle which consists of 140K Real and Fake Faces. Out of which 70K are Fake and 70K are Real Images. The dataset consists of images divided into training, validation, and test folders.

Technologies used to Generate Deepfakes

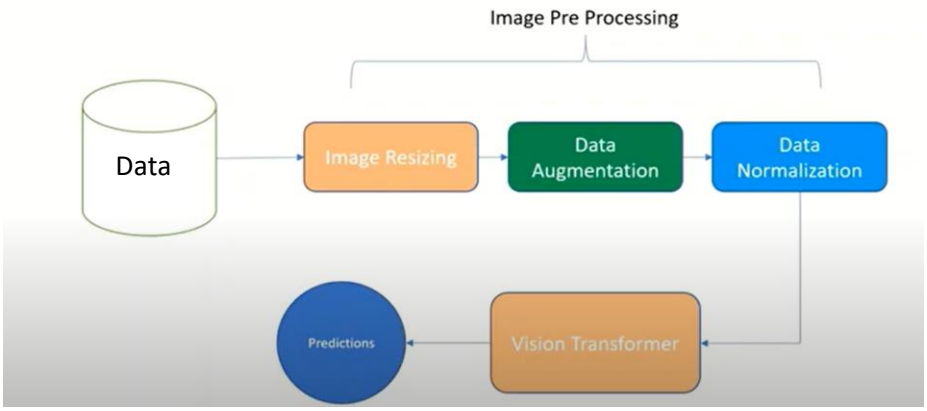
Autoencoders: It is made of a pair of two connected neural networks. The encoder finds and learns similarities between the two images and reduces them to their shared common features, compressing the images in the process. A second AI algorithm called a decoder is then taught to recover the faces from the compressed images.

GAN (Generative adversarial network): Is another deep learning technique to generate deep fake. It is made up of 2 parts a generator that learns to generate plausible data and a discriminator that learns to distinguish the generators fake data from real data

Proposed Solution

Transformer architecture – Vision Transformer

A pretrained Convolution Neural Network

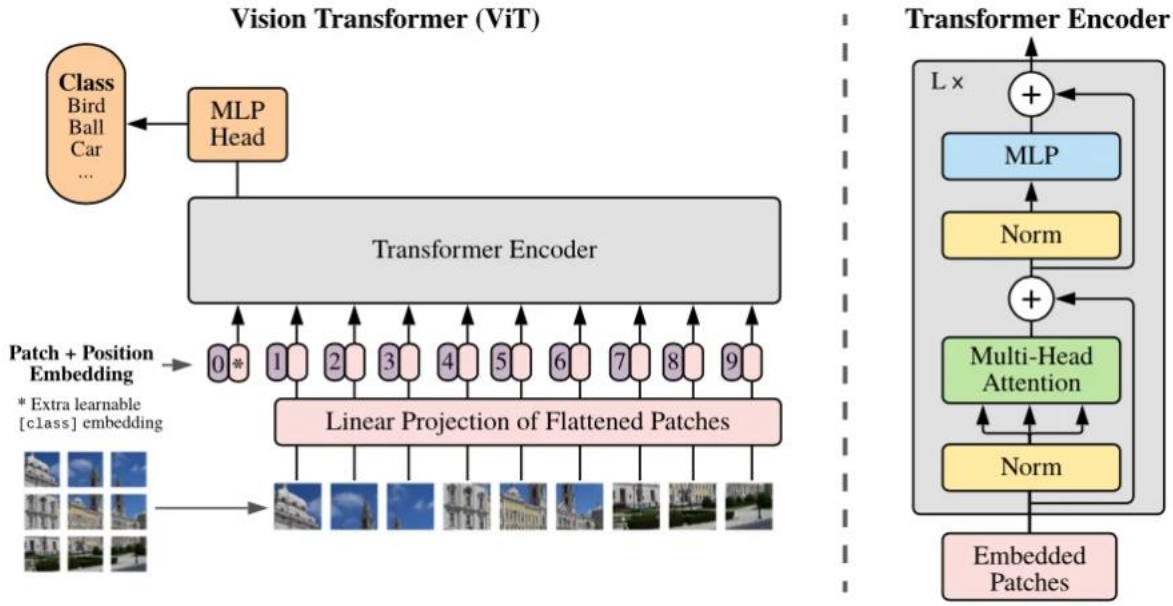


Difference between CNN and Vision Transformer

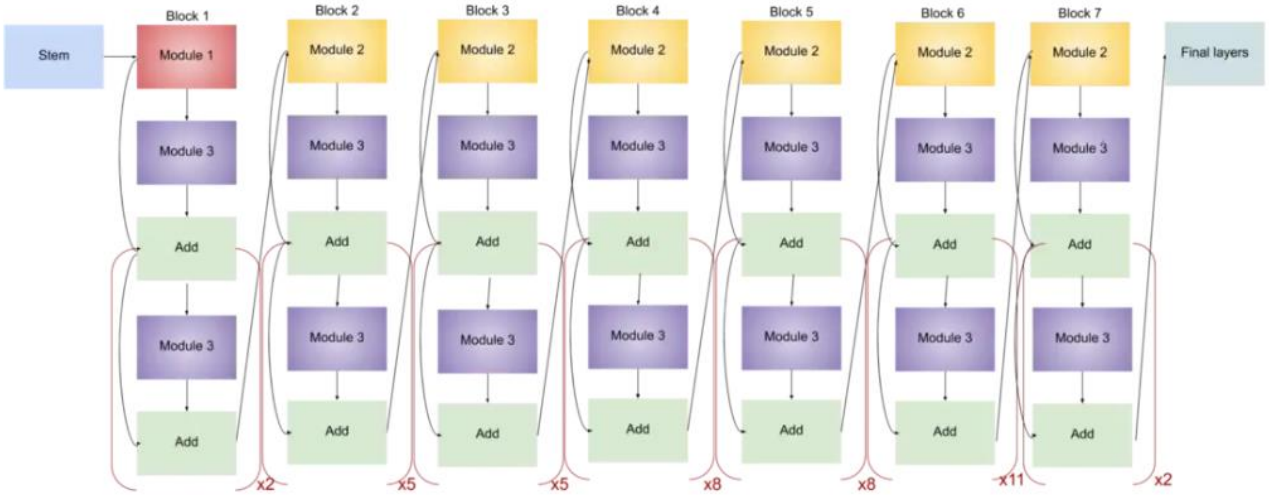
Vision Transformer (ViT) achieves remarkable results compared to convolutional neural networks (CNN) while obtaining fewer computational resources for pre-training. In comparison to convolutional neural networks (CNN), Vision Transformer (ViT) show a generally weaker inductive bias resulting in increased reliance on model regularization or data augmentation (AugReg) when training on smaller datasets.

Vision Transformer Architecture

1. Split an image into patches (fixed sizes)
2. Flatten the image patches
3. Create lower-dimensional linear embeddings from these flattened image patches
4. Include positional embeddings
5. Feed the sequence as an input to a state-of-the-art transformer encoder
6. Pre-train the ViT model with image labels, which is then fully supervised on a big dataset
7. Fine-tune the downstream dataset for image classification



EfficientNet-B7



Architecture of EfficientNet-B7

Before feeding our convolutional neural network with train and test samples, image samples must be pre-processed. The images are firstly resized to 224 by 224, and we convert them from greyscale to RGB space by repeating the intensity values across all three channels. The process then reads the image in RGB format and applies pixel normalization. Once the image pre-processing has been completed, our convolutional neural network is ready to accept the input data. Before feeding data to the CNN, the training data goes through data augmentation stage, which increases the diversity of dataset without the need to collect more data.

Data Preprocessing

To carry out training I extracted a total of 10000 images from the original dataset. The extracted images were equally divided into the respective classes in such a way that we had 4000 training images for real face image and 4000 images for deep fake images. The test set consist of 2000 images equally divided among the two classes, like the training set as shown in the table below.

Data Set	Real Face Images	Fake Face Images
Training Set	4000	4000
Test Set	1000	1000

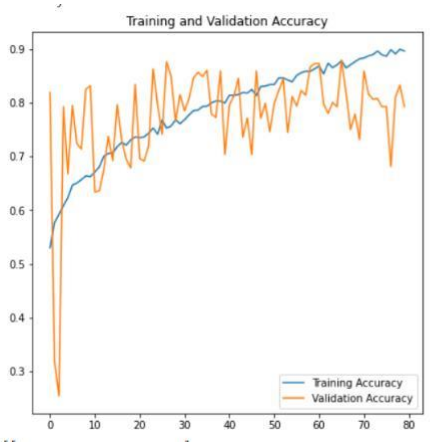
The images containing real and fake faces go through a normalization process, resizing by 224x224 pixels, and data augmentation process before being saved in a dataset that will be used by both the vision transformer, and the Efficient Net B07 Architecture.

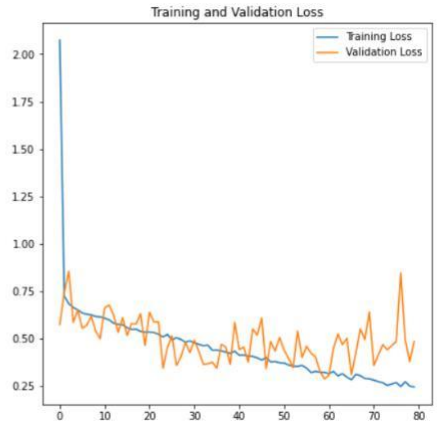
Vision Transformer Architecture Accuracy

The summary of accuracy scores is displayed in table 2 below. According to our results, the visions transformer architecture achieved an overall accuracy of 80.7 %. The vision transformer classifier achieved a precision, recall and f-1 score of 80.5 %.

The training and validation curves for our classifier.

	Precision	Recall	F1-Score
Real Face	80%	83%	81%
Fake face	82%	79%	80%

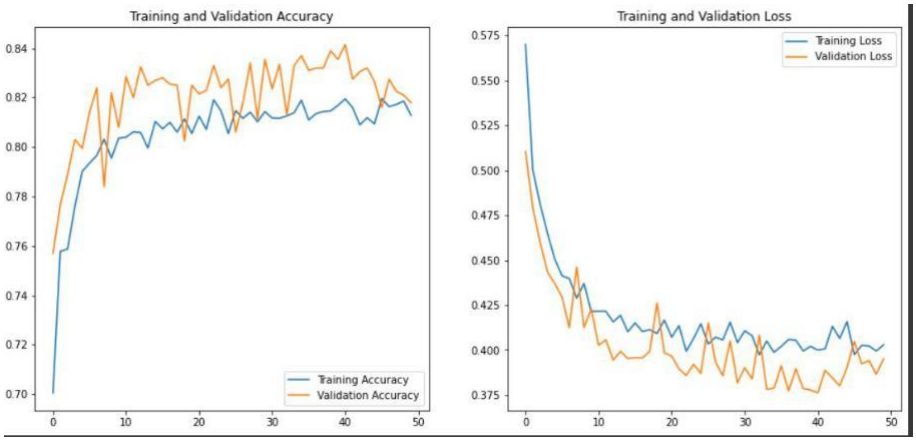




Efficient Net B07 Architecture Accuracy

Variations in our initial pipeline included a different architecture for classification purposes. We feed the images to a convolutional neural network based on Efficient Net B07 architecture using the same image pre-processing and normalization techniques. Table below shows the overall accuracy scores of our CNN architecture. Efficient Net classifier was able to achieve an overall accuracy of 81.8 %. This entails that the Efficient net architecture outperformed the vision transformer architecture by 1.1 % in terms of accuracy. Efficient Net architecture achieved an overall precision score of 83%, a recall of 82%, and an f1 score of 81.5 %. Figure 6 below shows us the training and validation curves for the Efficient Net B07 architecture.

	Precision	Recall	F1-Score
Real Face	89%	73%	80%
Fake face	77%	91%	83%



Conclusion

I used two distinct classifiers for deep fake recognition. The classifier in the initial method was a vision transformer. In contrast, a variant of the same pipeline that was employed for deep fake picture identification utilized the Efficient Net B07 architecture. The study offered two deep learning strategies with a range of benefits and a respectable accuracy rating. The accuracy of the vision transformer model was 80.7%, while the accuracy of the CNN-based architecture was 81.8%. Image and video manipulation detection is becoming more important in digital forensics. As a result, techniques for creating systems to identify deep fake photographs can also be the foundation for creating standardized tools for classifying and detecting forgeries as well as for detecting deep fake images. Although the outcomes of this study can be deemed adequate, we can get better results with more data and hyperparameter optimization.

References

https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142

https://viso.ai/deep-learning/vision-transformer-vit/

https://www.kaggle.com/code/ahmederaky/real-vs-fake-images/data