Deepfake Image Detection using Vision Transformer Models

GHITA, Bogdan, KUZMINYKH, Ievgeniia, USAMA, Abubakar, BAKHSHI, Taimur and MARCHANG, Jims <http://orcid.org/0000-0002-3700-6671>

Deepfake Image Detection using Vision Transformer Models

Bogdan Ghita
University of Plymouth
Plymouth, UK
bogdan.ghita@plymouth.ac.uk

Ievgeniia Kuzminykh
King's College London
London, UK
ievgeniia.kuzminykh@kcl.ac.uk

Abubakar Usama
University of Plymouth
Plymouth, UK
usamacheema0007@gmail.com

Taimur Bakhshi
Leeds Beckett University
Leeds, UK
t.bakhshi@leedsbeckett.ac.uk

Jims Marchang
Sheffield Hallam University
Sheffield, UK
jims.marchang@shu.ac.uk

***Abstract*—Deepfake images are causing an increasing negative impact on the day to day life and pose significant challenges for the society. There are various categories of deepfake images as the technology evolves and becomes more accessible. In parallel, deepfake detection methods are also improving, from basic features analysis to pairwise analysis and deep learning; nevertheless, to date, there is no consistent method able to fully detect such images. This study aims to provide an overview of existing methods of deepfake detection in the literature and investigate the accuracy of models based on Vision Transformer (VIT) when analysing and detecting deepfake images. We implement a VIT model-based deepfake detection technique, which is trained and tasted on a mixed real and deepfake images dataset from Kaggle, containing 40000 images.**

The results show that The VIT model scores relatively high, 89.9125%, which demonstrates its potential but also highlights there is significant room for improvement. Preliminary tests also highlight the importance of a large dataset for training and the fast convergence of the model. When compared with other deepfake machine learning and deep learning detection methods, the performance of the ViT model is in line with prior research and warrants further investigation in order to evaluate its full potential.

***Index Terms*—deepfake images, deepfake detection, Vision Transformer model**

I. Introduction

The recent advancements in artificial neural network (ANN) technologies have had a significant impact on multimedia content manipulation. AI-based software tools, allowing users to modify facial appearance, hairstyle, gender, age, and other personal attributes, have facilitated the creation of realistic fake images, videos, and audios. The widespread availability of these tools and the manipulated content they produced were coined in 2017 with the term "Deepfake", derived from "Deep Learning (DL)" and "fake," and it encompasses applying deep learning methods to generate very realistic appearing (fake) content. The advent of deepfakes was facilitated by the increased complexity and capabilities of computer vision and deep learning techniques. While it can be employed towards legitimate, creative use, it has been typically misused by users to create fake news or fake images. [1]

Its misuse was also followed by significant effort from the research community both to establish more realistic images as well as developing techniques to detect them. Deepfake images can be catalogued based on the focus and degree of change into entire face synthesis, identity swap, attribute manipulation, and expression swap. From a technology perspective, deepfakes use unsupervised ANNs named autoencoders, which are, in fact, used for both image manipulation as well as facial recognition, by being able to both synthesise the facial features into a set of characteristics as well as modify images based on their defining features. The encoding process can be further improved through the use of a generative adversarial network (GAN).

From a detection perspective, the early efforts were based on image features analysis, such as pixel similarity and noise, to identify artifacts of the modification process. Such approaches were effective towards early instances but were surpassed by Autoencoders and GANs to replicate the modification process. In parallel, a number of classification models may also be employed for deepfake detection purposes. One obvious example of such a model are Visual Transformers (ViT), a classification approach derived from natural language processing which allows interpretation of images as an analysis array. Although initially used for image classification, ViT can also be applied for deepfake detection, particularly given its algorithm-agnostic approach and ability to handle very large inputs. This study expands the work so far in ViT by applying the model for classifying an image dataset into real and deepfake images. We evaluate the model against a small image set and discuss its classification performance, as well as identify a number of limitations and potential for future work.

II. Deepfake generation and detection

A. Deepfake generation techniques

Since its inception, the generation of deepfake content has had a negative effect across all levels of society [2], from its prevalence into everyday social media [3] to its ability to disrupt international politics [4] [1].

Although there is a wide variety of approaches for generating deepfakes, they revolve around the concept of image (typically face) features extraction and reconstruction. The process, implemented through an autoencoder-decoder. uses multiple images for training to extract and recompose these features. As part of the process, the deepfake model is trained to parse a set of images, extract their salient features, then reconstruct each image as accurately as possible using the extracted features. Once the model is trained for extraction and reconstruction, it can be applied to cross-convert images. For example, in order to replace the face of a subject, it can extract the image features of subject X and then reconstruct the image using features from a different subject Y.

While autoencoder-based models will produce very good results, it is designed with efficiency in mind rather than performance and will therefore allow for either pixelation defects or blurriness, both detectable by users or automated methods. Generative Adversarial Networks (GANs) [5] have better intrinsic capabilities to identify and remove artifacts and were therefore proven to significantly reduce noise and improve the quality of the resulting deepfake images. [6].

More recently, newer deepfake approaches applied existing techniques for more efficient and realistic results. Variational Autoencoders (VAEs), proposed in [7] and [8], allow the inclusion of more complex models built from larger datasets. VAEs include a reconstruction loss monitoring component, which aims to minimise the loss value resulting from the process, and a *regulariser* component, that ensures diversity in the outcome. Similarly, an Adversarial Autoencoder (AAE) [9] draws in the benefits of GANs but relies on the autoencoder training to extract the distribution of the data rather than impose it on the output layer. As shown in the study that introduced the concept, AAE is net superior to VAE and incrementally better than GAN in terms of the error rate applied to standard evaluation images datasets.

B. Deepfake detection techniques

As pointed out in the previous subsection, deepfake generation has been through an evolutionary process; similarly, detection techniques followed that trajectory. The early methods were based on machine learning and aimed to detect artifacts and defects in the generation process. A typical such example is [10], where the authors aimed to identify the convolutional lines produced by a deepfake autoencoder and achieved a high accuracy of 93%. Similarly, [11] looked at artifacts introduced in the process such as global consistency, illumination, and geometry, focusing on the particular characteristics of face, such as iris and teeth characteristics. The method performs very well with an AUC-ROC of 0.83 when combining all observed features. The main challenge in both studies is the nature and availability of the images used, as the database is sufficient to train the detection models and capture all the artifact variances.

As content became more realistic and more effective at eliminating image artifacts, the detection methods aimed to replicate the generation ones and became deep learning-based approach. A typical example is AutoGAN, the approach from [12], which is essentially a GAN that replicates the deepfake process. AutoGAN takes the image used as input and a GAN-based generator to produce an image following the same principles as deepfake; the image is then compared to the original to detect spectral artifacts. The method relies on the GAN requiring a variant of upsampling, either transposed convolution or nearest neighbour interpolation, that do produce spectral artifacts that can be learned. The method delivers a tangible improvement to cycleGAN, reaching accuracy of over 95% for both transposed and nearest neighbour interpolation.

C. Conclusion

As highlighted by the results of the studies in the previous subsection, detection techniques mimic the deepfake generation models in order to expose artifacts, defects, or other types of data errors in the input images and successfully segregate real and fake images. The results are very good, with methods achieving accuracy of over 85-90% across the datasets tests.

One common point to the detection methods is their need and awareness of the model used for generating the fake images, as they are somewhat geared towards specific deepfake models behaviour. This issue is made clearer by some of the papers by highlighting that deepfake images generated through other or unknown methods may perform worse when analysed and, implicitly, the proposed methods may become obsolete with future variants of deepfake models. It is therefore worth exploring the wider field of image recognition techniques and investigate their potential for usage as a deepfake method-agnostic detection alternative.

III. VT-based deepfake detection

A. Introduction

Vision Transformer (ViT) is an image analysis approach proposed in [13], based on the concept of Transformer introduced by [14]. Transformers consist of alternative self-attention and multilayer perceptron blocks; they were initially aimed at natural language processing and relatively small models, but subsequent research demonstrated that it can be scaled to very large models of 10^{11} parameters [15]. An early attempt by [16] saw self-Transformers applied to resized images on a pixel-by-pixel basis. In their paper, Dosvitskiy et al. apply a variant of the Image Transformer but, instead of smaller images and pixel-by-pixel analysis, they split the image into fixed-size patches which are fed to a Transformer. In an NLP scenario, the information is provided as a 1D sequence; for ViT the image is converted into a linear projection vector. The Transformer itself follows the [14] design, with a slight adjustment to allow for position embedding.

Both papers acknowledged the ability of the technology to go beyond image recognition to identifying generated images as a possible application, with [16] also providing some preliminary results.

This aim of this research is to determine the efficiency of ViT-based algorithms when classifying real vs fake images. In

order to observe their performance, we used the image classification work described in [13] and evaluated its performance on the [17] image dataset from Kaggle.

B. Text and image classification

The ViT model is inspired by the the standard Transformer from [14]. The Transformer model has *self-attention* at its core, a multilayered stack of feed forward and multi-head attention blocks. Both the encoder and the decoder use a stack of 6 identical layers, each composed of a multi-head self-attention mechanism and connected to a feed-forward network. Amongst the two types of attention, the model uses *Scaled Dot-Product Attention* due to its more efficient computation. The authors exploited its ability to be scaled in a Multi-Head attention architecture, whereby the queries and results are parallelised and producing vectored values. As tested by the authors, Transformer models work very well with text, being able to outperform previous models in English-to-German translation.

While effective at processing text-based input, Transformers were not intrinsically able to handle 2D content; [16] extended the model with position embedding information and named the architecture Image Transformer. The authors proposed a pixel-by-pixel approach, whereby the representation q' of a channel for a pixel q is derived based on self-attention to the memory of the previously generated pixels. Unlike text-based input, scaling for Image Transformers was unattainable for a larger image, hence the authors biased the model with a level of locality, whereby pixel values were derived mostly from their vicinity rather than the entire picture, property termed *Local Self-Attention*.

In their paper, Dosovitskiy et al. revisited the concept of image transformer by dividing the image into patches. [13]. The overall image $\mathbf{x}$ of resolution (H, W) and C channels can be reshaped from $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ to an array $\mathbf{x}_p$ of $N = HW/P^2$ patches, with $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$.

It is interesting to note that ViT does not heavily rely on locality. Instead, the two components are adjusted to exploit their characteristics as follows: the multilayer peceptron layers are local and the self-attention layers are global. As a result, locality is shared between the two types of layers.

C. ViT implementation and dataset

We implemented the ViT model in python, following the ViT specification from [13]. The implementation takes a dataset as input, including a mix of real and fake images, then it applies patching to each image and encodes the positioning of each patch. The model includes the multi-head attention, normalisation, and MLP layers that take in the inputs.

The ViT model requires a set of parameters for training, testing, and speed optimisation, as listed in table I above. The size and number of inputs are dictated by $image\ size$ and $patch\ size$, which set the size of the input images and patch images, both measured in pixels.

At the core of the model performance are the $learning\ rate$, for which smaller values typically lead to slower but more accurate convergence, and the $weight\ decay$, which is the penalty for the loss function; the two parameters also influence the $number\ of\ epochs$ for training the model. The speed of the model is also influenced by the $batchsize$ which dictates the level of parallelism for the processing of the model.

TABLE I
VIT MODEL PARAMETERS

Parameter	Value
learning rate	0.001
weight decay	0.0001
batch size	256
number of epochs	800
image size	72
patch size	6
projection dimension	64
transformer layers	5

Fig. 1. Processed and resized image: full (left) and split into patches (right).

IV. MODEL PERFORMANCE

The model was trained and tested on subsets of images, including a mix of fake and real samples, from a Kaggle image dataset containing 190,345 images. [17]

All input samples are the same size, 256x256 pixels. The preprocessing takes each sample, converts it to 72x72 pixels, then splits it using the patch size, as per the specified ViT model parameter, into 144 patcehs, 6x6 pixels each. An example of the converted full image and set of patches is show in 1. Each patch is fed to the model with its positional information.

The model was trained and tested on a Google Colab L4 cloud instance, equipped with 64GB of RAM, 24GB of GPU RAM, and a powerful GPU, designed specifically for deep learning applications [18]. Each processed dataset was split 80/20 for training and testing.

A. Preliminary test

The training process is a rather computationally demanding one. The preliminary evaluation aimed to determine the learning speed of the model, to avoid excessive training and optimise the use of computational resources. For this, we used a subset that included 5850 images (2531 deepfakes and 3319 real). The reason for the subset is pragmatic - we aimed to evaluate the model on a manageable subset from an accuracy and training perspective.

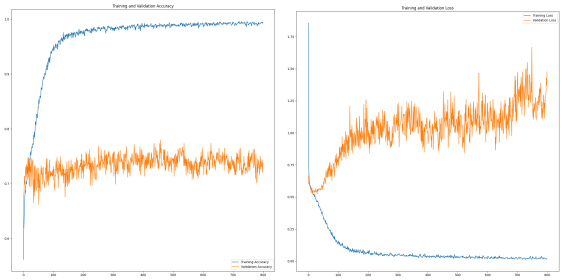

Fig. 2. Evolution of accuracy (left) and residual loss (right) during training of the preliminary dataset

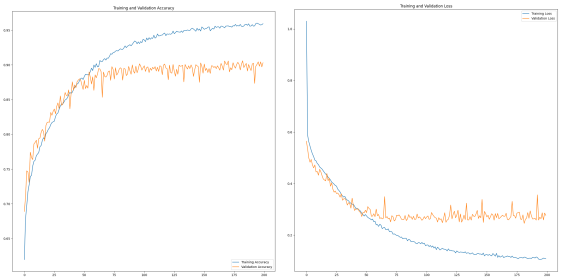

Fig. 3. Evolution of accuracy (left) and residual loss (right) during training of the complete dataset

The model was trained and tested subsets of 5850 images, including a mix of fake and real samples, from a Kaggle image dataset containing 190,345 images. The subset included 2531 deepfake images and 3319 real images. All input samples are the same size, 256x256 pixels. The reason for the subset is pragmatic - we aimed to evaluate the model on a manageable subset from an accuracy and training perspective.

The preprocessing takes each sample, converts it to 72x72 pixels, then splits it using the patch size, as per the ViT model, into 6x6 pixels patches. An example of the converted full image and set of patches is show in 1. Each patch is fed to the model with its positional information.

The training was set to 800 epochs and took just under 1 hour, with just over 4s per epoch.

The evolution of both accuracy and residual loss through the testing process is shown in Fig. 2. As it can be seen, the model performs very well, with accuracy close to 1, for the training dataset but, for the validation dataset, the accuracy remained rather low, 0.75. Looking at the loss values, the training and validation sets behave similarly for the first 100 epochs, then their evolution changes. The fit of the model to the training set is monotonously improving, reaching residual loss close to 0, while for the validation set the residual loss is increasing; as a result, while the loss stabilises for training, it becomes increasingly oscillating for validation. Coming back to the accuracy diagram, there is no tangible improvement in the accuracy for the validation dataset beyond 100 epochs, despite the better results for the training dataset.

The results confirm the behaviour of ViT models if we take into consideration the size of the dataset. ViT models require large datasets for training, followed by fine-tuning on smaller datasets. Compared to ILSVRC-2012 ImageNet, the smallest dataset [19] used by [13] which included 1.3 million samples and over 1000 classes, despite the decrease in the number of classes from 1000 to just two (real and deepfake), the ViT model still requires a larger training dataset for better accuracy.

B. Model evaluation

A larger subset of 40000 images, 20000 deepfakes and 20000 real, was used to fully train the implemented ViT model on an L4 Google Colab instance. We were not able to use the full Kaggle dataset due to two reasons - length of time for training and, more importantly, the L4 instance failing to process larger datasets. Given the results from the preliminary tests, we set the training to 200 epochs, which required 95 minutes of processing.

TABLE II
VIT FULL TESTING RESULTS

	precision	recall	f1-score	support
class				
real	0.87	0.93	0.90	4015
fake	0.93	0.87	0.90	3985
accuracy				
macro avg	0.90	0.90	0.90	8000
weighted avg	0.90	0.90	0.90	8000
Accuracy: 89.9125				

Accuracy and residual loss had a similar evolution to the preliminary test, as shown in Fig. 3. The overall full results are shown in Table II.

The model performs very well with an overall accuracy of 89.9125%. There are some minimal variations between the real and fake classes detection, but nothing significant. Overall, the model is slightly biased towards deepfakes, with a higher percentage of both accurate detection of deepfakes as well as identifying real images as deepfakes. Looking at the other studies in the area, we can compare our model with the summary provided by [20] which looked at the results from 62 studies aiming to detect deepfake images. According to the overall figures, deep learning models deliver an average accuracy of 89.73% and machine learning provide an average accuracy of 86.86%. Our accuracy therefore does match closely the deep learning category, but it is worth noting the dataset size limitations we encountered during training and therefore the likely possibility to reach better results with a larger dataset.

V. CONCLUSION AND FUTURE WORK

Transformer models are a combination of self-attention and multilayer perceptron blocks, notable for their ability to handle Natural Language Processing tasks. Vision Transformer models are an expansion of Transformer models, whereby the input is a serialised patches carrying locality information. This paper provides a practical evaluation of a Vision Transformer model

applied to the task of detecting deepfake images. We used an implementation that followed strictly the design of the ViT and was tested on a small dataset consisting of a combination of deepfake amd real images. The model delivered a 89.9125% accuracy, with slightly better results for the deepfake images.

The results showed that the accuracy of the model is significantly affected by the size of the dataset used and, due to implementation limitations, we were able to test only a subset of potential images. For future work, we aim to use larger datasets for evaluating the model, which are likely to deliver significantly better results, as well as further investigate the performance of the model when using different training parameters.

REFERENCES

[1] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security (july 14," Research Paper 692, 2018). 107 California Law Review 1753 (2019), U of Texas Law, Public Law, 2018-21.

[2] J. T. Hancock and J. N. Bailenson, "The social impact of deepfakes," *Cyberpsychol. Behav. Soc. Netw.*, vol. 24, pp. 149–152, Mar. 2021.

[3] S. Karnouskos, "Artificial intelligence in digital media: The era of deepfakes," *IEEE Transactions on Technology and Society*, vol. 1, no. 3, pp. 138–147, 2020.

[4] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," 2019.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[6] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, "Methods of deepfake detection based on machine learning," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 408–411, 2020.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[8] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.), vol. 32 of *Proceedings of Machine Learning Research*, (Bejing, China), pp. 1278–1286, PMLR, 22–24 Jun 2014.

[9] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[10] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[11] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, 2019.

[12] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2019.

[13] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[16] N. J. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning (ICML)*, 2018.

[17] M. Karki, *Deepfake and real images*. Kaggle, 2022.

[18] "Introducing G2 VMs with NVIDIA L4 GPUs — Google Cloud Blog — cloud.google.com." https://cloud.google.com/blog/products/compute/introducing-g2-vms-with-nvidia-l4-gpus. [Accessed 10-04-2024].

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[20] M. Rana *et al.*, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 1–1, 2022.