DeepFake Detection Using Wavelet Packets with Vision Transformer (WPT-ViT)

Osama Rawy and AbdulRahman AlTahhan

University of Leeds, School of Computing, ODL MSc in AI, UK.
WWW home page:https://github.com/osmahus/WPT-ViT

Abstract. The latest advances in generative algorithms have raised the quality of virtually created images and videos to the point that it has become very difficult to distinguish the real from the generated ones (Deep-Fake). This stimulated hot research to build better models to detect DF. Our paper proposes a new DNN model to detect DF images using a wavelet packet transformer and a vision transformer (WPT-ViT). The study shows that attention could be found between the WPT decompositions of an image even without slicing the image into spatial patches, which is a novel modification to the original VIT model. We showed that by using smaller model sizes and lower GPU and CPU requirements, we can achieve comparable results with previous work in this research area, The model was trained and tested using two datasets, "CIFAKE" and "140k Real and Fake Faces," which are generated using StyleGAN and Stable Diffusion algorithms.

Keywords: Deep Fake Detection, Wavelet Packets, Vision Transformer

1 Introduction

In 2014, Goodfellow et al. introduced the Generative Adversarial Network (GAN), marking the beginning of the generative AI (GAI) era. Since then, researchers have shifted their focus from discriminative learning to generative learning. This wave brought various vision-generative applications to the market, such as Midjourney, Firefly, DALL-E2, and Imagen [1]. Such applications were developed using state-of-the-art architectures like GAN, Variational Autoencoders, and Diffusion to generate images and videos with high fidelity and diversity, mimicking real-world photos and videos [2]. Vision-generative technologies have shown high value in several domains. In the entertainment industry, for example, they could generate complete scenes that would otherwise be very risky for actors to perform or prohibitively expensive to produce. In Education, they could bring historical characters to life to talk with students for an immersive learning experience; similarly, the list of positive uses continues in other fields like manufacturing and marketing. However, this ability to produce synthetic content with realistic flavor was termed Deep Fake due to the unfortunate incidents in which these technologies were used to attack people through identity theft, character assassination, and faked pornography. On a larger scale, deep fakes were also

used to spread misinformation, fake news, and communal hatred. According to a report released in April 2021 by Cybernews, deep fake content over the internet doubles every six months, posing a significant threat that needs to be addressed urgently [3].
For that reason, there has been significant attention in both academic and industrial fields on finding ways to detect deep fakes with high accuracy and performance. For example, Facebook, Microsoft, and Amazon collaborated to launch the Deep Fake Detection Challenge (DFDC) on Kaggle from 2019 to 2020. A survey conducted by (Liang & Xue, 2024) showed that the number of publications on deep fake detection surpassed the number of publications on deep fake generation in 2022 and 2023 [4].
This paper introduces a new deepfake detection tool that combines the strengths of wavelet analysis to extract important image features and Vision Transformer (VIT) to create lighter models with lower GPU and CPU requirements than CNN counterparts.

2 Literature Review

2.1 Deepfake Literature

When it comes to detecting deep fakes, it can be seen as a binary classification problem involving training a machine learning model on a dataset of real and fake examples. This includes extracting relevant features from the data and using these features to predict the authenticity of the content [3]. Previous research has suggested different ways to categorize work in the deepfake detection field. (Patel et al., 2023)[3] highlighted three approaches for detecting deepfakes in video and images.

The first approach involves using handcrafted algorithms to extract features of visual artifacts, such as inconsistent head poses or unusual eye blinking. The results of the feature extractor could then be passed to any classifier, such as SVM or NN, to perform the detection. An example of this approach is the work of Matern et al. (2019)[], which has an AUC of 0.866.

The second approach works on the pixel level to extract spatial features related to visual inconsistencies using local feature detectors (like SIFT and HOG) or steganography detectors. An example of this approach is the two-stream network proposed by (Zhou et al., 2018) with an AUC of 0.927. However, the effectiveness of the first two approaches has been reduced by the fact that the latest deepfake datasets were created using advanced image generation techniques, which decreases the likelihood of producing visual artifacts or detectable local features.

The third approach utilizes Deep Neural Networks to understand the intricate patterns and features present in the training dataset. The detection results are more accurate when a more relevant and comprehensive dataset is provided. According to Rana et al. (2022), previous deep fake detection models can be grouped into three categories: Machine learning, deep learning, and statistical.

Their research indicated that 77% of the work from 2018 to 2020 falls under the Deep Learning category, with 78% being CNN-based. Ultimately, it was demonstrated that deep-learning-based deep fake detection models outperform non-deep learning models (Rana et al., 2022).

In a study conducted by Wang et al. (2024), it was demonstrated that while Convolutional Neural Networks (CNNs) are commonly used in deepfake detection to capture spatial relationships within images, making them effective in identifying facial manipulations and other visual irregularities at the frame level, Vision Transformers (ViTs) have distinct advantages in analyzing and comprehending the intricate details of deepfake images and videos. ViTs are especially good at understanding the overall structure of an image to identify inconsistencies or anomalies suggestive of manipulation.

However, Wang also highlighted the challenges faced by standalone ViT models in deepfake detection, such as their struggle to generalize across diverse datasets, their need for extensive training data, their difficulty in maintaining temporal consistency in video deepfakes, their limited ability to capture local spatial information, and their potential inability to fully capture the temporal and sequential dependencies present in video data. To address these limitations, hybrid models that combine ViTs with other techniques, such as CNNs or RNNs, are often utilized.

2.2 Wavelet Related Work

According to a study conducted by Nadler et al., it was observed that while Deep Neural Networks (DNNs) are increasingly employed as models of human vision, they may disregard essential perceptual attributes, such as color, in favor of constructing high-dimensional abstractions. In contrast, human visual cognition retains and integrates color information with other representations. The research also indicated that a wavelet algorithm based on wavelet decomposition produced more consistent color embeddings, which yielded better alignment with human color judgments compared to all the DNNs included in the study. In their study, Walter et al. used wavelet-packet representation to analyze GAN-generated images from datasets such as CelebA and LSUN. The research revealed differences in the spatial frequency properties of real images compared to GAN-generated images. Specifically, the mean ln-db4-wavelet packet plots and mean Haar-wavelet packet representation showed variations in the frequency content of the images. GAN-generated images exhibited differences in mean and standard deviation of wavelet packets, especially at higher frequencies and edges of the images. These differences suggest distinct spatial frequency properties of GAN-generated images compared to real images, which can be used for their detection and differentiation from real images.

3 Methods

3.1 Multiresolution Analysis

The concept of multiresolution analysis (MRA) was developed to address the limitations of the Fourier transform when dealing with non-stationary signals. These limitations arise from the principle of uncertainty in the time and frequency domains. Specifically, seeking high resolution in the time domain leads to poor resolution in the frequency domain, and vice versa. MRA addresses this issue by incorporating increasing levels of time samples as we transition from the slower part of the signal to the faster part (Erick Axel, et al.). In order to achieve this, MRA (Multiresolution Analysis) utilizes a group of orthonormal basis functions to estimate the signals at a specific resolution.

If we have a discrete signal $f_m(t)$, with a resolution m then we can decompose this signal to two functions $f_{m-1}(t)$ and $e_{m-1}(t)$ at the lower resolution $m-1$, as follows:

$$f_m(t) = \sum_{n=0}^{2^m N-1} c_{m,n}\phi_{m,n}(t) \tag{1}$$

then:

$$f_m(t) = f_{m-1}(t) + e_{m-1}(t) \tag{2}$$

While:

$$f_{m-1}(t) = \sum_{n=0}^{2^{m-1} N-1} c_{m-1,n}\phi_{m-1,n}(t) \tag{3}$$

$$e_{m-1}(t) = \sum_{n=0}^{2^{m-1} N-1} \omega_{m-1,n}\psi_{m-1,n}(t) \tag{4}$$

In general:
Let V_m be the set of functions that can be expressed in terms of the basis $\phi_{m,n}$,and W_m be the set of functions that can be expressed in terms of the basis $\psi_{m,n}$

equation 2 can be written in more general form using the orthogonal sum as follows:

$$V_m = V_{m-1} \oplus W_{m-1} \tag{5}$$

by expanding equation 5 we drive the following formula:

$$V_m = V_0 \oplus W_0 \oplus W_1 \oplus ... \oplus W_{m-2} \oplus W_{m-1} \tag{6}$$

The last equation represents the m-level Discrete Wavelet Transform (DWT) which denotes the change of basis from ϕ_m to the following wavelet coefficients:

$$\phi_0 \oplus \psi_0 \oplus \psi_1 \oplus ... \oplus \psi_{m-2} \oplus \psi_{m-1} \tag{7}$$

ϕ is wellknown as "the scaling basis" and ψ is wellknown as "the wavelet basis".

4 Experiment Results

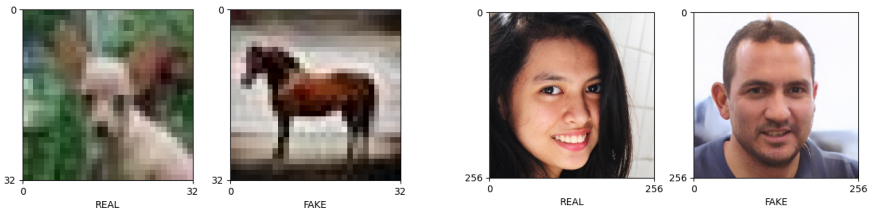
4.1 Used Datasets

The datasets currently used to train and test deepfake detection models, such as FaceForensic++, Celeb-DDF, and DFDC, were developed before the latest diffusion-based tools like DALE-2 and Midjourney came into existence. This raises concerns about the effectiveness of models trained on traditional datasets to detect newly created deepfake images.

Another important aspect of this study is the limitation of computational resources, which affects the size of the required datasets. Traditional datasets are larger than what is currently feasible. CIFAKE, which is an ideal deepfake dataset based on CIFAR-10, addresses these concerns. Firstly, it is created using the state-of-the-art stable-diffusion 1.3 algorithm. Secondly, it consists of a moderate total of 120000 images, with half being real and half being fake.

Additionally, a StyleGAN3-based dataset called "140k Real and Fake Faces" was used to cover a broader range of deepfake styles. Similar to CIFAKE, this dataset also has a moderate size of 140000 images, equally divided into real and fake.

In the model code, an option was provided to split the input dataset into training, validation, and testing according to desired split ratios. However, I chose to keep the original split as it is for a fair comparison with other models that are also using the same datasets.



(a) CIFAKE (b) 140k Real and Fake Faces

Fig. 1: The elected deepfake datasets for this study

4.2 Key Packages

The model is built using Python 3.11 , and Pytorch version 2.2.1 with cuda 12.1 The WPT stage in the WPT-ViT model is not only inspired by the work of Moritz et al. (2024), but also uses their open wavelet library called PTWT for image decomposition instead of the older PYWT developed by Lee et al. (2019). This choice was made after trying both, because of the significantly faster speed of PTWT, particularly due to its support for parallel processing using GPU (Wolter et al., 2024).

4.3 Experiments

Table 1 summarizes the trials that was made to train and evaluate our model, we will visit some of them in more details

Experiment	Dataset	wavelet fun	wavelet level	Paches per decompo-sition	Heads	encoder levels	Sliced?	Used Slices	Batch Size	Image Dimen-sion
1	CIFAKE	db2	3	1	18	1	0		500	32
2	RVSF	db2	3	1	18	1	0		500	32
3	CIFAKE	db2	3	1	18	1	0		1000	32
4	CIFAKE	db2	3	1	18	1	0		2000	32
5	CIFAKE	db2	3	1	18	1	1	[aah ,aha ,vaa ,vav ,dva ,dvv ,aaa ,aav ,aha]	1000	32
6	CIFAKE	db2	3	1	18	2	0		500	32
7	CIFAKE	haar	3	1	16	2	0		1000	32
8	CIFAKE	haar	3	1	16	1	1	[aah ,aha ,vaa ,vav ,dva ,dvv ,aaa ,aav ,aha]	1000	32
9	CIFAKE	db2	3	1	18	2	0		1000	32
10	RVSF	haar	3	1	16	2	0		1000	32

Table 1: Table Shows

Experiment1 In this experiment CIFAKE dataset is passed to the WPT stage for 3 levels of decompositions using Daubechies(db2) wavelet (figure 2)

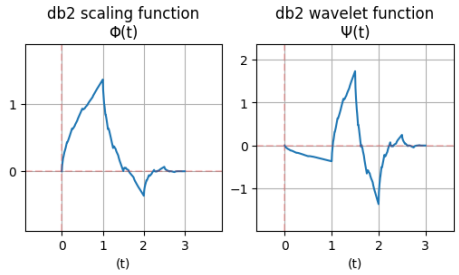
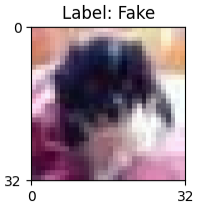
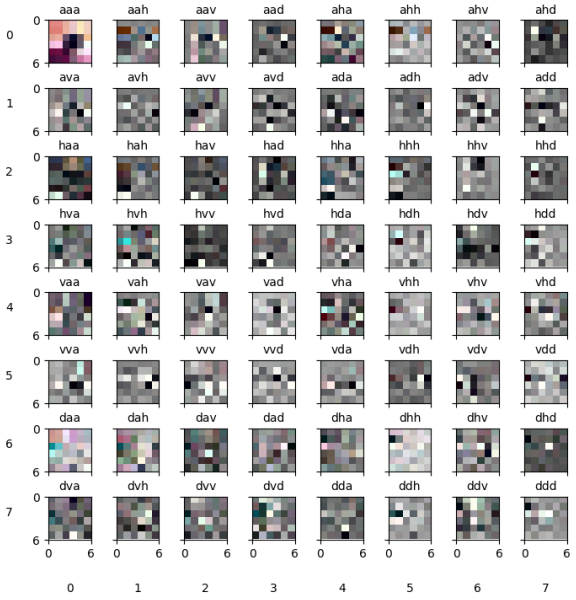


Fig. 2: Daubechies(db2) wavelet

This stage will analyze an input image (that has dimensions 32(H) x 32(W)) to its $4^3 = 64$ coefficients, as shown in figure (figure 3) ,aaa is the most approximate coefficient and ddd is most detailed coefficient which is considered a noise on the image



(a) Sample input



(b) 64 decompositions of the sample image

Fig. 3: Output of the WPT stage

In this experiment, it has been decided to retain all of the coefficients without discarding any. Subsequently, all coefficients will be forwarded to the patchify stage, where further determination will be made on whether to slice each decomposition vertically and horizontally as per the recommendation of the vanilla VIT model. Rather than slicing, the entire Wavelet Packet Transform (WPT) will be treated as a single patch (token) for the VIT stage. This VIT stage comprises a

single encoder with 18 parallel heads. Additionally, the dataset will be supplied to the model in batches of 500 samples. Stochastic gradient descent has been chosen as the optimizer with an initial learning rate of 0.1 and a momentum of 0.9. The model summary for this experiment is presented in the following diagram 5.

```
==========================================================================================================================
Layer (type (var_name))                            Input Shape          Output Shape         Param #              Trainable
==========================================================================================================================
vit (vit)                                          [500, 3, 32, 32]     [500, 2]             7,128                True
├─WPT2D (to_wpt2d)                                 [500, 3, 32, 32]     [500, 64, 3, 6, 6]   --                   --
├─Patch_Embed (to_patch_embedding)                 [500, 64, 3, 6, 6]   [500, 64, 108]       --                   True
│    └─Sequential (patch_embed)                    [500, 64, 3, 6, 6]   [500, 64, 108]       --                   True
│    │    └─Rearrange (0)                          [500, 64, 3, 6, 6]   [500, 64, 108]       --                   --
│    │    └─Linear (1)                             [500, 64, 108]       [500, 64, 108]       11,772               True
├─Dropout (dropout)                                [500, 65, 108]       [500, 65, 108]       --                   --
├─Transformer (transformer)                        [500, 65, 108]       [500, 65, 108]       --                   True
│    └─ModuleList (layers)                         --                   --                   --                   True
│    │    └─ModuleList (0)                         --                   --                   188,352              True
├─Identity (to_latent)                             [500, 108]           [500, 108]           --                   --
├─Sequential (mlp_head)                            [500, 108]           [500, 2]             --                   True
│    └─LayerNorm (0)                               [500, 108]           [500, 108]           216                  True
│    └─Linear (1)                                  [500, 108]           [500, 2]             218                  True
==========================================================================================================================
Total params: 207,686
Trainable params: 207,686
Non-trainable params: 0
Total mult-adds (Units.MEGABYTES): 100.28
==========================================================================================================================
Input size (MB): 6.14
Forward/backward pass size (MB): 421.21
Params size (MB): 0.80
Estimated Total Size (MB): 428.15
==========================================================================================================================
```

Fig. 4: WPT-ViT Model for Experiment 1

The graph shows the training progress over 90 iterations, reaching a peak validation accuracy of 86.56%.

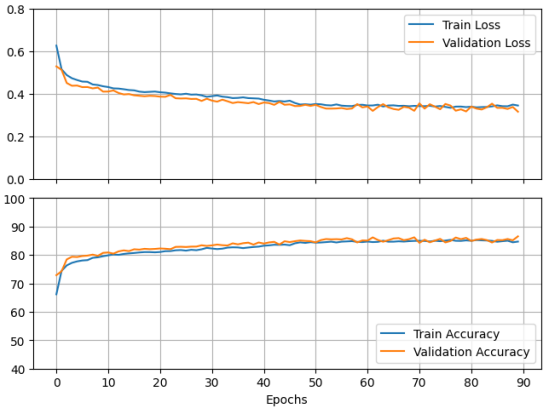


Fig. 5: Training Graph of Exp. 1

Experiment3 In this experiment CIFAKE dataset is passed to the WPT stage for 3 levels of decompositions using Daubechies(db2) wavelet (figure 2)

5 Conclusion

References

1. Bengesi, Staphord, et al. "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers." IEEE Access (2024).
2. Raut, Gaurav, and Apoorv Singh. "Generative AI in Vision: A Survey on Models, Metrics, and Applications." arXiv preprint arXiv:2402.16369 (2024).
3. Pei, Gan, Jiangning, Zhang, Menghan, Hu, Guangtao, Zhai, Chengjie, Wang, Zhenyu, Zhang, Jian, Yang, Chunhua, Shen, Dacheng, Tao. "Deepfake Generation and Detection: A Benchmark and Survey". arXiv preprint arXiv:2403.17881. (2024).
4. Gong, Liang Yu, Xue Jun, Li. "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges". Electronics 13. 3(2024): 585.
5. Heidari, Arash, Nima, Jafari Navimipour, Hasan, Dag, Mehmet, Unal. "Deepfake detection using deep learning methods: A systematic and comprehensive review". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14. 2(2024): e1520.
6. Patel, Yogesh, Sudeep, Tanwar, Rajesh, Gupta, Pronaya, Bhattacharya, Innocent Ewean, Davidson, Royi, Nyameko, Srinivas, Aluvala, Vrince, Vimal. "Deepfake Generation and Detection: Case Study and Challenges". IEEE Access. (2023).