Deepfake Detection using Parallel Vision Transformers

1st B Chetan Kumar
Department of Computer Science and Engineering
PES University
Bengaluru, India
pes1pg22cs007@pesu.pes.edu

2nd Shylaja S S
Department of Computer Science and Engineering
PES University
Bengaluru, India

***Abstract*—Deepfake technology, which uses artificial intelligence to create highly realistic synthetic media, poses significant threats to privacy, security, and the spread of misinformation. Traditional deepfake detection methods, primarily based on convolutional neural networks (CNNs), often fall short in effectively identifying these sophisticated forgeries. This project explores the use of Parallel Vision Transformers (PViTs) for deepfake detection, leveraging their advanced capabilities in modeling complex patterns and long-range dependencies in visual data understanding. We trained the PViT model on a dataset comprising of 140k real and fake faces using Google Colab with an Nvidia A100 GPU. Our results demonstrate that PViTs significantly enhance detection accuracy, precision, recall, and robustness, offering a promising solution for combating the challenges posed by deepfake technology attaining 91.92 accuracy.**

***Index Terms*—Deepfake Detection, Parallel Vision Transformer, AI-generated, Fake Content Identification, Transformers, Cybersecurity, Digital Forensics, Machine Learning, Deep Learning.**

I. Introduction

Deepfake technology refers to the use of deep learning methods to create highly realistic synthetic media, particularly images and videos that are nearly indistinguishable from real content. These deepfakes are generated through advanced multimedia manipulation techniques and artificial intelligence, making them a significant challenge to detect and mitigate. The rise of deepfake technology has been fueled by the development of generative models, such as Generative Adversarial Networks (GANs), which are capable of producing fake content that is nearly identical to real media.

The evolution of deepfakes has been marked by the development of sophisticated tools like Face2Face and FaceSwap, which focus on swapping facial regions and manipulating specific features to achieve realistic results. Among the most notable advancements is the open-source software "DeepFaceLab," which has become the state-of-the-art tool for creating deepfakes with high accuracy. Deepfakes are not limited to visual media; they also extend to audio deepfakes, where voice cloning technologies replicate a person's voice with high fidelity, further complicating the detection process.

As deepfake creation technology advances, it becomes increasingly difficult to distinguish between real and fake content. This necessitates the development of robust countermeasures, primarily through deep learning methods. Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Recurrent Neural Networks (RNNs) have been the traditional approaches used in deepfake detection. These methods involve feature extraction and pattern recognition to classify content as either real or fake.

However, recent advancements in deep learning have introduced Vision Transformers (ViTs) as a more powerful alternative for deepfake detection. Unlike traditional methods, Vision Transformers excel in processing and understanding large datasets, enabling them to capture complex patterns and long-range dependencies within images. This makes them particularly well-suited for the task of deepfake detection, where subtle inconsistencies in synthetic content need to be identified.

The role of Vision Transformers in deepfake detection is crucial as they provide a more sophisticated approach to counteracting the rise of deepfake technology. By leveraging their ability to process large-scale data and their robust learning capabilities, Vision Transformers represent a significant advancement in the fight against synthetic media. As deepfake technology continues to evolve, the importance of developing and refining these detection models cannot be overstated.

This paper explores the application of Parallel Vision Transformers (PViT) for deepfake detection, combining the strengths of transformer models to enhance detection capabilities. The contributions of this paper are:

- Introduction of the Parallel Vision Transformer (PViT) model for deepfake detection.
- Experimental results demonstrating the effectiveness of PViT in detecting deepfakes.

II. Literature Survey

A. Vision Transformers

Vaswani et al.'s "Attention is All You Need" introduced the transformer architecture, which has become foundational in natural language processing and computer vision. Vision Transformers (ViTs) leverage this architecture for image classification and other visual tasks, capturing long-range dependencies through self-attention mechanisms [1]. The success of ViTs in tasks like image recognition and segmentation has

led researchers to explore their potential in more complex applications, such as deepfake detection.

B. Deepfake Detection Techniques

Deepfake detection has evolved from simple statistical methods to advanced neural networks. Early methods focused on detecting visual artifacts and inconsistencies in facial features. More recent approaches utilize convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze temporal and spatial features in videos. Naeem et al. explored various techniques for detecting deepfake content, highlighting the challenges in differentiating real, fake, and synthetic faces using advanced neural networks [2].

C. Parallel Vision Transformers (PViT)

The Parallel Vision Transformer (PViT) extends the ViT architecture by incorporating parallel attention and feedforward blocks, allowing the model to process multiple perspectives simultaneously. This architecture has shown promise in improving accuracy and efficiency in deepfake detection by leveraging the self-attention mechanisms across multiple layers of transformers running in parallel. This approach is particularly beneficial for handling large datasets, which are common in deepfake detection tasks [4].

III. Methodology

A. Data Preprocessing

Images are divided into patches, each resized to 224x224 pixels. These patches are then transformed into feature vectors through a fully connected layer, preparing the data for further processing. Additional preprocessing steps include normalization, augmentation (e.g., random cropping, flipping, rotation), and balancing the dataset to mitigate class imbalance.

B. Parallel Vision Transformer (PViT) Architecture

The PViT model extends the ViT architecture by incorporating parallel attention and feedforward blocks within each layer. This parallel processing enhances the model's ability to learn from multiple perspectives simultaneously, improving its accuracy and efficiency. The architecture of PViT is illustrated in Figure 1.

C. Training and Evaluation

The PViT model was trained and evaluated on the 140k Real and Fake Faces dataset from Kaggle, which includes 70,000 real faces from Flicker face and 70,000 GAN-generated faces StyleGAN from NVidia. The dataset was split into training, validation, and test sets.

The model was trained using Google Colab equipped with an Nvidia A100 GPU, with a batch size of 64, a learning rate of 0.0001, and a binary cross-entropy loss function. The optimizer used was Adam, and a ReduceLROnPlateau scheduler was employed to adjust the learning rate. The training process involved 50 epochs, and early stopping was used to prevent overfitting.

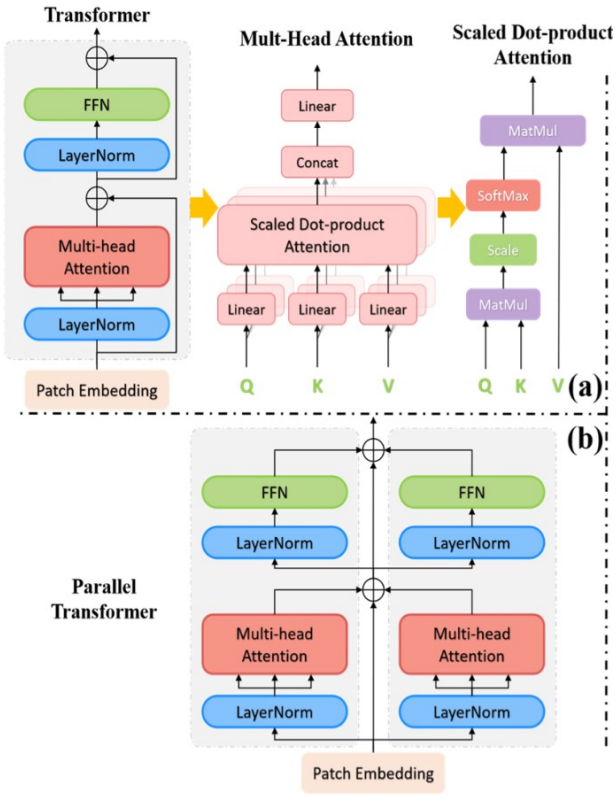

Fig. 1. Proposed Architecture of Parallel Vision Transformer (PViT) for Deepfake Detection.

D. Hyperparameter Tuning

The hyperparameters were optimized through grid search, evaluating combinations of learning rates, batch sizes, and the number of transformer layers. The best performing configuration was found with a learning rate of 0.0001, a batch size of 64, and 6 transformer layers.

IV. Results

The PViT model's performance was evaluated using several metrics, including validation accuracy, test accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). The performance metrics of the PViT model are summarized in Table I.

TABLE I
Performance Metrics of PViT Model

Metric	Value
Validation Accuracy	91.70%
Test Accuracy	91.92%
Precision	95.20%
Recall	88.21%
F1 Score	91.57%
AUC	97.90%

To further illustrate the model's performance, the confusion matrix and ROC curve are shown in Figures 2 and 3.

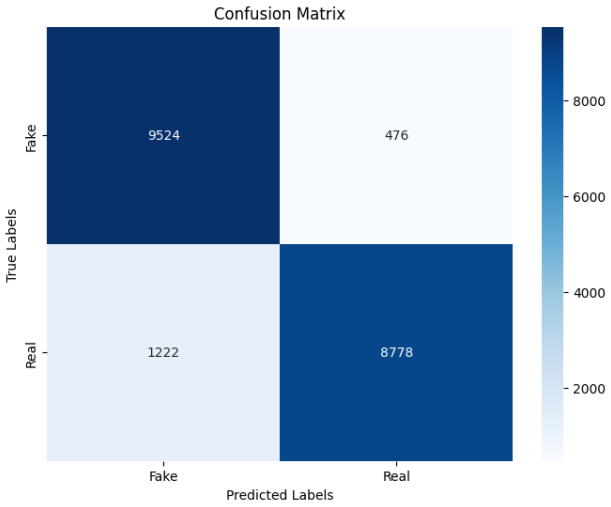

Fig. 2. Confusion Matrix of PViT Model

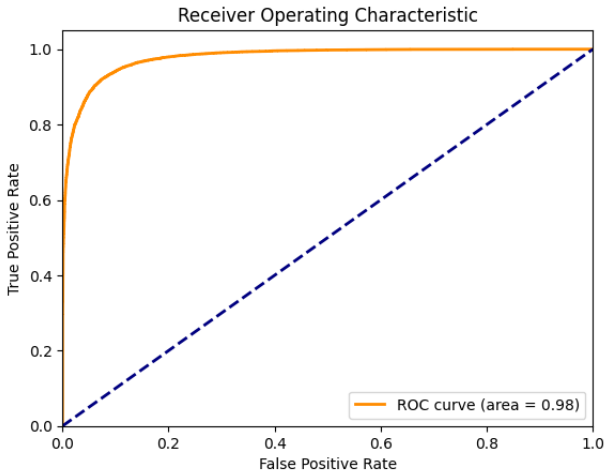

Fig. 3. ROC Curve of PViT Model

A. Comparison with Other Techniques

The performance of the PViT model was compared with existing deepfake detection techniques, including CNNs, RNNs, and traditional ViTs. The results, shown in Table II, indicate that PViT outperforms these models in terms of accuracy and AUC, particularly in handling large datasets and complex image patterns.

TABLE II
COMPARISON OF DEEPFAKE DETECTION MODELS

Model	Accuracy	AUC
CNN	88.54%	95.21%
RNN	86.75%	93.50%
ViT	89.91%	96.35%
PViT (Ours)	91.92%	97.90%

V. DISCUSSION

The results indicate that the PViT model outperforms several traditional deepfake detection methods. The high precision and AUC suggest that the model is particularly effective at minimizing false positives and correctly identifying deepfakes. The PViT model leverages the self-attention mechanism of transformers, allowing it to capture long-range dependencies and subtle inconsistencies in images.

A. Limitations and Challenges

Despite the promising results, the PViT model has limitations, such as the need for large datasets and significant computational resources for training. Additionally, the model's performance may degrade when applied to datasets with unseen or novel deepfake techniques. Future research should explore ways to improve the model's generalization capability.

B. Ethical Considerations

While deepfake detection technology is essential for mitigating the risks associated with deepfakes, it also raises ethical concerns. The development and deployment of such technology must be carefully managed to prevent potential misuse, such as the unjust targeting of individuals or the invasion of privacy.

VI. CONCLUSION

The implementation of the Parallel Vision Transformer (PViT) for deepfake detection has shown a significant improvement in accuracy and robustness in identifying synthetic images. The multi-head self-attention mechanism of the Vision Transformer, combined with parallel transformers, has effectively led to a test accuracy of 91.92%, precision of 95.20%, recall of 88.21%, F1 score of 91.57%, and an AUC of 97.90%. These results demonstrate that using multiple transformers in parallel can result in high effectiveness in distinguishing real and fake images. This approach can be applied in cybersecurity and digital forensics by analyzing fake content on the internet and social media. It can also be used in banking and online verification security, helping prevent the malicious use of deepfake technology for impersonation.

A. Future Work

Future work can focus on fine-tuning the model and training it on a wider deepfake dataset for real-world scenarios. The model's ability to handle large batch sizes, trained in less time and with more effectiveness than a single ViT model, makes it suitable for real-time applications. Additionally, exploring cross-modal integration with audio and visual datasets and improving the PViT model through quantum computing simulations could further enhance its capabilities.

Also, self-supervised learning can be use to train model in real time with the raw data and make the model more real time.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.

[2] Naeem, S., Al-Sharawi, R., Khan, M. R., Tariq, U., Dhall, A., & Al-Nashash, H. (2024). Real, fake and synthetic faces – does the coin have three sides? http://arxiv.org/abs/2404.01878.

[3] Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., & Yang, X. (2024). A Timely Survey on Vision Transformer for Deepfake Detection. http://arxiv.org/abs/2405.08463.

[4] Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., & Jégou, H. (n.d.). Three things everyone should know about Vision Transformers.

[5] Ghita, B., Kuzminykh, I., Usama, A., Bakhshi, T., & Marchang, J. (n.d.). Deepfake Image Detection using Vision Transformer Models.

[6] Mallet, J., Pryor, L., & Dave, D. R. (n.d.). Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM.

[7] S. A. Aduwala, M. Arigala, S. Desai, H. J. Quan and M. Eirinaki, "Deepfake Detection using GAN Discriminators," 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, United Kingdom, pp. 69-77, 2021, doi: 10.1109/BigDataService52369.2021.00014.

[8] Bansal, K., Agarwal, S., & Vyas, N. (2023). Deepfake Detection Using CNN and DCGANS to Drop-Out Fake Multimedia Content: A Hybrid Approach. 2023 International Conference on IoT, Communication and Automation Technology, ICICAT 2023. https://doi.org/10.1109/ICICAT57735.2023.10263628.

[9] Ramachandra, A. C., Nihal, K. P., Mishra, K., Sahithi P, M., Prajwal, A. v., & Jagadeesh, H. S. (2023). Hybrid Model for Detecting Deepfake Videos. 2023 International Conference on Data Science and Network Security, ICDSNS 2023. https://doi.org/10.1109/ICDSNS58469.2023.10245432.

[10] Jaiswal, G. (2021). Hybrid Recurrent Deep Learning Model for DeepFake Video Detection. 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCON 2021. https://doi.org/10.1109/UPCON52273.2021.9667632.

[11] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (n.d.). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. https://deepfakedetectionchallenge.ai.

[12] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE International Conference on Computer Vision, 2019-October, 1–11. https://doi.org/10.1109/ICCV.2019.00009.

[13] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. In IEEE Access (Vol. 10, pp. 25494–25513). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3154404.

[14] Falko Matern, Christian Riess, Marc Stamminger (n.d.). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. 2019 IEEE winter applications of computer vision workshop, https://github.com/shaoanlu/faceswap-GAN.

[15] Doshi, A., Venkatadri, A., Kulkarni, S., Athavale, V., Jagarlapudi, A., Suratkar, S., & Kazi, F. (2022). Realtime Deepfake Detection using Video Vision Transformer. IBSSC 2022 - IEEE Bombay Section Signature Conference. https://doi.org/10.1109/IBSSC56953.2022.10037344.

[16] Pipin, S. J., Purba, R., & Pasha, M. F. (2022). Deepfake Video Detection Using Spatiotemporal Convolutional Network and Photo Response Non Uniformity. ICOSNIKOM 2022 - 2022 IEEE International Conference of Computer Science and Information Technology: Boundary Free: Preparing Indonesia for Metaverse Society. https://doi.org/10.1109/ICOSNIKOM56551.2022.10034890.

[17] Ritter, P., Lucian, D., Anderies, & Chowanda, A. (2023). Comparative Analysis and Evaluation of CNN Models for Deepfake Detection. 2023 4th International Conference on Artificial Intelligence and Data Sciences: Discovering Technological Advancement in Artificial Intelligence and Data Science, AiDAS 2023 - Proceedings, 250–255. https://doi.org/10.1109/AiDAS60501.2023.10284611.