Video Deepfake Detection using Hybrid CNN-Transformer Models

Aryan Raj, Prince Singh, Kajal Gupta

Galgotias University, India

aryanrajsirsi@gmail.com, prince200424singh@gmail.com, kajal.gupta@galgotiasuniversity.edu.in

ABSTRACT

Deepfakes on a mass scale have been enabled by advancements in deep generative models. Conventional deepfake methods based on manually designed features or independent CNN models are found to have limitations in generalising across different types of modification methods and levels of compression. In this paper, a combined CNN-Transformer approach for exact deepfakes video detection has been used. In this combined approach, the Transformer component has been used for representing relations among different lifelike videos. On the other hand, for spatial mapping of registered artefacts in deepfake images, a CNN component has been used. This combined approach has been found to perform better at detecting deepfakes than other models when tested on the FaceForensics++, DFDC, and Celeb-DF image datasets. Bias in data consideration in testing and research ethics has been considered in this paper for future research on exact multimedia forensics.

Keywords: Deepfake Detection, Convolutional Neural Networks (CNN), Transformer Models, Video Forensics, Faceforensics++, DFDC, Multimedia Security.

1. Introduction

Owing to their ability to produce highly realistic forgeries, deepfake videos have significantly undermined online knowledge by enabling political manipulation, identity forgery, and reputation attacks in an efficient and widespread manner through social platform sharing. Given today's temporally consistent deepfakes, traditional methods that rely on detecting anomalies in physiology and visuals are no longer useful. Models based on RNNs/LSTMs solved the problem of short-term temporal dependencies, but they are not very efficient at handling long-term dependencies and are not good at detecting anomalies in visuals, as they still neglect temporal anomalies altogether. Transformers solved problems concerning global modelling by utilising self-attention, but they tend not to be very efficient in modelling spatial aspects and now obviously require large amounts of data, which is a severe disadvantage. To combine spatial and temporal knowledge and thus overcome the aforementioned limitations, hybrid architectures based on CNNs and Transformers are used to achieve greater efficiency and higher accuracy than their counterparts in deepfake detection.

2. Research Methodology

1. General Architecture:

There are three principal parts in the proposed hybrid model:

(i) Video preprocessing and frame extraction;

(ii) CNN-based spatial feature extraction;

(iii)Transformer-based temporal modelling and classification

Proceedings of ICICCS-2026 (Poster Papers), DMPedia Lecture Notes in Computer Science & Engineering.
ISBN: 978-81-993813-1-5

Video Input → Frame Extraction → , Face Detection & Alignment → CNN Feature Extraction → Transformer Encoder → Fully Connected Layer → Deepfake / Real Classification

2. Preprocessing Videos: For ensuring temporal consistency, the videos are sampled at a fixed rate of frames per second. To focus the network on relevant regions of the face, face identification and alignment are performed using a face detector trained on the face dataset. Before being passed on to the CNN layer, the frames are reduced and normalized.

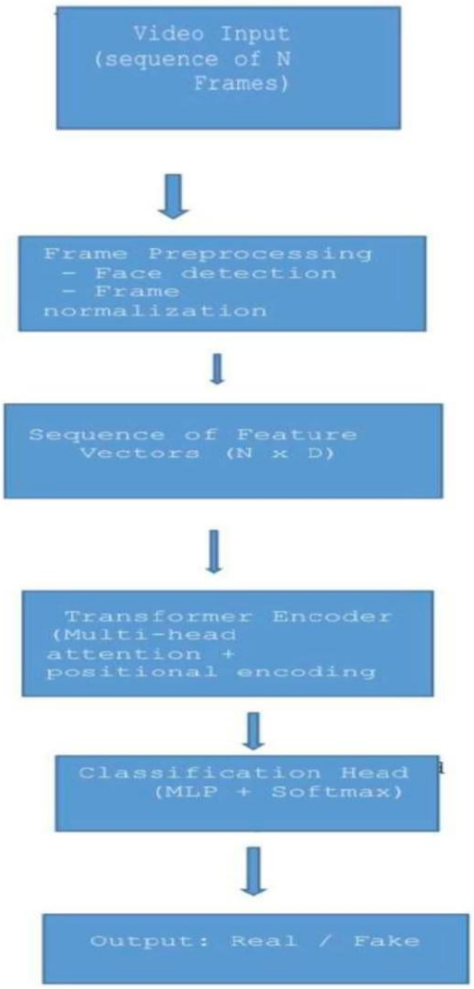
3. Spatial Feature Extraction Using CNN: High-level spatial information is extracted from each frame of the video using a deep CNN backbone like ResNet or EfficientNet. Local artefacts, such as boundary blurring, texture inconsistencies, and colour inconsistencies due to tampering, are captured by the CNN. Vectors of features corresponding to different frames are the CNN's output.

4. Temporal Modelling Using Transformer: The Transformer encoder is fed the CNN-extracted features. To provide temporal consistency, positional encoding is introduced. The model can identify long-range dependencies and temporal irregularities between frames, often indicative of a deepfake movie, using multi-head self-attention.

5. Layer of classification: For identifying whether a video is real or fake, a fully connected layer with a softmax activation function is applied on the summed Transformer output.

Experimental Setup: as per Figure 1

Figure 1. Model Pipeline (ASCI Flowchart):



3. Theory and Calculation

1. Quantitative Evaluation: A comparison of the proposed hybrid CNN- Transformers method with individual CNN and Transformers was made using FaceForensics++ (FF++) and DFC benchmarked datasets. The results highlight the effectiveness of combining the learning of temporal and spatial features in the task of deepfake detection, as illustrated in Table 1.

Table 1. Comparing Deepfake Detection and Model Performance

Model	**Accuracy**	**AUC**
CNN Only	85.3	0.91
Transformer - Only	81.7	0.88
Hybrid CNN- Transformer	92.5	0.95

4.Mathematical Expressions and Symbols

1.Data Preparation:

The model is tested and validated using popular benchmark datasets available for deepfake detection:

• FaceForensics++ (FF++): It consists of videos manipulated using a range of face editing algorithms, including DeepFakes, FaceFace, FaceSwap, and NeuralTextures.

• Deepfake Detection Challenge Dataset (DFDC):

This dataset is intended to help create effective models as it comprises a vast number of deepfake videos with different subjects, environments, and levels of compression.

2. The Loss Function:

The loss function that is common in binary classification tasks and also used for training the network is the Binary Cross-Entropy loss function:

[

L = -Ly \log(p) + (1)

]

where (y \in {0,1}) represents the label, and (p\in[0,1]) denotes the probability or likelihood that a frame or video is a deepfake.

5.Results and Discussion

The experimental evaluation provides several key findings on the role of temporal and spatial feature learning in video deepfake detection as follows:

CNN: Is able to successfully detect fine details of spatial tampering artefacts, such as blurring problems in a frame or pixel-level anomalies.

Transformer: Identifies discrepancies that an analysis per frame cannot capture concerning facial movement and continuity of expressions and transitions from frame to frame.

Fusion: With both CNN and Transformer features, the detection precision and AUC improve dramatically, as they are more robust towards different deepfake modification techniques.

6.Discussion:

• Joint Spatial: Temporal Analysis: This enables more precise identification by capturing both the discrepancies over time in video sequences and the spatial artefacts on individual images.

• Improved Detection Precision: This approach effectively distinguishes both local and global deepfake signals by merging CNN features and a Transformer architecture.

• Resistance to Varying Deepfake Tactics: Excellent generalisation capabilities against varying levels of manipulation tactics, such as facial expressions, reenactment, and face swap techniques.

• Effective Long Range Dependency Modelling: This develops more accurate methods for spotting very fine-scale movement and expression irregularities by applying transformers for modelling long-range dependencies.

7.Conclusions

Through the integration of the capabilities of CNNs to detect minute spatial cues with the ability of Transformers to learn about discrepancies in the temporal context over a prolonged period, the video deepfake detection capabilities are boosted. Despite the benefits brought about by the use of the combined capabilities of CNNs and Transformers, further research is needed to include other aspects that may boost their capabilities, such as authenticity in the audio segment, as well as other aspects, such as eye movements and heart rate changes reflected in face colours.

8.Acknowledgements

We would like to express our sincere gratitude to our faculty members, mentors, and institution for their guidance, support, and encouragement throughout this research work. We also acknowledge the availability of public datasets, open-source software libraries, and research resources that contributed to the successful completion of this study. Their assistance and valuable insights were instrumental in conducting this research on deepfake detection using Hybrid CNN–Transformer models.

9.Funding source

No funding was received for this study.

10.Conflict of Interest

The authors declare no conflict of interest.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza,B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, 2014, pp. 2672-2680.

[2] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," IEEE Access, vol. 6, pp. 67745-67759, 2018.

[3] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing Al-Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018, pp. 1-7.

[4] S. Agarwal, H. Farid, Y. Gu, M. He, K Nagano, and H. Li, "Protecting World Leaders Against Deep Fakes," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 38-45.

[5] G. Shrivastava, R.P. Ojha, S. Awasthi, H. Bansal, K. Sharma, eds., "Emerging threats and countermeasures in cybersecurity, "John Wiley & Sons, 2024.

[6] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artefacts, "IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3085-3097, 2020.

[7] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf.Learning Representations (ICLR), 2021.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A Video T. Mittal, U. Bhattacharya, R. Chandra, and R. R. Ramakrishnan, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues," in Proc. ACM Int. Conf. Multimedia (MM), Seattle, WA, USA, 2020, pp. 2823-2832.Vision Transformer," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, 2021, pp. 6836-6846.

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. "FaceForensics++: isthis, Leaning and M. Nießner, to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1-11.

[10] J. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[11] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020.

[12] E. Sabir, J. Cheng, A. Jaiswal, W.AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for FaceManipulation Detection in Videos," in Proc. IEEE Int. Conf. Image Processing Theory, Tools and Applications (IPTA), Paris, France, 2019, pp. 1-6.

[13] Y. Wang, O. B. Wright, and A. Shrivastava, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 8695-8704.

[14] P. Zhou, X. Han, V. I. Morariu, and L. S. "Two-Stream Neural Networks for Con Compace Detoni on Paleo erosio. Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 1831-1839.