Computing and Artificial Intelligence 2024, 2(2), 1409.
https://doi.org/10.59400/cai.v2i2.1409

Article

Predicting manipulated regions in deepfake videos using convolutional vision transformers

Mohan Bhandari[1,*], Sushant Shrestha[2], Utsab Karki[2], Santosh Adhikari[2], Rajan Gaihre[2]

[1] Department of Science and Technology, Samriddhi College, Lokanthali, Bhaktapur 44800, Nepal
[2] Department of Computer Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur 44700, Nepal
*** Corresponding author:** Mohan Bhandari, mail2mohanbhandari@gmail.com

CITATION

Bhandari M, Shrestha S, Karki U, et al. Predicting manipulated regions in deepfake videos using convolutional vision transformers. Computing and Artificial Intelligence. 2024; 2(2): 1409.
https://doi.org/10.59400/cai.v2i2.1409

ARTICLE INFO

Received: 30 May 2024
Accepted: 13 June 2024
Available online: 19 July 2024

COPYRIGHT

Copyright © 2024 by author(s). *Computing and Artificial Intelligence* is published by Academic Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.
https://creativecommons.org/licenses/by/4.0/

Abstract: Deepfake technology, which uses artificial intelligence to create and manipulate realistic synthetic media, poses a serious threat to the trustworthiness and integrity of digital content. Deepfakes can be used to generate, swap, or modify faces in videos, altering the appearance, identity, or expression of individuals. This study presents an approach for deepfake detection, based on a convolutional vision transformer (CViT), a hybrid model that combines convolutional neural networks (CNNs) and vision transformers (ViTs). The proposed study uses a 20-layer CNN to extract learnable features from face images, and a ViT to classify them into real or fake categories. The study also employs MTCNN, a multi-task cascaded network, to detect and align faces in videos, improving the accuracy and efficiency of the face extraction process. The method is assessed using the FaceForensics++ dataset, which comprises 15,800 images sourced from 1600 videos. With an 80:10:10 split ratio, the experimental results show that the proposed method achieves an accuracy of 92.5% and an AUC of 0.91. We use Gradient-Weighted Class Activation Mapping (Grad-CAM) visualization that highlights distinctive image regions used for making a decision. The proposed method demonstrates a high capability of detecting and distinguishing between genuine and manipulated videos, contributing to the enhancement of media authenticity and security.

Keywords: face detection; machine learning; vision transformer; convolution neural networks; Grad-CAM

1. Introduction

Technologies for altering images and videos are developing rapidly. The rise of fake technology has gained significant attention in recent years due to its ability to generate highly realistic, manipulated media. The different techniques and technical expertise needed to create and manipulate digital content are also easily accessible, as there is abundant reading material on the internet [1]. Currently, it is possible to seamlessly generate hyper-realistic digital images with a few resources and easy-to-follow instructions available online [2]. Deepfake is a technique that aims to replace the face of a targeted person with the face of someone else in a video. It is created by splicing the synthesized face region into the original image. The term can also mean to represent the final output of a hyper-realistic video created. Deepfakes can be used for the creation of hyper-realistic Computer-generated imagery (CGI), Virtual Reality (VR), Augmented Reality (AR), Education, Animation, Arts, and Cinema. However, since Deepfakes are deceptive, they can also be used for malicious purposes [3]. Deepfake detection is the task of identifying and exposing digital falsifications of images, video, and audio that are created with machine learning

techniques [4]. This task poses a formidable challenge to privacy, democracy, and national security, as deepfakes can be used to manipulate public opinion, deceive voters, undermine trust in institutions, exacerbate social divisions, endanger public safety, disrupt international relations, and jeopardize national security. Detecting deepfakes is not only technically difficult but also socially and legally complex. Technical solutions, such as forensic analysis, digital watermarking, and immutable authentication trails, face limitations in accuracy, scalability, and usability [5]. Social and legal solutions, such as media literacy, platform regulation, and legal liability face trade-offs between free expression, privacy, and accountability. Moreover, deepfake creators can adapt to detection methods and exploit cognitive biases that make people susceptible to believing and spreading false information. Therefore, deepfake detection requires a multidisciplinary and collaborative approach that balances the benefits and harms of deepfake technology [6].

The challenge of deepfake detection is the diversity and complexity of deepfake generation methods. There are various types of deepfake techniques, such as face swapping, face reenactment, lip-syncing, voice cloning, and text generation [7]. Each of these techniques requires different approaches and models to create and manipulate digital content. Moreover, the quality and realism of fake media vary depending on the data, algorithms, and parameters used for the generation process. Therefore, it is difficult to design a universal and effective deep fake detector that can handle all kinds of deep fake scenarios.

In this study, we propose to leverage the power of convolutional vision transformer (CViT) to develop a comprehensive and robust deepfake detection framework that can adapt to different types of deepfake techniques and media. By utilizing the capabilities of CViT and focusing on the inconsistency in pixel-level details, we aim to address the disadvantages of deepfake technology and provide a robust defense against its malicious usage. This study strives to contribute to the development of advanced deepfake detection techniques, enhancing the security and integrity of digital media in an increasingly vulnerable landscape [7].

2. Literature review

In “Deepfakes Detection with Automatic Face Weighting”, Montserrat et al. [8] proposed a novel method utilizing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to detect deepfakes. This approach extracts visual and temporal features from facial regions in videos for effective manipulation identification. The study uses the Deepfake Detection Challenge (DFDC) dataset, comprising over 100,000 videos with various facial modifications. The method employs CNNs and RNNs to detect and localize manipulated faces, showing competitive performance against existing techniques. It can handle videos with multiple faces, varying quality, and different manipulation methods, and provides a confidence score for each face region. The reported accuracy is 92.61% in detecting forgeries. However, the method struggles with highly realistic manipulations in blurry or low-quality images and does not incorporate audio information, which could enhance detection performance. In “MesoNet: A Compact Facial Video Forgery Detection Network,” Afchar et al. [9] present an efficient method for

detecting manipulated faces in videos, focusing on Deepfake and Face2Face techniques. They utilize two datasets: the Deepfake dataset, with 175 forged videos and frames extracted and aligned, and the FaceForensics-based Face2Face dataset, with over a thousand videos. Training and testing sets include 5111 forged and 7250 real images, 2889 forged and 4259 real images, respectively, for Deepfake; and 300 training and 150 testing videos for Face2Face. Traditional image analysis methods fail for videos due to compression issues, prompting the authors to propose two deep learning networks with few layers to analyze key features, achieving over 98% accuracy for Deepfake and 95% for Face2Face.

Wodajo et al. [3] proposed a CViT for detecting Deepfakes, integrating a CNN with a ViT. The CNN extracts learnable features, which the ViT then processes using an attention mechanism for categorization. Trained on the DeepFake Detection Challenge (DFDC) dataset, their model achieves 95.8% accuracy, an AUC of 99.30, and a loss value of 0.32. The key contribution is the integration of a CNN module into the ViT architecture, resulting in competitive performance on the DFDC dataset. This combination leverages the strengths of both CNNs and ViTs, enhancing feature extraction and classification accuracy in Deepfake detection.

Ha et al. [10] introduced a robust DeepFake detection method that combines ViT and CNN models. Experiments showed that the ViT model excels at processing side faces and low-quality videos. The method, which integrates the ResNeSt269 model with the DeiT model using a weighted majority voting ensemble approach, achieved a 97.66% accuracy, surpassing the 96.78% accuracy of the current state-of-the-art model in the DFDC. Additionally, when tested on a completely different dataset, the method demonstrated robustness and over 10% higher accuracy compared to the CNN model, thanks to ViT's high generalization performance.

3. Materials and methodology

Face extraction using MTCNN and data augmentation are performed on the extracted face images. CViT combines CNNs for feature learning and ViTs for deep fake detection. CViT processes standardized face images (224 × 224 RGB), splitting them into patches for analysis. Utilizing the features from CNNs and ViTs, CViT accurately detects deepfake manipulation within face images. The entire process of the study is shown in **Figure 1**.

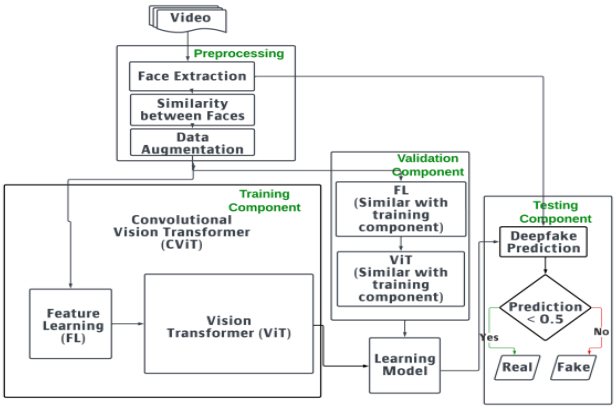

Figure 1. Methodology.

3.1. Dataset

FaceForensics++ [11] has 15,800 images extracted from 1,600 videos. The dataset is divided into training (72.38%, 11,448 images: 5835 fake, 5613 real), testing (19.62%, 3,103 images with a similar fake-real distribution), and validation (7.94%, 1252 images, evenly split between fake and real).

3.2. Preprocessing component

The preprocessing component plays a crucial role in preparing input data for the model. It consists of two key processes: face extraction and data augmentation. The face extraction component identifies and extracts faces from video frames, focusing the analysis on facial features. This step is vital, given that deepfakes often involve manipulations in this particular region. On the other hand, data augmentation enhances the model's ability to generalize by diversifying the training dataset. This involves applying random transformations like rotation, scaling, flipping, and sharpening to face images. To illustrate, the face extraction process outputs images in a standardized 224 × 224 RGB format. Simultaneously, data augmentation creates additional training samples with slightly modified versions of the original data. **Figure 2a** is an example of some of the frames. After obtaining the frames (224 × 224 RGB) as shown in **Figure 2b**, we calculated the facial region which is performed with the help of the MTCNN. After the face region has been obtained further processing and normalization are performed and the **Figure 3** are the images obtained after the normalization.

(a) **(b)**

Figure 2. Frames and detection of face. **(a)** Frames in video; **(b)** Detection of face.

Figure 3. After normalization.

3.3. Multi-task cascaded convolutional neural networks

- The Multi-task Cascaded Convolutional Neural Network

(MTCNN) algorithm used to detect face and face landmarks, works in three steps and uses one neural network for each process. The initial part is a proposal

network that will predict potential face positions and their bounding boxes just like an attention network in Faster R-CNN. The result of this process is a large number of face detection sandlots of false detections. The second part uses images and outputs of the first prediction, thus making a refinement of the result to eliminate most of the false detections and aggregate bounding boxes. The last part refines the predictions and adds facial landmarks predictions in the original MTCNN implementation. Experimental results have always demonstrated that while keeping the reliability of real-time performance, this method consistently outperforms the sophisticated conventional methods across most of the challenging benchmarks. This better performance for real-time is of great importance in a surveillance system [12]. The equations involved in the MTCNN algorithm are shown in Equations (1)–(3).

$$B = sigmoid\ (f_1(x, y, w, h)) \tag{1}$$

where B represents the bounding box coordinates, (x,y) are the coordinates of the top-left corner, (w, h) are the width and height of the bounding box, and f_1 is the neural network function.

$$\Delta B = f_2(B, I) \tag{2}$$

where ΔB represents the refined bounding box coordinates based on the initial bounding box B and the input image I, and f_2 is the neural network function.

$$(P, L) = f_3(B, I) \tag{3}$$

where P represents the facial landmarks and L represents the probability of the face being real, and f_3 is the neural network function.

3.4. Feature selection using CNN

Feature Learning (FL) is important in CNNs, especially for face recognition. It involves using blocks with convolutional layers to extract features from input face data. The features include the two eyes, nose, and the two sides of the mouth. These features are gradually learned and used as building blocks for higher-level analysis in the model. FL transforms raw data into meaningful representations, enabling more advanced processing in the neural network.

3.5. Vision transformer

The Vision Transformer (VT) within the CVIT framework is a key component that adapts transformer architecture, originally developed for natural language processing, to computer vision tasks. It processes learned features from the Feature Learning (FL) stage using self-attention, capturing global context information to understand relationships across different parts of the face. Following this, the MLP head, comprising fully connected layers and activation functions, refines these features for classification, distinguishing between real and fake inputs. The soft max function then assigns class probabilities based on raw scores from the MLP head, aiding in the final classification decision. The transformer encoder, meanwhile, handles linear projections of flattened patches from the Vision Transformer, refining features further by combining local and global information. These refined features are then fed into the MLP head for classification, enhancing the model's predictive accuracy and robustness. Additionally, the validation component assesses the model's performance on unseen data, incorporating FL and VT stages but operating

on a separate validation dataset to ensure an unbiased evaluation of the model's generalization capabilities.

3.6. Grad-CAM

Grad-CAM calculates the gradient of a differentiable output, such as class score, in relation to the convolutional features of a selected layer. Grad-CAM is most commonly employed for image classification tasks, but may also be utilized for semantic segmentation. The soft max layer of the proposed model outputs a score for each class for each pixel to aid in semantic segmentation. For a particular class *C* with *N* number of pixels and A^K as a feature map, Grad-CAM mapping is explained in Equation (4) [13].

$$M^c = ReLU \sum_K \propto_c^K A^K \tag{4}$$

where,

$$\propto_c^K = \frac{1}{N} \sum_{i,j} \left(\frac{d_y{}^c}{dA_{i,j}^K}\right) \tag{5}$$

3.7. Real time implementation

During testing, the user uploads the video, and after face extraction, the extracted features are loaded with our CViT model. Leveraging the validated model, this component aims to predict the authenticity of new content, such as videos or frames, by determining whether they are real or fake. During testing, a predefined threshold of 0.5 is established to serve as the decision boundary. It gives the prediction score which if it is less than 0.5 is a real video otherwise it is a fake video.

4. Experiments and analysis

The experiment is conducted using Python language with Intel(R) Core (TM) i5-13500H CPU, windows 11 operating system, with 8 GB RAM.

Table 1. Results of videos along with prediction score (Samples are from the dataset).

SN	Sample Inputs	Prediction Score	Result
1	Sample 1.mp4	0.051	Real
2	Sample 2.mp4	0.67	Fake
3	Sample 3.mp4	0.03	Real
4	Sample 4.mp4	0.15	Real
5	Sample 5.mp4	0.96	Fake
6	Sample 6.mp4	0.26	Real
7	Sample 7.mp4	0.02	Real
8	Sample 8.mp4	0.9	Fake

The system evaluates the uploaded video and determines whether it's authentic or fake based on the prediction score from the model. If the score exceeds a threshold of 0.5, the video is flagged as fake otherwise, it's considered real, with the capability to predict whether a video is real or fake achieved. The above **Table 1**

comprises the file name, the prediction score through which we can predict whether the video is real or fake, and the prediction.

Throughout the training process, the training accuracy achieved by the model on this dataset was 92.5%. The accuracy graph is shown in **Figure 4** and training loss is shown in **Figure 5**. Additionally, the model's capacity was assessed using a Receiver Operating Characteristic (ROC) curve. The Area Under the Curve (AUC) value, representing the area covered by the ROC curve, was determined to be 0.91.

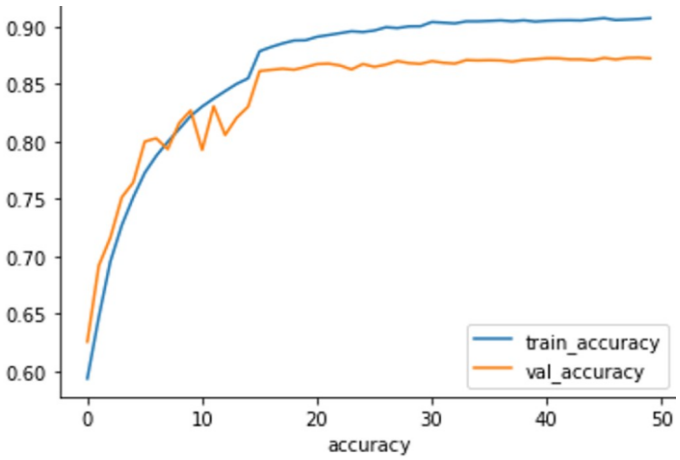

Figure 4. Training Accuracy.

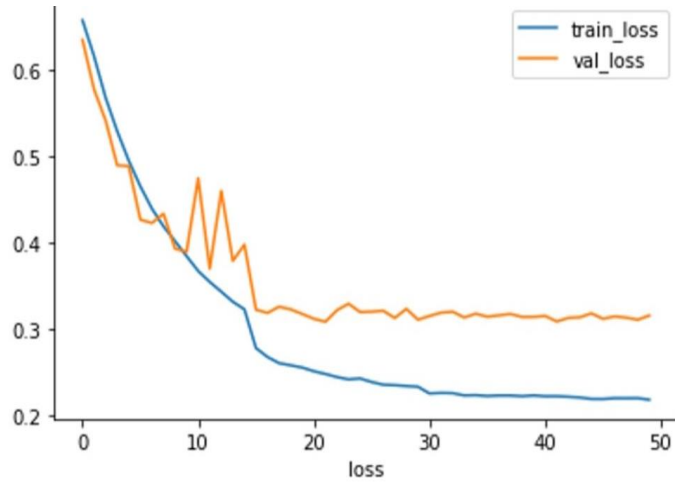

Figure 5. Training Loss.

A higher AUC suggests that the model has a strong ability to distinguish between the positive and negative classes. The ROC curve is shown in **Figure 6**. We conducted a 10-fold cross-validation and obtained an average accuracy of 92.5%, an average AUC of 0.91, an average precision of 0.91, and an average recall of 0.93. The K-fold result is shown in the **Table 2.**

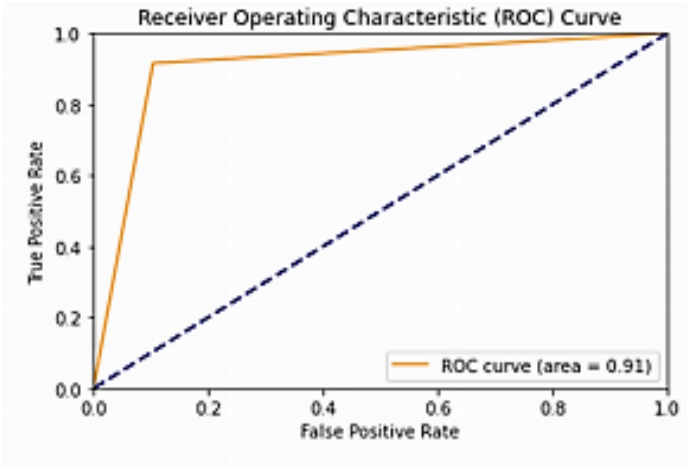

Figure 6. ROC curve.

Table 2. K-Fold cross-validation results.

Fold	Accuracy (%)	AUC	Precision	Recall
1	87.5	0.92	0.86	0.88
2	88.2	0.91	0.87	0.89
3	89.6	0.92	0.88	0.90
4	90.1	0.93	0.89	0.91
5	91.4	0.94	0.90	0.92
6	92.0	0.95	0.91	0.93
7	92.3	0.94	0.92	0.94
8	93.0	0.96	0.93	0.95
9	93.8	0.97	0.94	0.96
10	94.0	0.98	0.95	0.97
Avg	92.5	0.91	0.91	0.93

Grad-CAM is used to understand which parts of the input image are crucial for the deep learning model to determine whether an image or video is real or fake. For this we created a heatmap on the image to visualize the regions of interest, we can gain insights into how the model makes its decisions and potentially identify artifacts or inconsistencies indicative of manipulation. **Figure 7** shows the Grad-CAM over frames for fake video content.

Figure 7. Grad-CAM of fake image.

5. Conclusion

The study focuses on deepfake detection, employing a fusion of MTCNN architecture for feature extraction and Vision Transformer for video classification, which has yielded a noteworthy accuracy of 92.5% on the FaceForensics++ dataset, containing 15,808 images encompassing both genuine and fabricated instances. This outcome underscores the efficacy of our methodology.

To enhance future iterations, enlarging the dataset could bolster the model's capacity for generalization and resilience across diverse scenarios, potentially augmenting accuracy further. Moreover, integrating audio analysis alongside visual data offers a promising avenue for fortifying deepfake detection capabilities. By harnessing both visual and auditory cues, we can develop more comprehensive and dependable detection systems to counteract the escalating threat of media manipulation.

Author contributions: Conceptualization, SS, UK, SA, RG and MB; methodology, SS, UK, SA and RG; software, SS, UK, SA and RG; validation, SS, UK, SA and MB; formal analysis, MB; investigation, SS, UK, SA and RG; resources, SS, UK, SA and RG; data curation, SS, UK, SA and RG; writing—original draft preparation, SS, UK, SA and RG; writing—review and editing, MB; visualization, MB; supervision, MB; project administration, MB. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Karnouskos S. Artificial Intelligence in Digital Media: The Era of Deepfakes. IEEE Transactions on Technology and Society. 2020; 1(3): 138-147. doi: 10.1109/tts.2020.3001312
2. Grobler GD. Narrative strategies in the creation of animated poetry-film [PhD thesis]. University of South Africa; 2021.
3. Wodajo D, Atnafu S, Akhtar Z. Deepfake video detection using generative convolutional vision transformer. Available online: https://arxiv.org/abs/2307.07036 (accessed on 20 May 2024).
4. Heidari A, Jafari Navimipour N, Dag H, et al. Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery. 2023; 14(2). doi: 10.1002/widm.1520
5. Kearns L, Alam A, Allison J. Synthetic media authentication threats: Detection using a combination of neural network and blockchain technology. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4658121 (accessed on 20 May 2024).
6. Chesney R, Citron DK. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. SSRN Electronic Journal. 2018. doi: 10.2139/ssrn.3213954
7. Masood M, Nawaz M, Malik KM, et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence. 2022; 53(4): 3974-4026. doi: 10.1007/s10489-022-03766-z
8. Montserrat DM, Hao H, Yarlagadda SK, et al. Deepfakes Detection with Automatic Face Weighting. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020. doi: 10.1109/cvprw50498.2020.00342
9. Afchar D, Nozick V, Yamagishi J, et al. MesoNet: a Compact Facial Video Forgery Detection Network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018. doi: 10.1109/wifs.2018.8630761
10. Ha H, Kim M, Han S, et al. Robust Deep Fake Detection Method based on Ensemble of ViT and CNN. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing; 2023. doi: 10.1145/3555776.3577769
11. Hasan FS. FaceForensics-1600 videos-preprocess. Available online: https://www.kaggle.com/datasets/farhansharukhhasan/faceforensics1600-videospreprocess?rvi=1 (accessed on 23 May 2024).

12. Jose EMG, Haridas MTP, Supriya MH. Face Recognition based Surveillance System Using FaceNet and MTCNN on Jetson TX2. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). Published online March 2019. doi: 10.1109/icaccs.2019.8728466
13. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy. pp. 618-626. doi: 10.1109/iccv.2017.74