Deep Fake Image Detection using CvT Model

Dr. Swati Wavhal[1,2], Dr. Ashwin I Mehta[3], Vivek Prajapati[4]
[1]*HOD, Department of Computer Science Ismail Yusuf College, Mumbai: 400060, Mumbai, India*
[2]*Principal of Ismail Yusuf College, Mumbai: 400060, Mumbai, India*
[3]*Guide, Department of Computer Science Ismail Yusuf College, Mumbai: 400060, Mumbai, India*
[4]*UG Scholar, Department of Computer Science Ismail Yusuf College, Mumbai: 400060, Mumbai, India*

Abstract: Deep Fake have become a serious challenge to society for trusting digital media as they can contain manipulative content. In this work, we apply Microsoft Convolution Vision Transformer (CvT) model to detect deep fake using the recently introduced dataset DF40 dataset. Unlike traditional CNN-based methods, CvT combines the local feature extraction strength of convolutions with the global reasoning capabilities of transformers, allowing it to capture both fine-grained artifacts and broader semantic inconsistencies. We train the CvT model end-to-end on DF40 and evaluate its performance without relying on additional ensembles or handcrafted features. The proposed approach achieves an accuracy of 86.26%, demonstrating that CvT can serve as a strong baseline for deepfake detection. Our results highlight the potential of transformer-based vision architectures in building scalable, accurate, and adaptable deepfake forensics systems.

INTRODUCTION

A fast-expanding field, synthetic media is media created by technology. Because of this, artificial media may also be referred to as "AI-generated media". Some examples of synthetic media include music composed by AI, text generation, imagery and video, voice synthesis, and fake images. Deepfakes can be defined as a synthetic media and a deep learning approaches to create false photos and videos by overlaid one person's face on top of another person's face in an already existing image or video [1]. These manipulated media can convincingly depict individuals performing actions that they never performed, creating serious risk in area such as political disinformation, identity fraud. Since the threat of deepfake has already been identified, methods for identifying deepfake are necessary [9].

Early deepfake detection methods mostly relied on convolutional neural networks (CNNs), which are good at spotting pixel-level artifacts and subtle inconsistencies in facial features. While these models achieved promising results, they often fall short in capturing long-range spatial relationships and tend to struggle when applied to new datasets or different manipulation techniques. More recently, Vision Transformers (ViTs) have emerged as a strong alternative, using self-attention mechanisms to capture global context across an image. However, ViTs lack the built-in inductive biases of convolutions, which help models focus on relevant local patterns, and they typically come with high computational costs and a need for large amounts of training data to perform well.

To address these limitations, we explore the Convolutional Vision Transformer (CvT) architecture, proposed by Microsoft, for the task of deepfake detection. Convolutional vision Transformer (CvT) employs all the benefits of CNNs: local receptive fields, shared weights, and spatial subsampling, while keeping all the advantages of Transformers: dynamic attention, global context fusion, and better generalization [8] . This hybrid design enables the model to effectively capture both fine-grained artifacts and broader semantic inconsistencies present in manipulated faces.

We evaluate our approach on the DF40 dataset, a highly diverse and large-scale deepfake detection dataset called DF40, which comprises 40 distinct deepfake techniques (10 times larger than FF++) [10]. The proposed CvT-based model is trained end-to-end without the need for handcrafted feature engineering or multi-model ensembles. Experimental results show that our method achieves a peak accuracy of 86.26%, establishing a strong single-model baseline for deepfake detection on DF40.

The main contributions of this work are as follows:

1. We apply and evaluate the Microsoft CvT architecture for deepfake detection, leveraging its hybrid convolution–transformer design to capture both local and global features.
2. We conduct experiments on the DF40 dataset, demonstrating the model's ability to handle diverse and high-quality manipulations.
3. We establish a strong single-model baseline with an accuracy of 86.26%, highlighting CvT's potential for scalable and accurate deepfake forensics.

LITERATURE REVIEW

[1] This paper represents a deep learning approach for detecting deep fake image specially focusing on real and fake faces. The authors use and modify the EfficientNetB0, convolution architecture enhancing it with additional fully dense connected layers. The model is trained on 140K real and fake faces Kaggle dataset which include both GAN style images and NVIDIA Flickr dataset. Three models are tested: a basic transfer learning model, a version added with dense layer for improved performance, a final model with learning rate schedule. The final model achieved state-of-art accuracy gaining accuracy of 99.06 % and 0.0596 error rate. The study demonstrates that modifying EfficientNetB0 and using a learning rate schedule significantly enhances fake image detection performance. The main drawback is its Dataset Specificity.

[2] The paper addresses the growing problem of deepfake face images, which are synthetic images created using advanced AI techniques like Generative Adversarial Networks (GANs) which are difficult to distinguish from the real ones. The authors propose a machine learning-based approach to detect deepfake face images, focusing on a model that uses Support Vector Machine (SVM) classifiers, both with and without Principal Component Analysis (PCA) for feature selection. The process involves several steps: collecting a large dataset of real and fake face images, preprocessing the images (including color space conversion, gamma correction, and edge detection), and then classifying the images using SVM. The study compares the effectiveness of SVM alone versus SVM combined with PCA. Results show that using SVM with PCA achieves a high accuracy of 96.8%, while SVM alone achieves 72.2% accuracy. The findings suggest that PCA significantly improves the detection performance, and the proposed method outperforms several previous approaches. The paper concludes that combining PCA with SVM is a robust solution for detecting manipulated face images. The main drawback is Dataset Specificity.

[3] The paper investigates methods for detecting deepfake face images using deep learning, specifically Convolutional Neural Networks (CNN). The authors use a dataset of 70,000 real and 70,000 fake face images, applying preprocessing steps such as color space conversion, gamma correction, and edge detection. They compare two approaches: CNN with Principal Component Analysis (PCA) and CNN without PCA. The results show that using CNN alone, especially without preprocessing and with more training data, achieves the highest accuracy (up to 98.04%) in detecting fake images. The study concludes that CNNs are highly effective for this task, and that preprocessing and PCA can sometimes reduce detection accuracy. The findings suggest that direct use of raw images with CNNs and large datasets yields the best results for deepfake detection.

[4] The paper presents a novel automated method for detecting and classifying deep fake images using a combination of Error Level Analysis (ELA), deep learning, and machine learning techniques. The proposed framework first preprocesses images with ELA to identify digital manipulations at the pixel level. These processed images are then fed into pre-trained Convolutional Neural Networks (CNNs) such as GoogLeNet, ResNet18, and SqueezeNet for deep feature extraction. The extracted features are classified using Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), with hyper-parameter optimization to improve performance. The method was evaluated on a publicly available dataset and achieved the highest accuracy of 89.5% using ResNet18 features and KNN. The results demonstrate the robustness and efficiency of the approach, making it suitable for real-time deep fake image detection. The paper also compares its results with state-of-the-art methods, showing superior performance, and discusses the potential for future work on video-based datasets and real-life data

[5] The authors propose a novel deepfake predictor (DFP) approach that combines VGG16 (a pre-trained convolutional neural network) with additional CNN

layers to improve the detection of deepfake images. The study uses a publicly available deepfake dataset from Kaggle, containing both real and expertly photoshopped fake face images. The proposed DFP model is compared against several state-of-the-art transfer learning techniques, including Xception, NAS-Net, Mobile Net, and VGG16. Through extensive experiments and performance evaluations, the DFP approach achieves superior results, with 95% precision and 94% accuracy, outperforming the other models. The paper concludes that the hybrid DFP model is efficient, less complex, and highly effective for deepfake detection, offering valuable applications in cybersecurity.

[6] The authors propose a comprehensive approach to deepfake detection using deep learning, specifically by evaluating and comparing several Convolutional Neural Network (CNN) models: InceptionResNetV2, DenseNet201, ResNet152V2, and InceptionV3. A large balanced dataset of 140k images where used for training, validation and testing purpose. The InceptionV3 achieved the highest performance (99.87 % validation and 99.86 % Testing) which was verified by LIME algorithm for XAI. The author used LIME algorithm (Local Interpretable-Agnostic Explanation) to visually highlight which part of an image influenced the model's decision, providing transparency and interpretability. The results demonstrate that the proposed method is highly accurate, robust, and reliable for detecting deepfake images. The integration of XAI not only validates the model's predictions but also makes the system more trustworthy and user-friendly.

[7] This paper addresses the growing challenge of detecting deepfake images, which are artificially generated or manipulated photos that can convincingly replace real faces and have been used to spread misinformation, incite unrest, and harm reputations. The authors propose a novel approach that combines transfer learning and data augmentation techniques with deep learning models—specifically CNN, VGG16, VGG19, and InceptionV3. A large dataset of 190,335 RGB images (real and deepfake) from Kaggle was used, with extensive data augmentation applied to improve model generalization and reduce overfitting. Among the tested models, the fine-tuned VGG16 model achieved the best performance, with 90% accuracy, recall, F1-score, and AUC-ROC, and 91% precision. The study highlights the effectiveness of combining transfer learning and data augmentation for deepfake detection but notes that the lack of comprehensive data is a limitation to be addressed in future work.

[9] This paper presents a deep learning model for detecting deep fakes, which are highly realistic fake videos which are hard to detect from real ones. It consist of 3 steps: Preprocessing (including frame extraction, face detection, alignment, feature cropping), detection (using CNN for eye and nose prediction and combined with ViT for face detection) and prediction (using a majority voting approach to merge result from three models). The model was evaluated on FaceForensic++ and DFDC dataset. The CNN model achieved 97% accuracy while CViT model achieved 85% accuracy. The results demonstrate significant improvements in deepfake detection compared to recent studies, highlighting the effectiveness of combining multiple facial regions (face, eyes, and nose) and advanced deep learning techniques. The study also discusses the advantages, limitations, and potential managerial implications of implementing such detection systems, emphasizing the importance of reliable deepfake detection for digital security and integrity.

3.METHODOLOGY

1.1 Dataset

DF40 dataset provides 40 deepfake approaches with 4 different types: FS, FR, EFS, and FE. For FS and FR, we provide video format data (over 0.1M video clips in total), and image format for EFS and FE (over 1M+ images in total). We also introduce several latest/popular generation techniques (e.g., PixArt-α in 2024) and online software (e.g., HeyGen, DeepFaceLab). Furthermore, some classical and representative methods are also included, e.g., FOMM [10]. Original Data Collection and Deepfake Data Generation. (1) Real Data: We consider two mainstream and popular deepfake datasets: FF++ (c23 version) and CDF as our original data. The rationale behind this selection is that most previous research follows the evaluation protocol of training on FF++ and testing on CDF. By utilizing these datasets for training and evaluation, we can adhere to previous works and verify if their conclusions and methods still hold in the context of our new dataset. We also provide real data from other existing datasets (e.g., CelebA) to

facilitate the unknown domain evaluations. Fake Data: To guarantee the diversity of deepfake approaches in the proposed DF40, we introduce and implement 40 distinct deepfake techniques. [10]
We have curated for the classification of deep fake face images. It is derived from the DF40 dataset.
The dataset contain 32134 images. It is divided into two categories:

a) Fake Images: 16,060 images generated using various deep fake techniques
b) Real Images: 16,060 authenticate face images

Dataset Structure:

a) Train: Contain 25696 images (12848 each of real and fake)
b) Test: Contain 3212 images (1606 each of real and fake)
c) Val: Contain 3212 images (1606 each of real and fake)

Following is the structural image of CvT model [8] :

1.2 Data Augmentation

To enhance model robustness and mitigate overfitting, training images are rescaled in the range [0,1], followed by random horizontal flips, zooms (≤ 20%), rotations ($\pm 18^0$) and translations (≤ 10% along both axes). Augmented were applied using tf.keras preprocessing layers and applied on-the-fly during training.

1.3 CvT Model Architecture

This convolution operation in CvT aims to model local spatial contexts, from low-level edges to higher order semantic primitives, over a multi-stage hierarchy approach, similar to CNNs [8] where each stage consist of convolution Token Embedding followed by several convolution Transformer blocks.

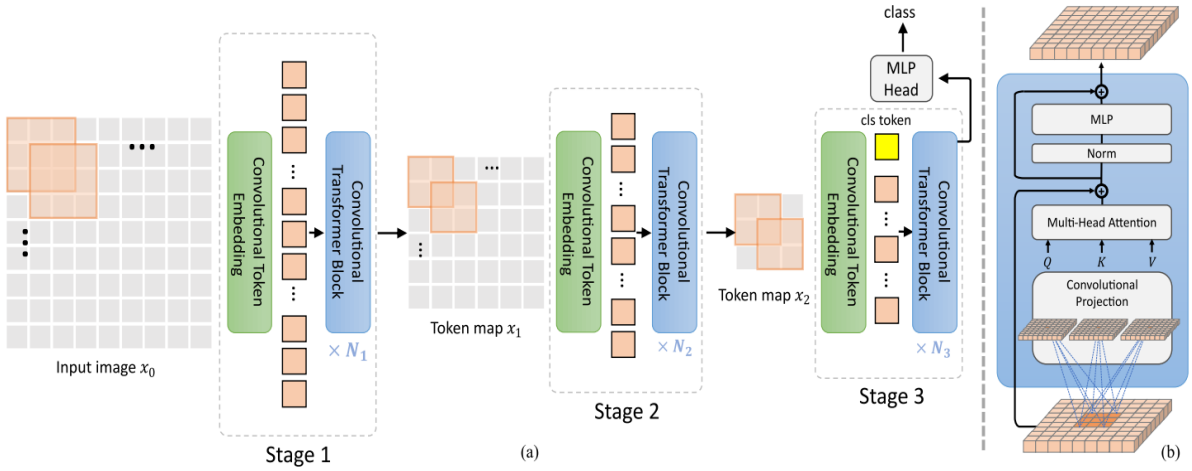

Figure 2: The pipeline of the proposed CvT architecture. (a) Overall architecture, showing the hierarchical multi-stage structure facilitated by the Convolutional Token Embedding layer. (b) Details of the Convolutional Transformer Block, which contains the convolution projection as the first layer.

The model consist of three stages. Each stage has 2 parts.
First, the input image (or 2D reshaped token maps) are subjected to the Convolutional Token Embedding layer, which is implemented as a convolution with overlapping patches with tokens reshaped to the 2D spatial grid as the input (the degree of overlap can be controlled via the stride length).
An additional layer normalization is applied to the tokens. This allows each stage to progressively reduce the number of tokens (i.e. feature resolution) while simultaneously increasing the width of the tokens (i.e. feature dimension), thus achieving spatial downsampling and increased richness of representation, similar to the design of CNNs.
Next, a stack of the proposed Convolutional Transformer Blocks comprise the remainder of each stage. Figure 2 (b) shows the architecture of the Convolutional Transformer Block, where a depth-wise separable convolution operation, referred as Convolutional Projection, is applied for query, key, and value embeddings respectively, instead of the standard position-wise linear projection in ViT. Additionally, the classification token is added only in

the last stage. Finally, an MLP (i.e. fully connected) Head is utilized upon the classification token of the final stage output to predict the class [8].

Following is the in general structure of CvT Model [8] :

Stages and Layers

- Stage 1:
  - Convolutional Token Embedding: 7×7 kernel, stride 4
  - Convolutional Projection (Conv. Proj.): 3×3 kernel,
  - Multi-Head Self-Attention (MHSA): R(feature dimension expansion ratio) = 4
  - MLP (Feed-forward layer): R = 4
- Stage 2:
  - Convolutional Token Embedding: 3×3 kernel, stride 2
  - Convolutional Projection: 3×3 kernel
  - MHSA: R = 4
  - MLP: R = 4
- Stage 3:
  - Convolutional Token Embedding: 3×3 kernel, stride 2
  - Convolutional Projection: 3×3 kernel
  - MHSA: R = 4
  - MLP : R = 4
- Head:
  - 1×1 Linear layer for classification

Our implementation is based on CvT architecture, with structural modification involving changes in no. of transformer blocks per stage and adjustment of architectural parameters such as embedding dimension, no. of attention heads. These modifications were made to better align the capacity of model with DF40 dataset characteristic, aiming to balance computational efficiency.

Following is the our table representation of layers:

Table A

Stage	Operation	Parameters	Output Shape
Input	–	–	224×224×3
Stage 1	Conv Embedding	3×3, stride=2, filters=64	112×112×64
	Transformer ×1	heads=1, MLP ratio=4	112×112×64
Stage 2	Conv Embedding	3×3, stride=2, filters=128	56×56×128
	Transformer ×2	heads=2, MLP ratio=4	56×56×128
Stage 3	Conv Embedding	3×3, stride=2, filters=256	28×28×256
	Transformer ×4	heads=4, MLP ratio=4	28×28×256
Head	GAP + Dense (256) + Dropout	–	1×1×256
Output	Dense (1, sigmoid)	–	1
Total Params		3.18	

4.RESULT

a) Accuracy Graph

Figure 4.1 Represent the Accuracy graph of the model. The blue line indicates the Training Accuracy and orange line indicates the Validation accuracy . As shown in figure training accuracy increase from 63.5% to 86.26 % from zero to 20 epoch respectively. Validation Accuracy Remain higher than training accuracy indicating strong generalization. Early stop was executed at epoch 21 since the validation accuracy decreased over time while training accuracy increased over time signaling the onset of overfitting due to which the best weight was saved giving validation accuracy of 89.9% and training accuracy of 86.26% (peak value).

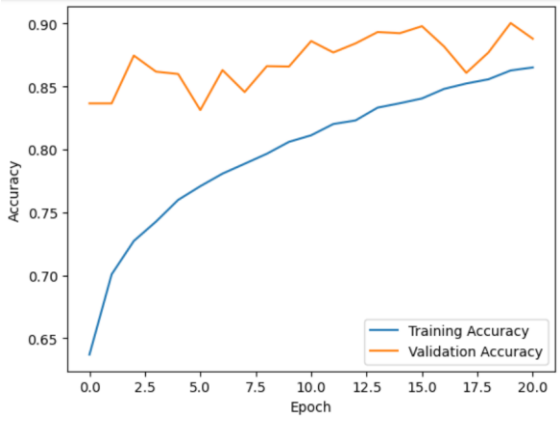

Figure 4.1: Accuracy Graph

2) Loss Graph

Figure 4.2 indicates the Loss Graph. Blue line indicates Training loss and Orange line indicates validation loss. The training loss is decreased from 0.64 all the way down to 0.29 indicating consistent learning without abrupt fluctuations from 0 epoch to 20 epoch respectively. The validation loss also decreases from 0.42 reaching as low as 0.24 at epoch 15.

The validation loss remains lower than training loss for most of the training process, suggesting that regularization techniques effectively prevents the model from overfitting. Minor fluctuations are seen at the end of epochs due to batch variance. However at 21 epoch, validation loss increases while training loss decreases indicating the onset of overfitting due to which early stopping was trigger saving the best weights.

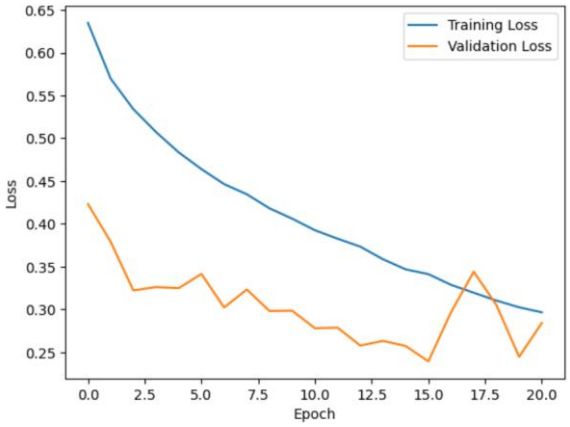

Figure 4.2: Loss Graph

3) Classification report

Table B represent the classification report of model on DF40 dataset. The model achieved an overall accuracy of 90% with balanced precision, recall and F1 score.

Table B

	Precision	Recall	F1 Score	Support
Fake	0.94	0.85	0.89	1606
Real	0.87	0.95	0.90	1606
Accuracy			0.90	3212
Macro Avg	0.90	0.90	0.90	3212
Weighted Avg	0.90	0.90	0.90	3212

4) Confusion Matrix

Figure 4.3 is heat map of confusion matrix. from this it can be stated that model perform slightly better at detecting real sample (approximately 95%) than fake samples (85%).

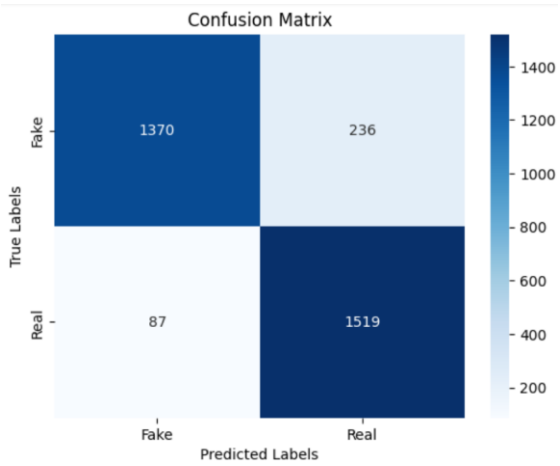

Figure 4.3

5.CONCLUSION

In this work, we applied Convolution Vision Transformer model proposed by Microsoft on the DF40 dataset. The result implies that the model work well with accuracy of 86.26% with balanced precision, recall and F1 score. It indicates that the model captures local as well global features enabling it detect subtle manipulation while maintaining strong generalization. The Confusion matrix indicates that the model work slightly better for real images than fake images. These finding highlights CvT's potential as a robust, single-model baseline for deep fake forensic without the need of complex ensembled or handcraft features.

Future Scope will explore the video dataset and Model will be able to detect deep fake videos. Finally, we aim to create a faster version of CvT that can run in real time on devices with limited resource and power like mobile phone.

REFERENCE

[1] Khudeyer, Raidah Salim, and Noor Mohammed Almoosawi. "Fake Image Detection Using Deep Learning." Informatica 47, no. 7 (2023).

[2] Altaei, Mohammed Sahib Mahdi. "Detection of deep fake in face images-based machine learning." Al-Salam Journal for Engineering and Technology 2, no. 2 (2023): 1-12.

[3] Altaei, Mohammed Sahib Mahdi. "Detection of deep fake in face images using deep learning." Wasit Journal of Computer and Mathematics Science 1, no. 4 (2022): 60-71.

[4] Rafique, Rimsha, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri. "Deep fake detection and classification using error-level analysis and deep learning." Scientific reports 13, no. 1 (2023): 7422.

[5] Raza, Ali, Kashif Munir, and Mubarak Almutairi. "A novel deep learning approach for deepfake image detection." Applied Sciences 12, no. 19 (2022): 9820.

[6] Abir, Wahidul Hasan, Faria Rahman Khanam, Kazi Nabiul Alam, Myriam Hadjouni, Hela Elmannai, Sami Bourouis, Rajesh Dey, and Mohammad Monirujjaman Khan. "Detecting deepfake images using deep learning techniques and explainable AI methods." Intelligent Automation & Soft Computing 35, no. 2 (2023): 2151-2169.

[7] Iqbal, Farkhund, Ahmed Abbasi, Abdul Rehman Javed, Ahmad Almadhor, Zunera Jalil, Sajid Anwar, and Imad Rida. "Data augmentation-based novel deep learning method for deepfaked images detection." ACM Transactions on Multimedia Computing, Communications and Applications 20, no. 11 (2024): 1-15.

[8] Wu, Haiping, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. "Cvt: Introducing convolutions to vision transformers." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 22-31. 2021.

[9] Soudy, Ahmed Hatem, Omnia Sayed, Hala Tag-Elser, Rewaa Ragab, Sohaila Mohsen, Tarek Mostafa, Amr A. Abohany, and Salwa O. Slim. "Deepfake detection using convolutional vision transformers and convolutional neural networks." Neural Computing and Applications 36, no. 31 (2024): 19759-19775.

[10] @article{yan2024df40, title={DF40: Toward Next-Generation Deepfake Detection}, author={Yan, Zhiyuan and Yao, Taiping and Chen, Shen and Zhao, Yandan and Fu, Xinghe and Zhu, Junwei and Luo, Donghao and Wang, Chengjie and Ding, Shouhong and Wu, Yunsheng and Yuan, Li}, journal={arXiv preprint arXiv:2406.13495}, year={2024} }