Indian Institute of Technology Roorkee

CSN-371

Artificial Intelligence

Research Paper Implementation

Deepfake Video Detection Using Convolutional Vision Transformer

Roopam Taneja
22125030
Computer Science and Engineering

Introduction

Deepfakes refer to hyper-realistic videos generated or synthesized using deep learning models. While these powerful video manipulation techniques offer potential benefits in fields like entertainment and education, they also pose significant threats like identity theft and spreading misinformation.

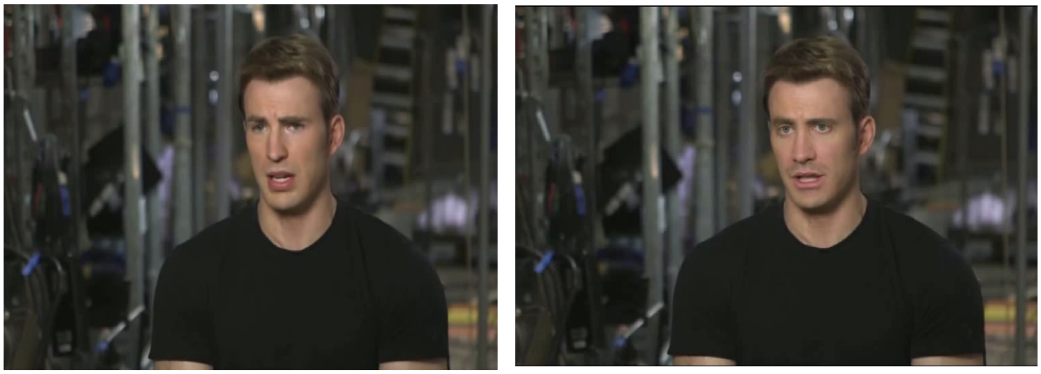

Figure 1: Left: Original - Image of Chris Evans. Right: Deepfake - Face-swapped with Jake Gyllenhaal. Part of Celeb-DF dataset.

This report details the implementation of a deepfake detection model based on the architecture proposed by Wodajo and Atnafu [1]. * The paper highlights limitations in contemporary deepfake detection models, particularly their inability to work across multiple and unseen spoofing techniques [2] and emphasizes the importance of data preprocessing in improving model efficacy [3].

The implemented model, named Convolutional Vision Transformer, leverages the strengths of both Convolutional Neural Networks (CNNs) for learning local features and Vision Transformers for capturing global dependencies via attention mechanisms.

The project involved implementing this architecture. The model was trained not just on the DeepFake Detection Challenge (DFDC) dataset [4] mentioned in the paper but on a combined dataset derived from DFDC and Celeb-DF datasets [5].

Proposed Solution: Convolutional Vision Transformer

The implemented deepfake detection system consists of two main components: Preprocessing and Detection.

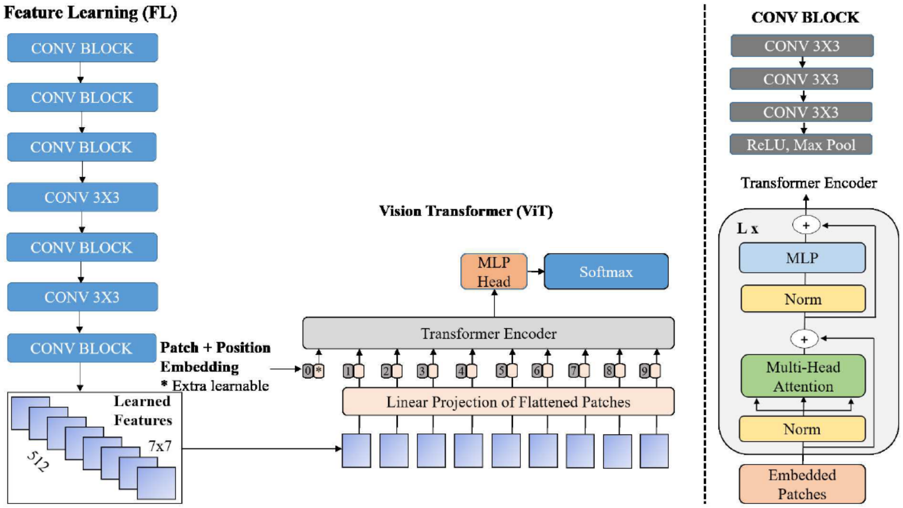

Figure 2: Convolutional Vision Transformer Architecture [1]

*The source code for this implementation is available at: GitHub

Preprocessing

This component processes the raw video data for the detection model.

1. **Face Extraction:** Faces are extracted from the input videos. BlazeFace library [6] is used for initial face localization, and the `face_recognition` library helps filter out false positives. Extracted faces are saved as 224x224 pixel RGB images.
2. **Data Augmentation:** To enhance model robustness against various image transformations, the training dataset images undergo an augmentation pipeline. This involves applying a mix of geometric and photometric augmentations using the `albumentations` library.

Detection

This component consists of training, validation, and testing phases. The model integrates a feature learning component and a Vision Transformer component.

- **Feature Learning Component:** This deep CNN extracts learnable features from the input face images. Its architecture is inspired by the VGG network [7].

  It comprises of 17 convolutional layers using 3x3 kernels, stride 1, and padding 1. ReLU activation and Batch Normalization are applied after each convolution. Five max-pooling layers (2x2 window, stride 2) are used, halving the feature map dimensions each time. Input dimensions are 224x224x3 while the output of this component is a 512x7x7 tensor.

- **Vision Transformer Component:** It processes the feature maps obtained from the previous component. Its architecture is based on the Vision Transformer proposed in [8].

  It uses a Transformer encoder architecture consisting of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Softmax function is applied to the output of the MLP head to produce probabilities between 0 and 1 for classification. An output probability $y < 0.5$ indicates `Real`, while $y \geq 0.5$ indicates `Fake`.

Datasets

To train and evaluate the model effectively, a diverse dataset is crucial. We utilized a combination of two prominent datasets:

- **DeepFake Detection Challenge (DFDC) Dataset:** At the time of its creation by Facebook AI, this was the largest publicly available face-swap video dataset [4].
- **Celeb-DF Dataset:** This dataset features deepfakes of celebrities, and is known for high visual quality and challenging artifacts. Including this dataset improves the input data quality and exposes the model to more subtle manipulations often targeting public figures [5].

Training and Evaluation

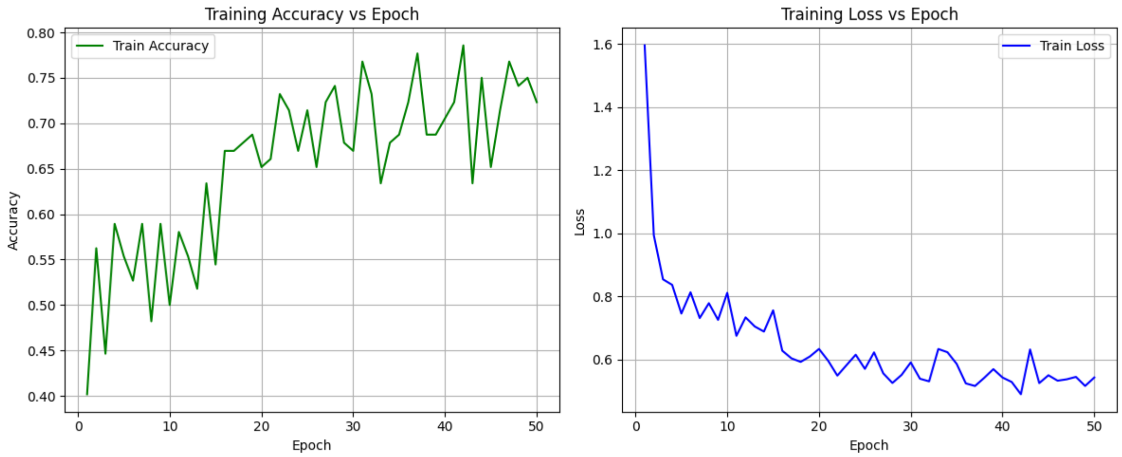

Figure 3: Plots of Training Accuracy and Training Loss

The model is implemented and trained using hyperparameters as specified in the paper. The model is trained for 50 epochs with a batch size of 32. Binary cross entropy loss function and Adam optimizer are used. The training statistics have been plotted in Figure 3.

The performance of the trained model was evaluated on the test dataset and the following metrics were calculated :

- **Accuracy:** The model achieved an accuracy of 81.25% on the test set.
- **Log Loss:** The log loss obtained was 0.6522.
- **AUC Score:** AUC score was 0.8889.

Conclusion

This project successfully implemented the Convolutional Vision Transformer architecture proposed in [1]. The model was trained and evaluated on a diverse dataset. The implementation achieved an accuracy of 81.25% and an AUC score of 0.8889 on the test dataset, demonstrating its potential for identifying deepfakes.

References

[1] Deressa Wodajo and Solomon Atnafu. *Deepfake Video Detection Using Convolutional Vision Transformer*. 2021. arXiv: 2102.11126 [cs.CV]. URL: https://arxiv.org/abs/2102.11126.

[2] Joshua Brockschmidt, Jiacheng Shang, and Jie Wu. "On the Generality of Facial Forgery Detection". In: *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*. 2019, pp. 43–47. DOI: 10.1109/MASSW.2019.00015.

[3] Polychronis Charitidis et al. *Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task*. 2020. arXiv: 2006.07084 [cs.CV]. URL: https://arxiv.org/abs/2006.07084.

[4] Brian Dolhansky et al. *The DeepFake Detection Challenge (DFDC) Dataset*. 2020. arXiv: 2006.07397 [cs.CV]. URL: https://arxiv.org/abs/2006.07397.

[5] Yuezun Li et al. "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics". In: *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 2020.

[6] Valentin Bazarevsky et al. *BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs*. 2019. arXiv: 1907.05047 [cs.CV]. URL: https://arxiv.org/abs/1907.05047.

[7] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV]. URL: https://arxiv.org/abs/1409.1556.

[8] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.