PROJECT REPORT

COMPARATIVE STUDY OF VIT WITH LSTM AND VIT WITH TCN FOR DEEP FAKE VIDEO DETECTION

MAXIMILLIAN MALEAKHI BUDIONO
22.K3.0001

Faculty of Computer Science
Soegijapranata Catholic University
2026

ABSTRACT

Deepfake videos are manipulated videos generated by artificial intelligence, deepfakes cause some dangerous problems, they can make people look like they are doing something they have not actually done, spread misinformation, defamation, and digital fraud. Therefore, it is important to develop a deepfake video detector with high stability and accuracy. In this study, we present a comparative study of 1 baseline model and 2 hybrid models, namely, Vision Transformer (ViT) combined with Long Short Term Memory (LSTM), and Vision Transformer (ViT) combined with Temporal Convolutional Networks (TCN). In order to assess its effects on detection performance, these hybrid models use the same ViT for feature extraction. UADFV and FF++ dataset was used, each video extracted into 32 sequence frames, These frames were passed into ViT, resulting in a 768 dimensional feature vector per frame. These extracted feature are then fed into temporal models LSTM or TCN. Both models evaluate using binary classification accuracy, binary cross entropy loss, and visualization metrics especially AUC ROC curve. Results indicate that ViT + TCN has higher AUC value with 0.7228 and more stable training than Vit baseline with 0.6741 and ViT + LSTM with 0.5833 AUC values. This study provides insight of the trade off between temporal modeling in real world scenarios.

Keyword: Deepfake Detection, Vision Transformer, LSTM, TCN, Temporal Modeling