Improving Deep Fake Detection: Integrating Spatial, Frequency, and Gradient Analyses

Aarav Patel
Alliance Academy for Innovation

Abstract

The increasing sophistication of deepfakes [Figure 6] represents a major threat to information integrity and public trust. While detection options do exist, many are often unable to generalize new forgery techniques. This project presents and evaluates a powerful hybrid spatio-temporal architecture designed to address this issue by learning both spatial and temporal artifacts from video data. The model uses EfficientNet-B0 to extract discriminative spatial features from individual video frames, and then employs a Gated Recurrent Unit (GRU) to model the temporal coherence of these features. The model achieves a final test accuracy of 87% and an AUC of the ROC Curve of 0.907 on the unseen and challenging Caleb-DF-v2 [1] dataset. These results validate this approach, but also highlight a generalization gap for improvement.

Introduction

In recent years, the global creation of deepfakes has significantly increased (21% in the last year) due to the rise in the availability of accessible and powerful generative models [2]. These hyper-realistic forgeries represent a challenge by enabling the spread of disinformation and decreasing public trust in the media. Current deepfake detection systems are limited by their inability to generalize. Models trained on specific datasets often learn to identify artifacts unique to the methods present in the set, and when presented with “wild” videos, often see a performance degrade. This shows that an effective detector needs to learn the features fundamental to the synthesis process.

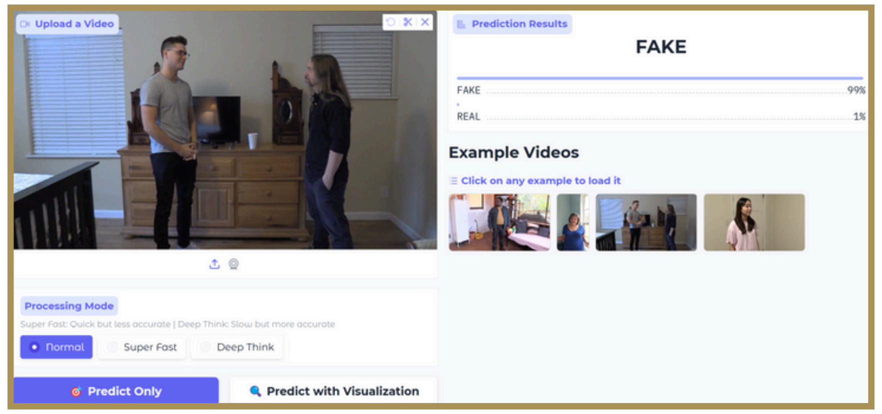
Upload a Video
Prediction Results
FAKE
FAKE 99%
REAL 1%
Example Videos
Click on any example to load it
Processing Mode
Super Fast: Quick but less accurate | Deep Think: Slow but more accurate
Normal
Super Fast
Deep Think
Predict Only
Predict with Visualization

Figure 1. Web App Example *(https://b-a-r-a-p-deepguard.hf.space)*

Methodology

Dataset and Preprocessing: A training dataset of 20,000 real and fake videos was used. A 2-stage preprocessing method was used: first, dlib [3] face tracker was used to detect and save facial bounding boxes from all videos. Second, 32 frames per video were extracted and processed into 3 feature streams: spatial (RGB), frequency (via Discrete Cosine Transform in Figure 5), and gradient (via Sobel operator), which were saved as .npz files.

Model Architecture: The model uses a hybrid spatio-temporal architecture, which consists of an EfficientNet-B0 for per-frame feature extraction and a Gated Recurrent Unit (GRU), which analyzes the sequence of features over time to learn temporal inconsistencies. The output from the GRU is passed to a small multi-layer perceptron (MLP), which has 2 linear layers, a ReLU activation, and dropout. This unit will produce a final output.

Training Procedure: The model was trained on 10 epochs on an NVIDIA RTX 3060 (12GB) using AdamW optimizer and Cosine Annealing learning rate scheduler. To operate within hardware constraints, training leveraged Automatic Mixed Precision (AMP) and gradient accumulation to achieve an effective batch size of 32.

$$F(u,v) = C(u)C(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} f(i,j)\cos\left[\frac{(2i+1)u\pi}{2N}\right]\cos\left[\frac{(2j+1)v\pi}{2N}\right]$$

Figure 5: Discrete Cosine Transform

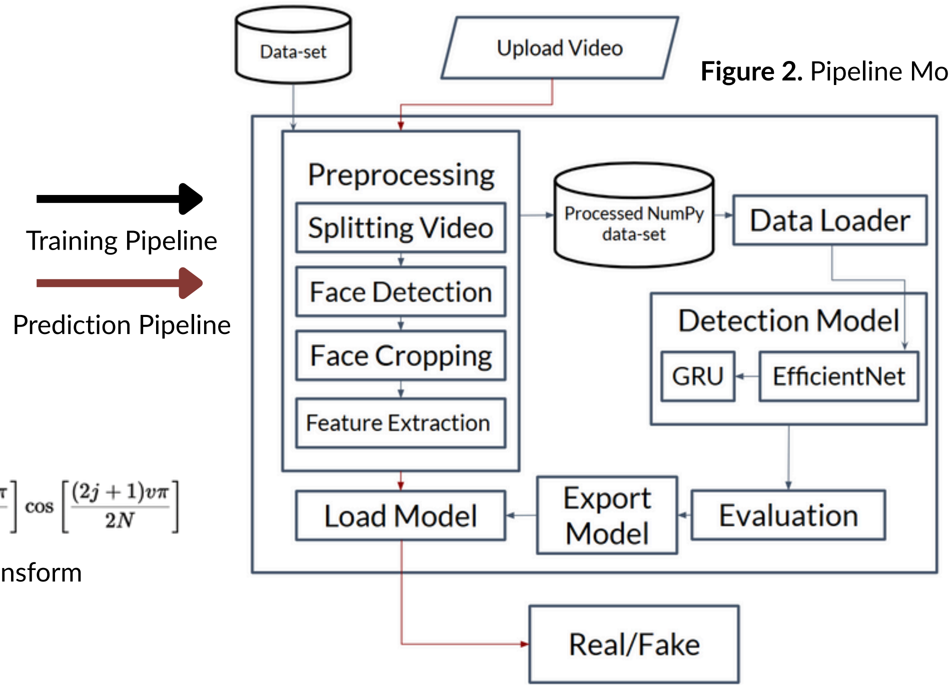

Figure 2. Pipeline Model

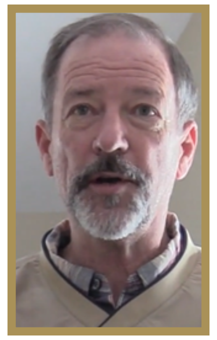

Figure 6: Deepfake Example *(This is fake)*

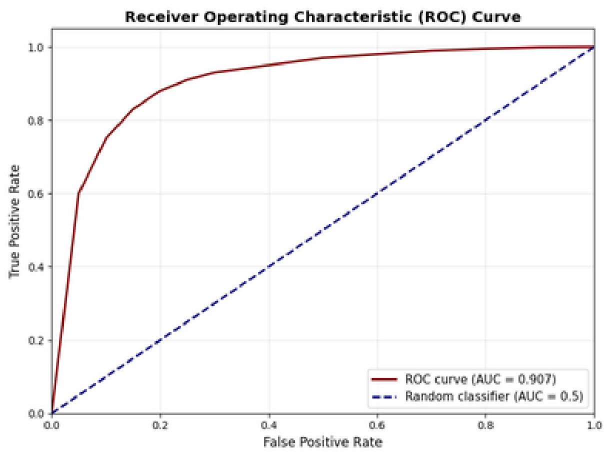

Figure 3. ROC Curve of model

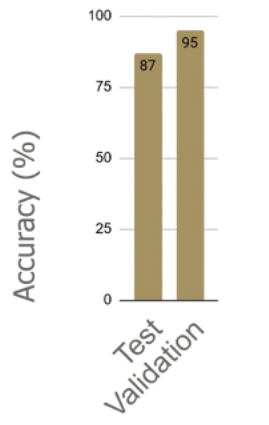

Figure 4. Model AUC

Results/Discussion

The model's performance was evaluated on the unseen Caleb-DF-v2 [1]. The model achieved a final test accuracy of 87% on this dataset. The Receiver Operating Characteristic (ROC) curve, shown in Figure 3, depicts the model's excellent ability to distinguish between real and fake classes. The Area under the Curve (AUC), a key performance indicator for classification performance, was 0.907, demonstrating a quality classifier. However, the drop in accuracy from validation accuracy to test accuracy highlights a possible generalization gap. It shows that while effective, it is not fully adapted to all of the various methods of deepfake generation.

Conclusion/Future Work

This research successfully implemented and evaluated a robust spatio-temporal method for deepfake detection. It demonstrated that a hybrid architecture using EfficientNet-B0 and a GRU can achieve a strong, generalizable performance on a challenging and unseen test set. Future work would be focused on closing this generalization gap by identifying more key, fundamental features to extract from samples.

View Source here: https://github.com/Barap1/Deepfake-Detection

References

1. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. ArXiv. https://arxiv.org/abs/1909.12962
2. Real-time deepfake fraud in 2025: AI-driven scams. Veriff. (2025, June 19). https://www.veriff.com/identity-verification/news/real-time-deepfake-fraud-in-2025-fighting-back-against-ai-driven-scams
3. Dlib C++ library. dlib C++ Library. (n.d.). https://dlib.net/