CIS 542 Digital Forensic

Topic: DeepFake Detection

Final Report

Team Members:

Ajmal Abbas

Vishnu Selvaraj

Abstract:

The extensive use of deepfake movies, which employ advanced generative techniques to edit or create hyper-realistic digital content, has raised significant concerns in domains such as digital forensics, media integrity, and cybersecurity. Deepfake detection is an important field of study because of its implications for fraud, deceit, and privacy concerns. In this study, we provide a robust deep learning-based methodology for detecting deepfake films. The model integrates Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and attention processes to effectively evaluate spatial and temporal features of video input. To train and evaluate the model, we employ both real and artificial video datasets, specifically from the Celeb-DF dataset. We employ preprocessing techniques such as data augmentation and others to normalize input data. The model's exceptional accuracy in distinguishing between real and fake videos demonstrates its efficacy. By showing how deep learning may be applied to tackle deepfake problems and provide a scalable technique for detecting manipulated data, this study advances the field of digital forensics.

Goal:
The aim of this project is to use cutting-edge deep learning techniques to create a reliable and scalable framework for identifying deepfake films. The increased difficulty of differentiating between real and modified digital content is what this approach attempts to address. The study specifically aims to improve media integrity and cybersecurity by efficiently analysing spatial and temporal information in videos to detect altered material.

Scope:
Included

- Deployment of a hybrid deep learning system that blends attention mechanisms, LSTMs, and CNNs.
- For better deepfake identification, video data's temporal and spatial properties are analysed. utilizing the Celeb-DF dataset for training and assessment, which comprises both synthetic and real films.
- Methods for preprocessing video data in order to prepare it for model training.

Exclusive:

- Detection of speech synthesis or audio modification.
- Analysing and implementing the model in real-time in commercial settings.
- Taking into account hostile attacks or identifying hidden deepfake types outside of the Celeb-DF dataset.

Methodology:

Frameworks and Techniques

- Transfer Learning: To extract spatial features, pre-trained base networks based on the xception and VGG16 models were employed.
- LSTM Network: Video frames with temporal relationships can be captured using LSTM networks.
- Attention Mechanism: By recognizing crucial frames, the attention mechanism improves the representation of temporal features.

Frameworks for Tools: Keres, TensorFlow
Libraries: Matplotlib, OpenCV, and NumPy
Hardware: rapid training in a GPU-accelerated environment. (Ryzen 7 Processor, Cluster Computer and Carnie Super Computer, Mac M1)

Steps:

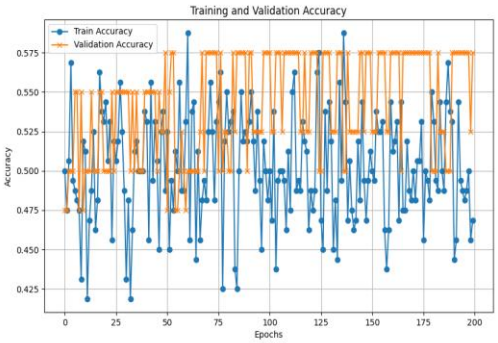
- Preprocessing includes normalization, 224x224 resizing, and frame extraction.
- CNN-LSTM and attention-based models for feature extraction and classification are developed as part of the model architecture.
- Training: Using the Adam optimizer and the Celeb-DF dataset with binary cross-entropy loss.
- Evaluation: Training and validation accuracy as well as learning curve visualization are used for assessment.

Results:

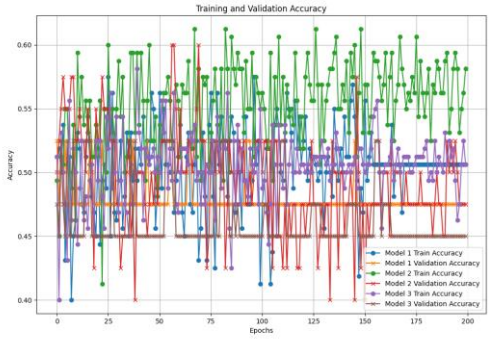
The Proposed deepfake detection framework achieved:

- Training Accuracy: ~75%
- Validation Accuracy: ~60%

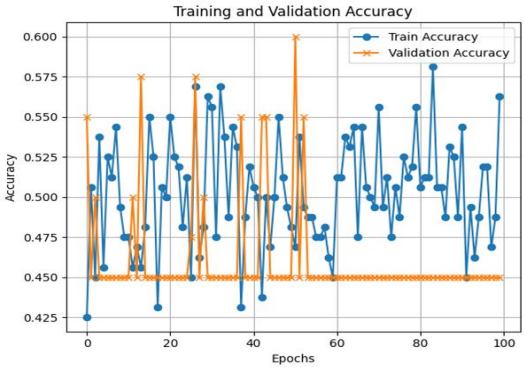
While the model Successfully learned spatial and temporal feature, validation performance indicated potential overfitting and room for improvement in generalization.



CNN-LSTM

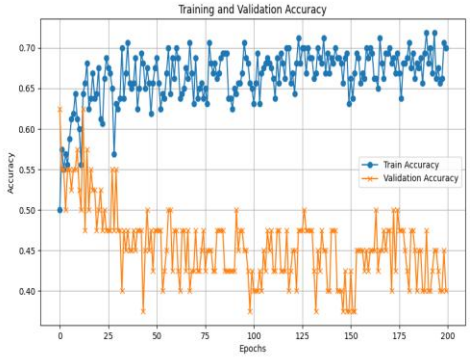


Ensemble Model

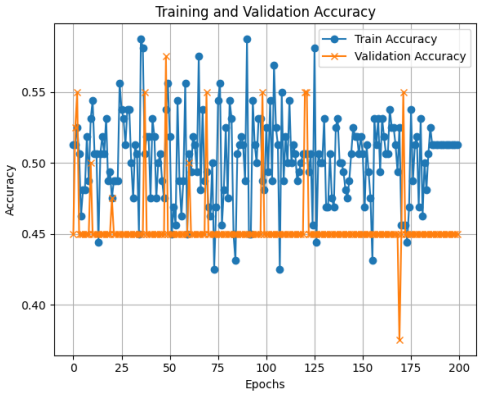


Final Ensemble Model

Transfer Learning in Exception



Hybrid Transfer Learning in Xception and VGG16

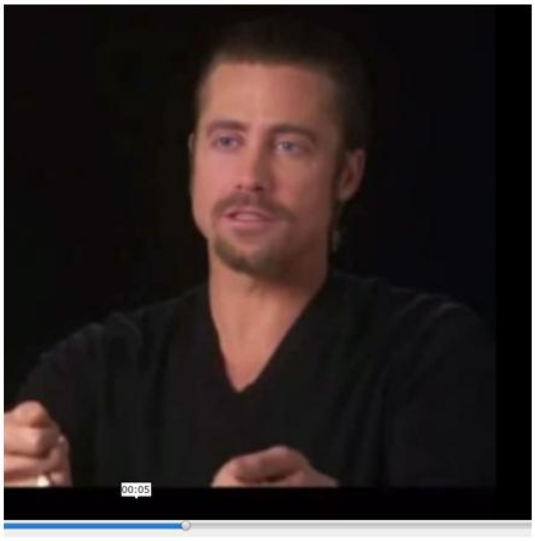


CNN-LSTM & Attention Mechanism

OUTPUT


```
# Testing with a video
uploaded_video_path = '/content/drive/MyDrive/UMASS_D/First Semsters/Digital forensics/Celeb-DF/Celeb-synthesis/id0_id16_0000.
result = predict_video(uploaded_video_path)
print(f"The uploaded video is predicted to be: {result}")

1/1 [==============================] - 0s 140ms/step
The uploaded video is predicted to be: Real
```




```
uploaded_video_path = '/content/drive/MyDrive/UMASS_D/First Semsters/Digital
result = predict_video(uploaded_video_path)
print(f"The uploaded video is predicted to be: {result}")

1/1 [==============================] - 0s 140ms/step
The uploaded video is predicted to be: Fake
```

Challenges Limitations in the dataset encounter:

- The Celeb-DF dataset has little diversity, which affects the model's capacity to generalize.
- High training accuracy combined with comparatively poor validation accuracy is known as overfitting.
- Computational Demand: Processing and training video data requires a lot of resources.

Solutions:

- Regularization methods such as data augmentation and dropout layers.
- To cut down on computational cost, transfer learning layers might be frozen.
- To avoid overfitting during training, stop early.

Conclusions

This experiment shows how CNNs, LSTMs, Transfer Learning and attention mechanisms can be combined to detect deepfakes. Despite the framework's encouraging performance in identifying tampered media, issues like overfitting and dataset constraints point to areas that require more study. Expanding the dataset, investigating adversarial defences, and refining the model for real-time applications could all be part of future research.

PROJECT LINK:

Project Github link