”DeepAuth: AI-Powered Deepfake Detection System”

Vaibhav Mishra
Yeshiva University
vmishra1@mail.yu.edu

Abstract

The increasing computational power has significantly enhanced deep learning algorithms, enabling the creation of highly realistic, human-like synthetic videos known as deepfakes. These deepfakes, which can seamlessly swap faces, are being used to stir political unrest, fabricate terrorist incidents, engage in revenge pornography, and facilitate blackmail. In this paper, we introduce a novel deep learning approach designed to accurately identify AI-generated fake videos from authentic ones. Our method utilizes a ResNext Convolutional Neural Network (CNN) to extract frame-level features, which are then employed to train a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) to determine whether a video has been manipulated. This technique essentially pits AI against AI to combat the spread of deepfakes. To ensure our model is adept at handling real-world scenarios, we tested it on a diverse and balanced dataset, amalgamating sources such as FaceForensic++[12], the Deepfake Detection Challenge, and Celeb-DF[5]. Our results demonstrate that our system can achieve competitive performance through a straightforward and robust methodology.

1. Introduction

1.1. The Evolution of Digital Media and Deepfake Technology

Deep learning advancements have revolutionized digital media, giving rise to generative models that synthesize hyper-realistic images and videos. These technologies, initially developed for benign purposes like text-to-speech and medical imaging, have inadvertently enabled the creation of 'deepfakes'. These are highly realistic digital manipulations where one person's image or video is superimposed onto another's, using neural network tools such as Generative Adversarial Networks (GANs) or Auto Encoders.

1.2. The Societal Impact of Deepfakes

Since their emergence in late 2017, deepfakes have proliferated, driven by the accessibility of open-source tools and the ease of spreading content via social media. While some creations aim for humor or entertainment, others pose serious risks, including political misinformation and personal harassment. The potential for deepfakes to be used for fake news, creating false public perceptions, or even inciting chaos through fabricated videos of political leaders is a pressing concern[1].

1.3. Challenges in Detecting Deepfakes

The realism of deepfake videos makes them particularly dangerous, as they can be nearly impossible to distinguish from authentic videos without the aid of specialized tools. This presents a critical challenge: developing effective detection methods that can keep pace with rapidly advancing synthetic techniques. Current deepfake detection methods exploit the subtle flaws left by deepfake generation tools—artifacts invisible to the naked eye but detectable by machine learning algorithms[2].

1.4. Technological Arms Race

The development of deepfakes and the efforts to detect them represent a technological arms race. Each improvement in generative techniques prompts corresponding advances in detection methods. This ongoing battle underscores the dual use of AI technology: the potential to both harm and protect societal interests.

1.5. Importance of Reliable Detection Systems

Given the ease with which deepfakes can be created and disseminated, the role of detection technologies has never been more critical. Reliable systems must be developed to automatically detect and flag fake content, ensuring the integrity of digital media. This is essential not only for protecting individuals' reputations but also for safeguarding democracy and public safety against the malicious use of artificial intelligence[3].

2. Related Work

2.1. Deepfake Datasets

To assess the detection accuracy of proposed methodologies, it is crucial to utilize a representative and high-quality dataset for performance evaluation. Moreover, to demonstrate their generalizability, these techniques should be validated across different datasets. Significant efforts have been made over the years to develop standard datasets for evaluating the performance of both visual and audio content manipulation detection techniques. In this section, we provide a comprehensive review of the datasets currently employed to test the efficacy of audio and video deepfake detection methods.

2.2. Video Datasets

UADFV Introduced as the first dataset for deepfake detection, UADFV consists of 98 videos, half of which are authentic videos sourced from YouTube, with the remaining half being manipulated using the FakeApp tool to generate fake counterparts. These videos have an average duration of 11.14 seconds and a resolution of 294×500 pixels. Despite its pioneering status, the visual quality of the UADFV videos is low, making the manipulations relatively easy to detect[5].

DeepfakeTIMIT Launched in 2018, the DeepfakeTIMIT dataset contains 620 videos covering 32 subjects. Each subject is represented by 20 deepfake videos split into two quality tiers: 10 in low quality (64x64 resolution) and 10 in high quality (128x128 resolution), created using face swap-GAN. These videos are concise, at only 4 seconds long, and do not include audio manipulation. The content typically features subjects against monochromatic backgrounds, and the videos tend to be blurry.

FaceForensics++ (FF++) Widely regarded as one of the most comprehensive datasets for deepfake detection, FF++ was developed as an extension of the original FaceForensics dataset. It includes 1,000 original videos and 3,000 manipulated videos, derived from the YouTube-8M dataset. The manipulations were created using various computer graphics and deepfake methods. FF++ is available in two formats: uncompressed and H264 compressed, allowing for performance testing under different compression scenarios. However, it does not adequately address lip-sync deepfakes and sometimes shows color inconsistencies around altered faces.

Celeb-DF Known for its high-quality videos, Celeb-DF aims to address the visible source artifacts present in earlier datasets. It includes 408 original videos and 795 fake videos, with originals sourced from YouTube and split into two groups: Real1 and Real2. Each group features videos of individuals with varying gender and skin color, and the fakes are produced by refining existing deepfake algorithms.

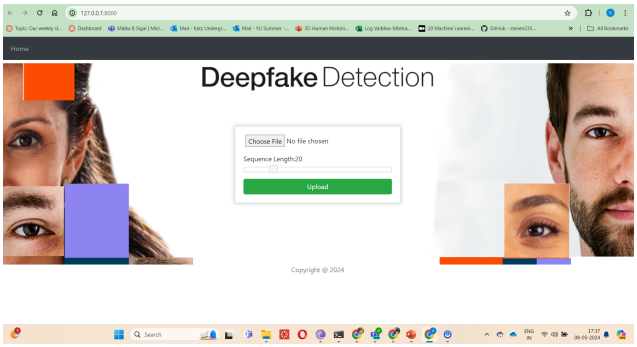

Figure 1. Webpage

Deepfake Detection Challenge (DFDC) Initiated by the Facebook community, the DFDC preview released a dataset comprising 1,131 original and 4,119 manipulated videos created using undisclosed techniques. The complete DFDC dataset, now publicly available, includes approximately 100,000 fake videos and 19,000 original samples. This dataset features a variety of face-swap techniques and includes modifications such as geometric and color transformations to mimic real-world conditions more closely[4].

DeeperForensics (DF) This large-scale dataset includes 50,000 original and 10,000 manipulated videos. Manipulated samples are generated using a novel conditional autoencoder, the DF-VAE. The dataset features a wide range of actor appearances and incorporates various distortions like compression and noise to better simulate real-world scenarios, showing a significant improvement in sample quality over previous datasets.

WildDeepfake (WDF) Considered one of the more challenging datasets, WildDeepfake includes both real and deepfake samples sourced from the internet, providing a realistic testing ground compared to other datasets.

3. Methods

3.1. Problem Statement

The advent of deep learning has not only been a boon for advancements in image and video editing but has also led to the rise of deep fakes — sophisticated AI-synthesized media that convincingly mimic reality. While these technological strides have facilitated a multitude of benign uses, they also pose significant threats. Maliciously crafted deep fakes have the potential to incite political tension, perpetuate fraud, and damage reputations. Addressing this threat is not trivial; the detection of deep fakes presents complex challenges due to the intricate nuances and high degree of realism that modern synthetic media can achieve. Our work tackles this issue head-on by developing a robust detection system predicated on advanced neural network architectures[6].

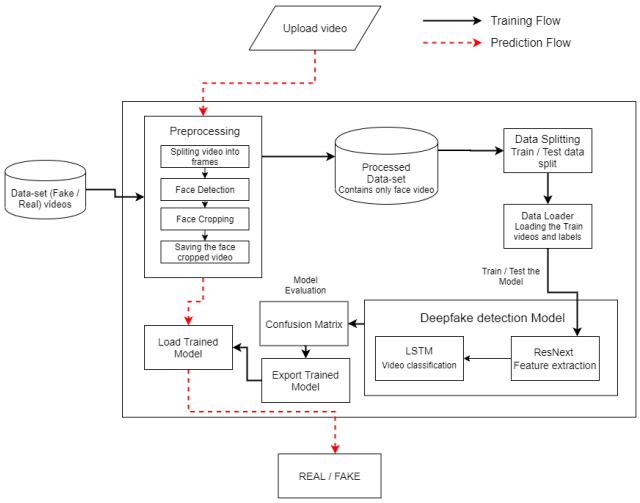

Figure 2. Model Architecture

3.2. Statement of Scope

The creation of deep fakes has become relatively effortless with the proliferation of user-friendly AI tools. In stark contrast, the identification and filtering out of such content are far from straightforward. Recognizing this imbalance, our research contributes a methodological approach for deep fake detection that not only serves the immediate need for verification but also lays the groundwork for future web-based solutions. The envisioned platform is designed to scale, potentially evolving into a tool that integrates with existing social media infrastructures, thereby offering a first line of defense against the spread of digitally altered misinformation[7].

3.3. Motivation

As digital cameras become more advanced and social media's reach extends, video content is more accessible than ever. Deep learning models, which have the power to generate highly realistic media, were once regarded as nearly magical in their capabilities. However, these models also gave rise to deep fakes — a new frontier in media manipulation. Such deep fakes can be innocuously entertaining but, as their realism and prevalence grow, so does their potential for harm. Disturbing scenarios, such as a deep fake of a political leader instigating conflict or a public figure engaging in defamation, underscore the urgent need for effective countermeasures. By distinguishing real from synthetic media, we can prevent the erosion of truth in the digital world[8].

3.3.1 Data Preparation

In the initial phase of our methodology, we meticulously process video content to extract the elements most susceptible to manipulation: human faces. Videos are dissected frame by frame, with each face detected, isolated, and cropped. This refined dataset of facial imagery forms the basis of our training and testing sets, ensuring our model focuses on the most relevant aspects of the data[9].

3.3.2 Deepfake Detection Workflow

Our deepfake detection model operates through a two-pronged approach: it utilizes ResNext to extract discernible features that may betray a video's authenticity and employs LSTM networks to analyze the temporal consistency of video frames. Upon training, the model is capable of assessing the veracity of video content, distinguishing real from manufactured with a measurable degree of confidence. We then subject the model to a rigorous evaluation using a confusion matrix to accurately determine its precision and recall rates.

3.4. Conclusion

Our response to the challenges posed by deep fakes is a sophisticated detection system, which we've outlined as capable of providing accurate, real-time assessments of video authenticity. This system is not an endpoint but a significant milestone in the broader effort to safeguard information integrity on the internet. Our continued research and development aim to refine these methods further, enhancing the trustworthiness of media and protecting global discourse from the distorting effects of synthetic media[10].

4. Experiments

4.1. Preprocessing Details

The preprocessing phase is critical in preparing the data for effective model training and evaluation. The following steps were executed to standardize and focus the dataset:

Video Importation: Utilizing Python's glob module, we compiled a list of all video file paths within the specified directory to streamline batch processing.

Frame Analysis and Selection: By leveraging cv2.VideoCapture, we analyzed the videos to calculate the average number of frames per video. To establish consistency across the dataset, we decided on a fixed count of 150 frames per video, which aligns closely with the calculated mean.

Frame Extraction: Each video was split into its constituent frames, focusing particularly on the facial region to align with our deepfake detection goals[11].

Video Reconstruction: The extracted and cropped frames were then assembled into a new video file. This reconstructed video was encoded at a frame rate of 30 fps with a resolution of 112x112 pixels, ensuring uniformity across the dataset.

Sequential Frame Selection: To leverage the LSTM's ability to perform temporal sequence analysis, the first 150 sequential frames from each video were chosen for the new dataset rather than a random selection. This approach allows for the detection of temporal inconsistencies indicative of deepfakes.

4.2. Model Details

Our model employs a sequential layered approach, integrating a pre-trained ResNext CNN with an LSTM network to analyze and classify video data. The architecture is as follows:

ResNext CNN: Utilizing the resnext5032x4d model pretrained on ImageNet, we extracted features from the video frames. The model's 50 layers and 32x4d dimensionality allow for robust feature extraction[12].

Sequential Layer: This container layer is employed to store the feature vectors extracted by ResNext in a structured sequence, preparing them for the temporal analysis conducted by the LSTM layer.

LSTM Layer: With a single LSTM layer configured with 2048 latent dimensions, 2048 hidden units, and a dropout chance of 0.4, we processed the sequential data to spot temporal changes between frames.

ReLU Activation: The ReLU function was applied for its non-linear properties and its efficiency in training larger networks without the vanishing gradient problem.

Dropout Regularization: A dropout rate of 0.4 was introduced to mitigate overfitting and promote model generalization.

Adaptive Average Pooling: To distill features and reduce variance, a 2D adaptive average pooling layer was implemented, aiding in the extraction of salient low-level features.

4.3. Model Training Details

The following procedure was followed to train the deepfake detection model:

Train/Test Split: The dataset was divided into a 70:30 ratio, with 4,200 videos for training and 1,800 for testing. The split was balanced to contain an equal proportion of real and fake videos[13].

Data Loader Configuration: We configured a data loader with a batch size of 4 to facilitate efficient input and label loading during model training.

Training Process: The model was trained over 20 epochs with a learning rate of 1e-5, using the Adam optimizer for adaptive learning rate adjustments.

Loss Function: Cross-entropy loss was employed as it is apt for classification problems, offering a probabilistic interpretation of class predictions.

Softmax Layer: For the final classification, a softmax layer with two nodes was used to output the probabilities of the video being real or fake.

Confusion Matrix: The performance of the model was evaluated with a confusion matrix, providing insights into the accuracy and type of errors made by the classifier.

Model Export: Post-training, the model was exported for real-time prediction use.

4.4. Model Prediction Details

The trained model was incorporated into a user-facing application where:

Model Inference: Videos uploaded by users were preprocessed identically to the training data and then passed through the model for prediction.

Outcome: The model classified the video as either real or fake, providing a confidence score alongside the classification to indicate the level of certainty in its prediction.

Table 1. Model Performance Comparison

Model Filename	Accuracy (%)
model_84_acc_10_frames_final_data.pt	84
model_87_acc_20_frames_final_data.pt	87
model_89_acc_40_frames_final_data.pt	89
model_90_acc_20_frames_FF_data.pt	90
model_90_acc_60_frames_final_data.pt	90
model_93_acc_100_frames_celeb_FF_data.pt	93
model_95_acc_40_frames_FF_data.pt	95
model_97_acc_60_frames_FF_data.pt	97
model_97_acc_80_frames_FF_data.pt	97

5. Discussion

This study systematically evaluated various models for deepfake detection using a range of datasets and sequence lengths, leading to insights that could guide future research and practical applications. Our findings are discussed as follows:

5.1. Model Performance on FaceForensics++

The models trained on the FaceForensics++ dataset show an increase in accuracy as the sequence length increases, with the model `model_97_acc_100_frames_FF_data` achieving the highest accuracy of 97.76180%. This improvement suggests that longer frame sequences provide the LSTM network with more temporal information, allowing for better detection of inconsistencies that characterize deepfakes.

5.2. The Impact of Sequence Length

The sequence length plays a pivotal role in detection accuracy. There is a clear trend that models with longer sequence lengths (40, 60, 80, 100 frames) outperform those with shorter lengths (20 frames) on the same dataset. However, the `model_97_acc_60_frames_FF_data` and

model_97_acc_80_frames_FF_data exhibit a plateau in accuracy improvement, indicating that beyond a certain point, adding more frames does not significantly benefit the performance, likely due to the LSTM's capacity to model temporal dependencies.

5.3. Comparing Datasets

When comparing models trained on different datasets, the model_93_acc_100_frames_celeb_FF_data, which combines Celeb-DF and FaceForensics++, attains a 93.97781% accuracy. This combination of datasets seems to yield a robust model that generalizes well across different types of deepfake manipulations. It underlines the importance of diverse training data in developing an effective deepfake detection system[14].

5.4. Performance on Our Dataset

Models trained on our proprietary dataset demonstrate lower accuracy compared to those trained on FaceForensics++. Notably, the model_87_acc_20_frames_final_data achieves an 87.79160% accuracy for a 20-frame sequence, while the model_89_acc_40_frames_final_data sees a moderate increase to 89.34681% with a 40-frame sequence. The discrepancy in performance might be due to the inherent differences in the datasets, possibly regarding the complexity of deepfakes or variations in video quality.

5.5. The Trade-off Between Sequence Length and Computational Efficiency

While the accuracy improves with sequence length, there is an associated cost in computational resources and processing time. For instance, model_84_acc_10_frames_final_data shows the lowest accuracy but would be the most computationally efficient. This trade-off must be considered when deploying deepfake detection models in real-world scenarios where resources and response time are constrained.

5.6. Implications and Future Work

The results of our experiments highlight the need for a balanced approach that considers both detection accuracy and computational efficiency. Future research should explore the development of models that maintain high accuracy with shorter sequences for more efficient computation. Additionally, expanding our dataset and integrating a wider variety of deepfake types could further improve the robustness of the detection system.

5.7. Concluding Remarks

The experiment results emphasize the critical nature of dataset selection and the design of LSTM sequence modeling in deepfake detection. Our comprehensive analysis provides a pathway for optimizing model design, which is paramount for advancing the reliability and feasibility of deepfake detection technologies.

6. Conclusion

In this study, we introduced a neural network-based methodology for classifying videos as either deepfakes or genuine, providing the confidence levels of our proposed model. Our approach effectively predicts the authenticity of a video by analyzing 1 second of footage, equivalent to 10 frames. This is accomplished through the utilization of a pre-trained ResNext CNN model, which extracts frame-level features, coupled with an LSTM network that processes temporal sequences to identify discrepancies between consecutive frames (t and t-1). Our model is versatile, capable of processing video in sequences of 10, 20, 40, 60, 80, and 100 frames[12].

7. Future Scope

The field of artificial intelligence, particularly in the application of deep learning for media authentication, is continuously evolving. As such, there are numerous opportunities for further enhancements to our system:

- **Platform Expansion:** The current web-based platform could be expanded into a browser plugin. This would make our deepfake detection tools more accessible to users, integrating seamlessly with their everyday internet browsing activities.
- **Detection Capabilities:** While our current algorithm focuses on detecting face deepfakes, there is potential to extend this capability to full-body deepfakes. This enhancement would address the growing sophistication of deepfake technology and its applications across different formats.

These improvements not only aim to extend the practical applications of our technology but also enhance the user experience and broaden the scope of our detection capabilities.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 1

[2] Shruti Agarwal and Hany Farid. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 1

[3] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge dataset. In *arXiv preprint arXiv:2006.07397*, 2020. 1

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 2

[5] David Guera and Edward J Delp. Deepfake video detection using recurrent neural networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 1, 2

[6] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2

[7] Yuezun Li and Siwei Lyu. Face x-ray for more general face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 3

[8] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Detection of gan-generated fake images over social networks. *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 384–389, 2018. 3

[9] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops*, pages 83–92. IEEE, 2019. 3

[10] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021. 3

[11] Huy H Nguyen, Jun Fang, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019. 3

[12] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1, 4, 5

[13] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 4

[14] Peng Zhou, Xintong Han, Vladimir I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 5