Deepfake Video Forgery Detection

APURV JINDAL
Maharaja Agrasen Institute of Technology

Abstract- ***With the proliferation of digital video content and the advancements in video editing tools, video forgery has become a critical concern in various domains, including journalism, surveillance, and legal proceedings. Detecting forged videos is a challenging task due to the increasing sophistication of forgery techniques. Traditional techniques include forensic watermarking, temporal inconsistencies analysis, and sensor pattern noise analysis. This research paper proposes a novel deepfake detection approach combining ResNet-50 and LSTM networks, a deep learning architecture known for its excellent performance in image recognition tasks. The proposed method leverages the strengths of both spatial and temporal modeling to enhance the detection accuracy. The ResNet-50 model is utilized to extract spatial features from individual frames, capturing visual cues and inconsistencies introduced by deepfake manipulations. The LSTM network is employed to model the temporal dependencies between frames, enabling the detection of subtle temporal artifacts that may indicate the presence of deepfakes. To train the model, a large-scale dataset comprising both real and deepfake videos is utilized. The dataset is carefully curated, ensuring a diverse range of deepfake manipulations and real-world scenarios. The ResNet-50 backbone is pre-trained on a large image dataset, allowing it to learn generic visual representations that are then fine-tuned for deepfake detection. The LSTM network is trained to capture the temporal dynamics and patterns specific to deepfake videos.***

Indexed Terms- ***CNN, Deepfake Detection, LSTM, ResNet50, Video Forgery Detection***

I. INTRODUCTION

Deepfake technology, driven by advancements in machine learning and computer vision, has raised significant concerns regarding the potential for malicious manipulation and dissemination of visual media. Deepfakes refer to synthetic videos that convincingly depict individuals engaging in actions or uttering statements they never actually performed. These manipulated videos have the potential to deceive viewers and undermine trust in visual content.

The need for effective deepfake detection methods has become paramount to mitigate the risks associated with this technology. Traditional techniques, relying on manual inspection and visual artifacts, are inadequate in the face of increasingly sophisticated deepfake algorithms. Consequently, researchers have turned to advanced machine learning and deep learning approaches to develop robust detection mechanisms.

This research paper focuses on the detection of deepfakes using a combination of ResNet-50 and LSTM architectures. ResNet-50, a deep convolutional neural network (CNN), excels at capturing spatial features and has demonstrated exceptional performance in various computer vision tasks. LSTM, a recurrent neural network (RNN), is specifically designed to model sequential and temporal dependencies, making it suitable for analyzing the temporal consistency of video sequences.

By combining ResNet-50 and LSTM, this paper aims to leverage the complementary strengths of both architectures to improve deepfake detection accuracy. The ResNet-50 component enables the extraction of spatial features from individual frames, effectively capturing visual cues and irregularities introduced by deepfake manipulations. The LSTM component processes the temporal dynamics between frames, capturing subtle temporal artifacts that might be indicative of deepfakes.

II. METHODOLOGY

1. DATA COLLECTION

FaceForensics++

FaceForensics++ is a forensics dataset consisting of original video sequences that have been manipulated with automated face manipulation methods: This dataset was developed by the researchers at Technical University of Munich. The data has been sourced from youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries. As we provide binary masks the data can be used for image and video classification as well as segmentation. The used dataset was downloaded using lossless compression rate factor of 23 using the h264 codec original as well as the altered videos had the same compression factor. All these videos were silent videos and had about 300-400 frames on average in each video with a duration of about 10 seconds.

Fig 1: Stills from videos in the dataset FaceForensics++

- Celeb-DF

Celeb-DF dataset includes original videos collected from YouTube with subjects of different ages, ethnic groups and genders, and corresponding DeepFake videos.This dataset was made by researchers Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu who wrote a paper titled "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics"[7].

Fig 2: Stills from videos in the dataset Celeb-DF

III. DATA PREPROCESSING

In pre-processing, we perform the following steps on the videos present in the database:

1. Frame Extraction: The first step is to extract frames from the video. This involves reading the video file and extracting individual frames at regular intervals. Each frame represents a single image in the video sequence. Here, we have chosen the first 150 frames in each video.
2. Region of Interest (ROI) Selection: In our case, we have chosen ROI as the face present in the video as we are detecting only deepfakes where the face has been swapped.
3. Image Enhancement: Enhanced the quality of frames by denoising and improving image clarity.
4. Normalization and Scaling: From the above framed new video is created by combining the frames. This video is created at 30fps with a resolution of 112x112. This helps us in normalizing all the videos.

These preprocessing steps help prepare the video data for subsequent forgery detection algorithms. By extracting relevant features and addressing noise or inconsistencies, these steps enhance the accuracy and effectiveness of video forgery detection systems.

IV. MODEL

The model architecture for the Deepfake Video Detection project uses two main components: an LSTM (Long Short-Term Memory) network and ResNext.

We used ResNext-50 with 32x4d. ResNext-50(32x4d) is a 4 dimensioned model with 32 nodes in each dimension and 50 layers. It's capable of learning 25 x 106 parameters. After passing through a cooling layer, we pass the feature vector to the sequential layer. The sequential layer passes the model to the LSTM layer with 2048 latent dimension and 2048 hidden layers with a chance of dropout rate of 0.4. The output of LSTM is further processed and passed on to an adaptive pooling layer which passes it to the softmax layer which predicts 'FAKE' or 'REAL'.

Multiple frames of videos were taken to train this model.

V. RESULT

Model Name	No of videos	No of Frames	Accuracy
model_84_acc_10_frames_final_data.pt	6000	10	84.21461
model_87_acc_20_frames_final_data.pt	6000	20	87.79160
model_89_acc_40_frames_final_data.pt	6000	40	89.34681
model_90_acc_60_frames_final_data.pt	6000	60	90.59097
model_91_acc_80_frames_final_data.pt	6000	80	91.49818
model_93_acc_100_frames_final_data.pt	6000	100	93.58794

Table 1: Accuracy of different frame model

Model Performance: The model performance improves as the number of frames used for prediction increases. This is evident from the increasing accuracy values associated with the models trained on different numbers of frames. The accuracy ranges from 84.21% for 10 frames to 93.587% for 100 frames. Increasing the number of frames used for prediction allows the model to capture more temporal information and better understand the video dynamics. This leads to improved accuracy in detecting deepfake videos. From the given results, it appears that the accuracy continues to improve up to 100 frames. This suggests that using 100 frames per video provides the highest accuracy achieved among the tested frame counts (10, 20, 40, 60, 80, and 100).

Model Selection: Depending on the desired trade-off between accuracy and computational efficiency, one can choose the appropriate model. If higher accuracy is the priority, the model trained with 100 frames (model_93_acc_100_frames_final_data.pt) achieves the highest accuracy of 93.59%. However, if computational efficiency is a concern, models with fewer frames could be considered.

Overall, the results suggest that increasing the number of frames used for prediction improves the model's ability to detect deepfake videos. The optimal frame count may vary based on the specific requirements and constraints of the application.

CONCLUSION

The project on Deepfake Detection using LSTM and ResNext CNN demonstrates the effectiveness of combining temporal modeling with spatial feature extraction for detecting deepfake videos. The model architecture, which includes an LSTM network and a ResNext CNN, achieves promising results in accurately identifying manipulated videos. The combination of LSTM and ResNext CNN leverages the strengths of both architectures. The ResNext CNN efficiently extracts spatial features from individual frames, while the LSTM captures temporal dependencies between frames. This comprehensive approach enables the model to identify both spatial and temporal inconsistencies introduced by deepfake manipulations.

We also made a website where user can verify their videos

Fig 3: Video uploading

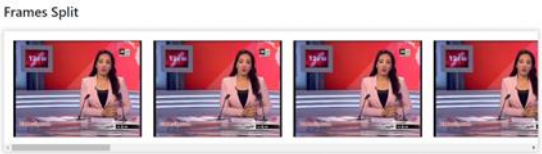

Fig 4: Frames split

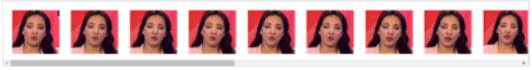

Fig 5: Face Crop from the extracted frames

Fig 6: Video prediction

REFERENCES

[1] Chih-Chung Hsu, Tzu-Yi Hung, Chia-Wen Lin and Chiou-Ting Hsu, "Video forgery detection using correlation of noise residue," 2008 IEEE 10th Workshop on Multimedia Signal Processing, Cairns, Qld, 2008, pp. 170-174, doi: 10.1109/MMSP.2008.4665069.

[2] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[3] Shelke, N.A., Kasana, S.S. A comprehensive survey on passive techniques for digital video forgery detection. Multimed Tools Appl 80, 6247–6310 (2021). https://doi.org/10.1007/s11042-020-09974-4

[4] A. W. A. Wahab, M. A. Bagiwa, M. Y. I. Idris, S. Khan, Z. Razak and M. R. K. Ariffin, "Passive video forgery detection techniques: A survey," 2014 10th International Conference on Information Assurance and Security, Okinawa, Japan, 2014, pp. 29-34, doi: 10.1109/ISIAS.2014.7064616.

[5] A. V. Subramanyam and S. Emmanuel, "Video forgery detection using HOG features and compression properties," 2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP), Banff, AB, Canada, 2012, pp. 89-94, doi: 10.1109/MMSP.2012.6343421.

[6] M. Aloraini, M. Sharifzadeh and D. Schonfeld, "Sequential and Patch Analyses for Object Removal Video Forgery Detection and Localization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 3, pp. 917-930, March 2021, doi: 10.1109/TCSVT.2020.2993004

[7] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[8] Deng L, Suo H, Li D. Deepfake Video Detection Based on EfficientNet-V2 Network. Comput Intell Neurosci. 2022 Apr 15;2022:3441549. doi: 10.1155/2022/3441549. PMID: 35463269; PMCID: PMC9033321.

[9] Stehouwer J, Dang H, Liu F, Liu X, Jain A. On the detection of digital face manipulation. arXiv preprint. arXiv:1910.01717 (2019)

[10] Rahmouni N, Nozick V, Yamagishi J, Echizen I. Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE workshop on information forensics and security (WIFS). IEEE; 2017. p. 1–6.

[11] Jing Zhang, Yuting Su, and Mingyu Zhang. 2009. Exposing digital video forgery by ghost shadow artifact. In Proceedings of the First ACM workshop on Multimedia in forensics (MiFor '09). Association for Computing Machinery, New York, NY, USA, 49–54. https://doi.org/10.1145/1631081.1631093