Deepfake Video Detection With Optical Flow and Recurrent Neural Network

Berker Tomaç
Department of Computer Engineering
TED University
Ankara, Turkey
tomaciberker@gmail.com

***Abstract*—In today's world, the capabilities of generative models are becoming better and more each day. It continues to evolve and it can be scary because it becomes challenging to detect if a video is generated with artificial intelligence or real. This problem can cause serious issues in our world such as forgery, copyright infringement and much more. This research paper directly addresses the problem discussed and offers an automatic detecting system by using machine learning and computer vision algorithms with a total of 800 deepfake and original videos by taking the leverage of the FaceForensics++ dataset. An ensembled architecture of two CNN layers one for optical flow and one for the videos fed into an RNN to detect if a video is fake or real. The score of this proposed architecture has a precision of 0.85 for deepfake videos while it has 0.82 for real. Future studies with more videos and more frames analyzed for a video need to be accomplished in order to fully understand the potential of the proposed architecture.**

I. Introduction

Usage of artificial intelligence is becoming an essential part of most people's daily routines. It makes our lives significantly easier. However, the evolution of deep learning comes with problems too. Since the quality becomes significantly enhanced each day, it becomes harder to identify whether a video or a photograph is generated with artificial intelligence or not. This situation raises a concern among people. Innocent people's faces can be used on a crime video they didn't know existed. It is one of the biggest problems faced in human history because, with the continuous evaluation of these models, it might become very hard to identify if that video is generated with AI or not [1]. Some websites and applications are capable of generating deepfake videos which solidify the issue presented in the earlier part. With the evolution of these programs, it becomes very difficult to manually recognize if a video is generated with artificial intelligence or a real-life video [2]. Therefore, it becomes compulsory to develop new, innovative ideas and solutions to protect people otherwise it might ruin lives [1]. In response to these problems and challenges that are identified, this research study offers the development of a robust identification for deepfake videos generated with artificial intelligence with the help of FaceForensics++ Dataset which has been sourced from 977 YouTube videos and provides approximately 500 GB of videos that are h264 compressed with the rate factor of raw/0 consisting of both original and deepfake videos [1].

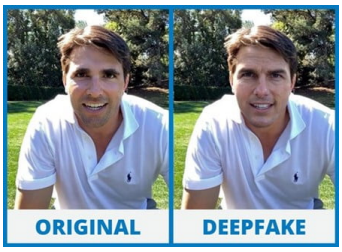

Fig. 1. Original and Deepfake Video

due to the limited number of resources, only 800 videos(400 deepfake, 400 original) are used for this study. Our approach uses machine learning and computer vision techniques to automatically identify the origin of a video. Our methodology offers an ensemble approach, which takes leverage of optical flow and EfficientNetb0 as a feature extractor and provides the information to a recurrent neural network(RNN). In the final step, the loss function maps the output of the recurrent neural network between 0 and 1 which leads to the identification of deepfake and real video [2].

A brief introduction and general summary are provided in this section. The following parts will dive into the literature review, methodology, results and discussion.

II. Literature Review

The rapid advancement of deepfake technology poses a critical threat to digital media authenticity, demanding robust detection mechanisms. The literature in deepfake detection is vast, encompassing an array of approaches each grappling with the challenges of discerning real from counterfeit content. There are several different approaches and methods to this problem such as transformers [3], self supervised learning techniques, fully unsupervised techniques [4], convolutional neural network models [5], recurrent neural network models and much more. However, this study is built upon two specific technique and research papers. These two techniques are usage of optical flow and recurrent neural networks. It also leverages the technique of ensembled version off convolutional neural network with a survey study to gain more insights about the techniques that are used for this problem.

A landmark study by Amerini et al. [6] delves into the realm of optical flow features for identifying subtle manipulations in video sequences. Their work is predicated on the principle that while facial manipulations can be visually convincing, they often introduce anomalies in the flow of motion between frames that can be meticulously detected through optical flow analysis. By harnessing such discrepancies, Amerini [6] have laid the groundwork for models that can differentiate between natural and artificial facial movements, adding a crucial layer of detection that complements spatial analysis with 81.16 and 75.46 accuracy scores by using VGG16 and ResNet50 cnn architectures. [6].

A seminal study by Guera and Delp explores the use of recurrent neural networks (RNNs) for the detection of deepfake videos [7]. Their research is grounded in the observation that while deepfake videos can appear visually seamless, they often introduce temporal inconsistencies that can be detected by analyzing the sequence of frames. The authors propose a novel detection system that employs a convolutional neural network (CNN) to extract frame-level features, which are then analyzed by an RNN to capture and classify temporal anomalies indicative of video manipulation. This dual-stage approach leverages the strengths of both CNNs and RNNs, providing a robust mechanism for identifying deepfake content. Guera and Delp evaluated their method using the HAHO dataset with 600 videos [8], half of which were deepfakes, and demonstrated that their system achieves high accuracy in detecting manipulated videos, significantly outperforming baseline methods. Their work represents a critical advancement in digital media forensics, addressing the growing threat of deepfake technology by providing a reliable tool for automated detection [7].

Complementing this approach, Bonettini et al explored the power of CNN ensembles in video face manipulation detection. Leveraging a modified version of the EfficientNetB4 architecture, they introduced an attention mechanism that allows the network to focus and learn from the most informative regions of the input frames, thereby enhancing the classification accuracy. Their methodology employed an ensemble of networks trained with different focuses, including siamese training strategies that are adept at capturing high-level semantic information from various models to outperform baseline methods [9].

Yu et al. survey the field extensively, categorizing the existing detection strategies while examining the trajectory of technological evolution in deepfake generation and detection methods. They emphasize the shift toward deep learning, which offers promising results due to its capacity for high-level data representation learning, outperforming traditional methods that relied heavily on handcrafted features [10].

In the midst of technical advances, the socio-cultural dimensions of deepfake technology cannot be overlooked [11]. As deepfakes permeate various aspects of society, the consequences of their misuse become a tangible concern, from spreading disinformation to violating individual privacy. This calls for a detection approach that is not only technologically sound but also cognizant of the ethical landscape, balancing the act of innovation with responsibility.

Given the complexity of deepfake detection, our research aims to contribute to the studies that will be conducted in the feature.

III. Methodology

As it stated in the literature review section, this study develops a system with an ensemble approach of different studies. There are 4 stages of the method offered in this step. At first, collecting and organizing the dataset, then calculating the optical flow of each video. Following with providing two EfficentNetb0 models one for the videos and one for the optical flows. In the third step, the outputs of the feature extractor are given to a recurrent neural network(LSTM) with 4 layers. In the last step, the output of LSTM is mapped with a sigmoid function that will successfully determine the type of video.

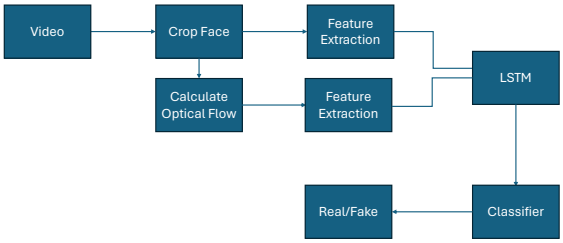

Fig. 2. Proposed Architecture (0 is the output of original, 1 is fake)

A. Data Collection

After conducting detailed research and also with the help of the survey [10] mentioned in the literature review part, it is observed that FaceForensics++ [12] was the most commonly used dataset for deepfake video detection models. Because of that reason, this research is made with FaceForensics++. However, as stated in the introduction part, only 800 videos which are divided equally into two parts (400 deepfake, 400 original) used for this study. In the literature review part it is observed that most studies used 600-1000 videos. Because of that reason, 800 videos are chosen randomly from the dataset [12]. These 800 videos first went through a face detection and cropping process since the different parts of the deepfake videos are actually in the face. By using only face-cropped video it is aimed to have a better accuracy than the usage of whole video. These new 800 cropped face videos

Fig. 3. FaceForensics++ Dataset

are divided into train, validation and test parts with 72,14,14 percentage. For each of these parts the number of original and fake videos provided equally to prevent any imbalanced classification problem.

B. Optical Flow Calculation

After collecting these 800 face-cropped videos for 3 different parts, each of them is provided to the system to calculate the optical flows of these videos. These optical flow videos are then stored in the current folder structure next to the cropped face videos. This means under the train folder there will be two sections both for original and fake ones and there will be two more sections under both the original and fake ones with face cropped video folder and optical flow video folder. It actually means that the number of videos is going to 1600 instead of 800 since both optical flow and face-cropped videos will provided to two different feature extractors. In Figure 4, you can observe the difference in optical flow for an original and deepfake video.

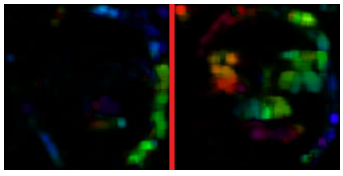

Fig. 4. Difference of optical flow for original and deepfake video

C. Feature Extraction

With the calculation of the optical flow videos, the system was ready in terms of the needed data. Two adjustments need to be made before continuing. At first, the videos are resized as 224x224 for a standardised structure and faster training process. Following resizing, the number of frames to be examined is set to 10 instead of the whole video due to the lack of resources. Limiting the number of frames makes the training process 20 times faster but it decreases the chance of learning new things for feature extractors. After making these adjustments, the data are provided to feature extractors. For this study, two EfficentNetb0 feature extraction models are used. One for the face-cropped videos and one for the optical flow videos. As an output, 512 features are provided from each of these models.

D. Recurrent Neural Network

In this part of the proposed method, the information from the two feature extractors is provided to the LSTM with an input size of 1024. This study developed an LSTM model with 4 layers. For the activation function, leaky ReLU is used instead of regular ReLU function since it addresses the dying ReLU problem by allowing a small, non-zero gradient when the unit is active, thus preventing neurons from becoming permanently inactive and improving the model's ability to learn complex patterns. The loss function that is used in the developed architecture uses sigmoid and it maps the output provided by the LSTM between 0 and 1. Since the original and deepfake videos are labeled as 0 and 1 in the developed model classification is completed with the provided output.

IV. Results

After the classification is completed from the model developed, Accuracy,Precision, Recall and F1-Score is calculated to test how successful is the proposed method. These results are provided in the TABLE 1.

TABLE I
DEEPFAKE DETECTION CLASSIFICATION RESULTS

	Precision	Recall	F1-Score	Accuracy
Real	0.82	0.88	0.79	0.83
Deepfake	0.85	0.89	0.79	0.83

V. Discussion

The scores are higher than the optical flow study [6] but lower than the system developed with LSTM [7] and an ensemble of CNN's [9]. It should be remembered this study is completed with a limited number of resources with an i3 CPU and rtx3060 GPU. It was only conducted with 800 videos from the dataset and only 10 frames of the video were observed due to lack of hardware. Even with these problems, this study provides decent accuracy and precision for the deepfake detection problem and it can contribute to different studies that will be conducted in the future. You can see the comparison of the results of the studies in TABLE 2

TABLE II
SUMMARY OF STUDIES ON DEEPFAKE VIDEO DETECTION

Study	Architecture	Data set name	Data Set Size	Metrics
[7]	CNN and RNN	Mix	600	0.97
[6]	Optical Flow	FaceForensics	1000	0.81
[9]	Ensemble CNN	DTC	50,000	0.93
Ours	O.Flow + RNN+ CNN	FaceForensics	800	0.85

VI. Conclusion

In conclusion, the development of an ensemble architecture of past studies demonstrates the potential of deep learning against current-day problems. This study provides a satisfactory result for the deepfake video detection problem with a limited number of resources. To test this model's real potential with real-world scenarios; more data, a computer with a minimum GPU of RTX 3070 and a CPU with i7 is suggested since this study couldn't increase the number of frames analyzed for a video more than 10.

REFERENCES

[1] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A comprehensive review of deepfake detection using advanced machine learning and fusion methods," *Electronics*, vol. 13, no. 1, 2024.

[2] L. Passos Júnior, D. Jodas, K. Costa, L. Souza Jr, D. Rodrigues, J. Del Ser, D. Camacho, and J. Papa, "A review of deep learning-based approaches for deepfake content detection," 10 2023.

[3] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *CoRR*, vol. abs/2102.11126, 2021.

[4] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and non-aligned double jpeg detection using convolutional neural networks," *Journal of Visual Communication and Image Representation*, vol. 49, 08 2017.

[5] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," pp. 2952–2956, 05 2020.

[6] I. Amerini, L. Galteri, R. Caldelli, and A. Bimbo, "Deepfake video detection through optical flow based cnn," pp. 1205–1207, 10 2019.

[7] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2018.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[9] N. Bonettini, E. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," pp. 5012–5019, 01 2021.

[10] Y. Peipeng, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, 04 2021.

[11] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[12] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.