Deepfake Detection Using Optical Flow and Recurrent Neural Networks

Joshua Mathew

***Abstract*—This paper introduces a novel method for deepfake video detection using dense optical flow and a convolutional LSTM neural network. This detection system can analyze the apparent motion in a video over time and distinguish between real and fake videos. Deepfake detection methods work only in the spatial domain looking at one frame of a video at a time, this system exploits this and analyzes both the spatial and temporal information of a video. The model achieved 97% and 70.5% training and validation accuracies on a dataset made of 400 real and 400 deepfake videos. The model achieved 71.4% accuracy on a testing dataset of 100 real and 100 deepfake videos. With further model tuning, this detection system can be a robust tool for deepfake detection.**

I. Introduction

Deepfakes are images and videos in which the original subject's likeness is replaced with someone else's using deep neural networks. There have been many recent advances in the development of deepfakes, the best deepfakes now are extremely realistic and undetectable by the human eye.

Deepfakes have the potential to severely disrupt the political, social, and economic sectors across the globe. They can be used to depict politicians and celebrities in compromising or pornographic images. In ordering to combat the threat that deepfakes pose, many governments and private companies are heavily investing in deepfake detection research. Google, Facebook, Microsoft, Amazon, and many universities have collaborated to create deepfake datasets and detection challenges to spur progress in this field.

This paper will introduce a novel method for deepfake video detection utilizing dense optical flow and recurrent neural networks. The motivation behind this method is that the optical flows will characterize the motion of the video subject's face and the recurrent neural network will learn about how this motion changes over time. Because deepfake generators work only to create a realistic image in the spatial domain for each frame individually, the recurrent neural network will be able to spot temporal inconsistencies or patterns in a deepfake video not found in real videos.

II. Proposed Method

Fig. 1. *Traffic Optical Flow*

The first part of this deepfake detector is the generation of optical flow fields. An optical flow field is a vector field showing the apparent motion between each frame of a video. Optical flow is calculated by assuming that the pixel colors associated with an object are persistent across each frame.

The optical flow fields were generated using the Farneback method [1] implemented in the OpenCV computer vision library. An example flow field showing the motion of cars on the highway is depicted in Fig. 1. The colors in the flow field are a function of the speed and direction of each object.

Layer (type)	Output Shape	Param #
flow_input (InputLayer)	(None, 40, 200, 200, 2)	0
conv_lst_m2d_1 (ConvLSTM2D)	(None, 40, 200, 200, 20)	15920
batch_normalization_1 (Batch	(None, 40, 200, 200, 20)	80
max_pooling3d_1 (MaxPooling3	(None, 40, 100, 100, 20)	0
conv_lst_m2d_2 (ConvLSTM2D)	(None, 40, 100, 100, 10)	10840
batch_normalization_2 (Batch	(None, 40, 100, 100, 10)	40
max_pooling3d_2 (MaxPooling3	(None, 40, 34, 34, 10)	0
conv_lst_m2d_3 (ConvLSTM2D)	(None, 34, 34, 5)	2720
max_pooling2d_1 (MaxPooling2	(None, 17, 17, 5)	0
flatten_1 (Flatten)	(None, 1445)	0
dense_1 (Dense)	(None, 512)	740352
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 1)	513

Fig. 2 – Neural Network Structure

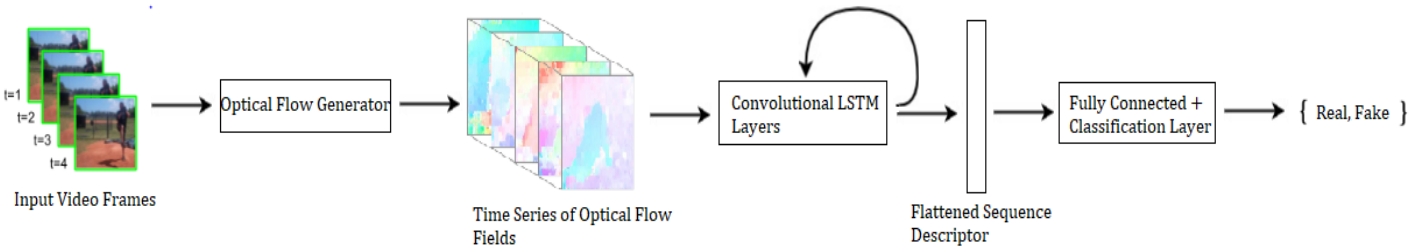

Fig. 3 – Overall Detection System Architecture

Dense optical flow fields were generated for the first 40 frames of each video. Each flow field was then decomposed into a magnitude and direction matrix. The matrices were then resized to have a shape of 200×200 and stacked on top of each other, creating a 200×200×2 array. Thus, for every video, a 40×200×200×2 array was generated. Stacking the matrices in this way makes sense because every stacked magnitude and direction matrix element is associated with the same pixel in the original video frame.

The second part of the detection system is a convolutional LSTM neural network [2]. The convolutional LSTM layers replace the matrix multiplication of a normal LSTM layer with convolution operations. Therefore the 200×200×2 magnitude and direction matrices can be inputted into the network without flattening them into a vector.

The neural network structure is composed of 3 convolutional LSTM layers, 3 batch normalization layers, 3 max pooling layers, a 512-neuron fully connected layer, and finally a sigmoid classification layer. The first 2 LSTM layers output a time series of arrays and the final LSTM layer outputs a single array which is flattened and fed into the fully connected layer before being classified. The full structure is depicted in Fig. 2. The Adam optimizer with a learning rate of 10^{-4} and binary cross-entropy loss function were used for training.

III. Experimental Results

Optical flow fields were generated for 800 videos taken from the Celeb-DF deepfake database [3]. Half of these videos were real and the other half deepfakes. The model was trained on these videos using an 80/20 training/validation split. The model was then tested on the optical flows generated from 200 other videos also taken from Celeb-DF and equally split between deepfake and real videos.

Model	Training acc. (%)	Validation acc. (%)	Testing acc. (%)
Conv-LSTM, 40 frames	95.30	70.50	71.40

Table 1 – Classification Results

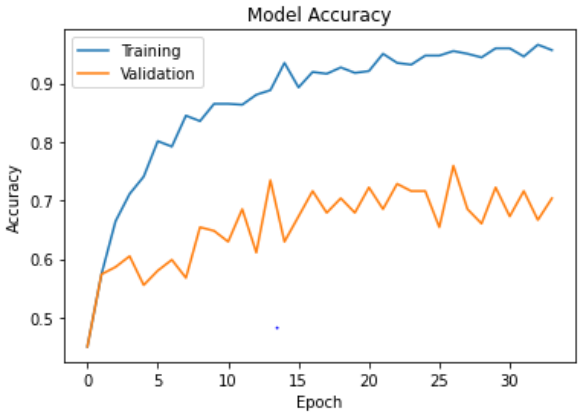

Fig. 5 - Model Accuracy Training History

Fig. 6 - Model Loss Training History

IV. Summary and Ongoing Work

Looking at Table 1 and Fig. 5 it is clear that the model is overfitting to the training data based on the large discrepancy between training and validation accuracy. Fig. 6 illustrates how the model begins to overfit around epoch 11, this is when the validation loss starts increasing as the training loss rapidly falls.

In an attempt to improve the model accuracy a grid search was done over certain model hyperparameters (learning rate, batch size, hidden layer nodes) but no combination of parameters resulted in any improvement in accuracy.

After visualizing the generated optical flow fields, it became apparent that these optical flow fields were not consistent with what would be expected for subtle facial movements. The flow fields over the face contained a large amount of noise and were not smooth. This is likely a result of the inaccuracies associated with how the Farneback method tracks the pixels, as the facial pixels are all of similar color.

The Farneback method used to generate the optical flows was developed seventeen years ago and has been improved on by other methods. A new state-of-the-art method was employed to generate the flow fields, the LiteFlowNet convolutional neural network [4][5].

A face detector was also added using a Haar Cascade classifier [6] to extract only the optical flow over the facial region. This is to remove any noise associated with the background which is not useful for learning.

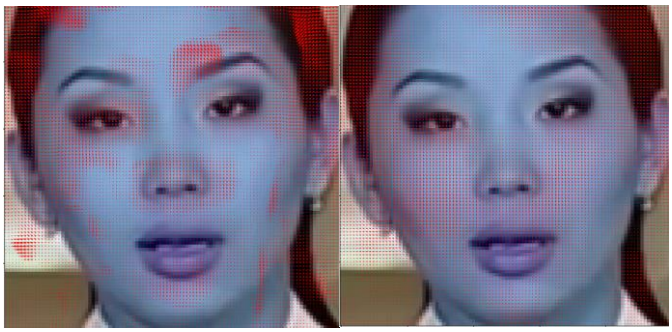

Fig. 7 – Optical flow fields using Farneback method (left) and LiteFlowNet (right)

Looking at Fig. 7 it is clear that the optical flow generated by LiteFlowNet is much smoother and less noisy than the Farneback method. The cleaner optical flow should make it easier for the neural network to learn to distinguish between real and deepfake videos.

Another issue is that when taking just one 40 frame clip from each video there were only 800 data points for training and validation, considering each clip as a data point. This is not enough data to train a deep neural network from scratch. In order to augment the dataset, all the videos were split into multiple clips. This increased the size of the data set by almost tenfold.

The next steps for this project are to generate the optical flows with the larger dataset using LiteFlowNet and added face extractor, and then train the convolutional LSTM network. Given a larger and better pre-processed dataset, the model should be able to better learn to distinguish between the real and fake videos and the accuracy should improve.

REFERENCES

[1] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," *Image Analysis Lecture Notes in Computer Science*, pp. 363–370, 2003.

[2] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," Neural Information Processing Systems, 2015.

[3] "Celeb-DF (v2): A New Dataset for DeepFake Forensics," *CVML Celeb-DF dataset*, 16-Mar-2020. [Online]. Available: http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics.html. [Accessed: 27-Mar-2020].

[4] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[5] S. Niklaus, "A Reimplementation of LiteFlowNet Using PyTorch," GitHub, 2019. [Online]. Available: https://github.com/sniklaus/pytorch-liteflownet. [Accessed: 23-Apr-2020].

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.