DeepFakeDetection with Convolutional and Recurrent Neural Networks

Stanford University

Yumeng Yue, Zewen Zhang

yuey3@Stanford.edu, zwzhang@stanford.edu

Introduction

DeepFakeVideo: Deep learning based generative models are capable of synthesizing hyper-realistic images, speeches, music, and even videos.

Figure : DeepFake Videos

While many of them are intended to be humorous, others can actually be malicious, having potential of being misused, like blackmail, spread fake news or fake terrorism event.

Dataset

We work with **DeepFakeDetection** dataset released along with **FaceForensics++**, provided by Google and Jigsaw, which is the most up-to-date dataset on forgery videos.

Figure : DeepFakeDetection Video

The dataset contains over **363 original videos** and **3068 manipulated videos** from 28 actors in 16 different scenes. Original videos are recorded with paid and consenting actors, and manipulated videos are generated via DeepFakes. Here we worked with visually lossless compression rate factor of 23 using the h264 codec.

We work with a total of **726 videos** with equal number of original videos and manipulated videos, where **582 videos for training** and **144 videos for test**.

Data Processing

Figure : Face Extraction Size

Figure : Face Extraction for original videos

Figure : Face Extraction for fake videos

Since DeepFake mainly focus on face manipulation, we extract faces from each frame with FaceNet's MTCNN face detector module. By varying different margin size, we choose to crop **image size of 256*256** with a **margin size of 100**, we'd crop the entire face without getting too much background information.

Models

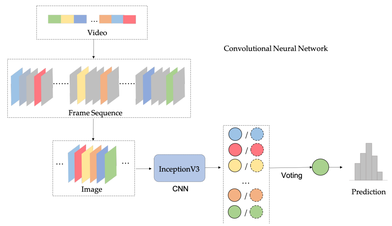


Figure : Schematic of Image Fake Detection Based Method

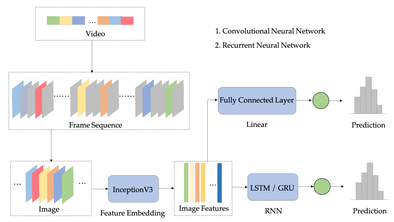


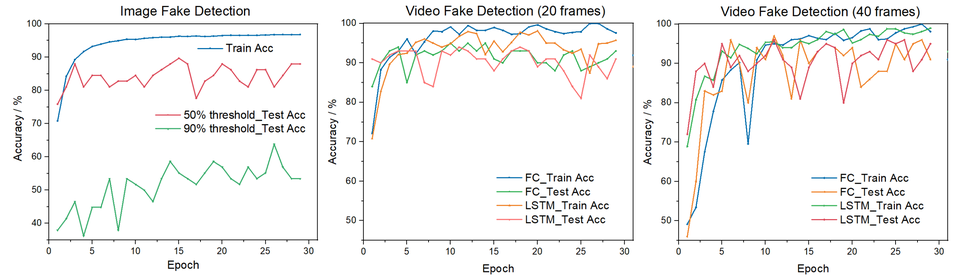
Figure : Schematic of Video Fake Detection Based Method

For **image fake detection method**, we aim to evaluate if we could achieve desirable detection accuracy based on only few individual frames from a video. The input will be individual frames. Our model will classify input frames as fake or original, and a classification on video will be made based on all its frame. Different threshold positive percentage will be evaluated here.

For **video fake detection method**, we concatenate all individual frame from a video as a sequence input. All image embeddings are from InceptionV3. Both linear classification and recurrent neural networks(LSTM) are examined here to understand how does temporal information affect model efficiency.

Results

Category	Model	Input	Detection Accuracy
Image Fake Detection	InceptionV3 (threshold 90%)	40	88
	InceptionV3 (threshold 50%)	40	53
Video Fake Detection-Linear	InceptionV3 + FC Layer	40	93
	InceptionV3 + FC Layer	20	91
Video Fake Detection-LSTM	InceptionV3 + LSTM	40	91
	InceptionV3 + LSTM	20	**95**



- For **image fake detection method**, if we set 90% right accuracy threshold, we get only 53% accuracy in test set, however with 50% right accuracy threshold, we could get 88% accuracy.
- For **video fake detection method**, if we work with 40 frames for each video, we would get 91% accuracy with only fully connected layer after embedding, but with recurrent neural network LSTM after embedding, we get 95% accuracy. Furthermore, if we only work with 20 frames, we get 93% accuracy with fully connected layer, but only 91% for LSTM network.

Analysis

Figure : Images from Fake Video Predicted to be Fake

Figure : Images from Fake Video Predicted to be Original

- Deepfake video manipulation could be highly deceptive, where even could not be noticed by human eyes. For current models, we could not capture these differences. Analysis of why embedding concatenation work would be important.
- High detection accuracy shown here might come from limited data available here. All these fake videos are made by face swapping, however we only have limited faces in the dataset, making the case easy, while in real scenarios we might need more powerful and general methods.

Conclusion

- Video deepfake detection is similar to its image counterpart, where simple image detection pipeline could achieve high detection accuracy.
- With all frame embedding concatenation, we could further improve video detection accuracy by around 5%.
- With increased frame numbers used for detection, there's not much difference in detection accuracy, indicating we might be able to achieve high accuracy with smaller models and more efficient pipeline design.
- Recurrent neural networks have limited effects on detection accuracy in our case. Intuitively, temporal information should help, however it's not clear here. Possible explanation might be all the frames are similar in each video, making temporal information less important in this case.