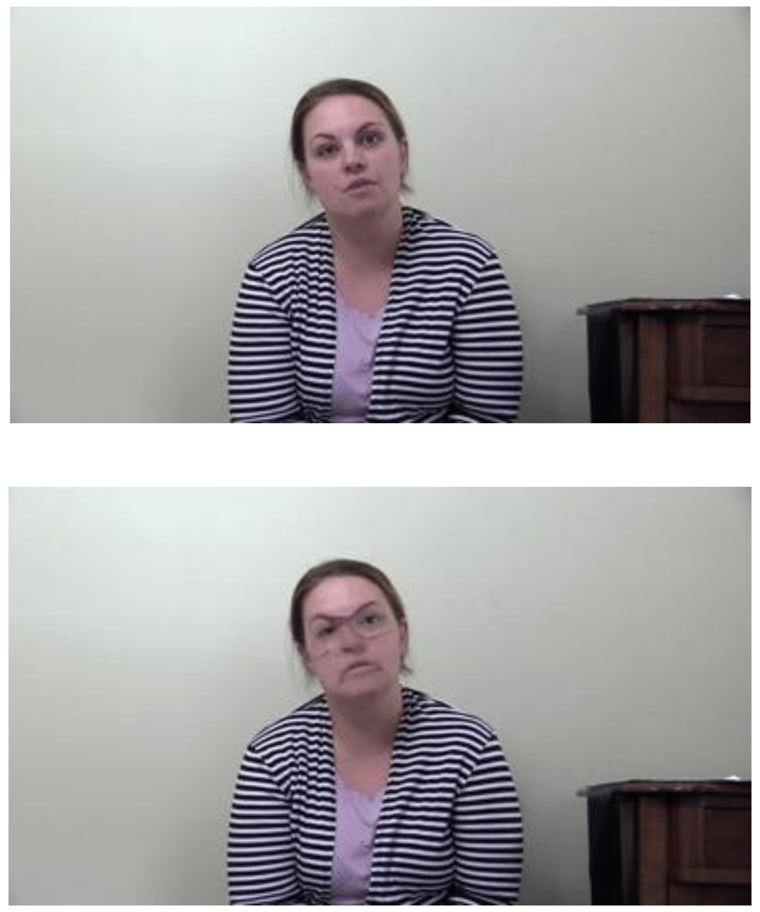
Investigating Deepfake Detection in Videos

Emily Hannon, Yogesh Rawat, Rahul Ambati
UCF Center for Research in Computer Vision, 4328 Scorpius St. Suite 245
Orlando, FL 32816-2365
Email Address: emilyhannon@knights.ucf.edu

Abstract

- A deepfake is a face/identity swap where machine learning software is used to replace a face/feature in an original image or video with the likeness of someone/something else.
- With this technology being open source, it is relatively easy to create a deepfake. However, deepfakes are becoming more realistic and more challenging to detect. With the current prevalence of social media in everyday life, deepfake technology can aid in disseminating disinformation or creating "fake news".
- While multiple methods exist to detect deepfakes, new technology makes certain fakes extremely difficult for both the human eye and computers to detect.
- Our goal was to explore the different methods that can be used to detect face swap deepfakes on Facebook's Deepfake Detection Challenge (DFDC) Dataset and compare the results to the contest participants on Facebook's Kaggle challenge. This data set was selected as one of the most comprehensive and ethically obtained datasets of deepfake face swap videos currently available.
- We are training a basic model on two different forms of the dataset. The first contains jpeg images of faces extracted from the videos. The second contains the mp4 videos from the original dataset. Since there are distractions added to the test set, we expect that working with full videos as opposed to images will produce better results. Once we obtain findings from training on the model with the videos, we will proceed to work on improving the model.



Example of frames from real vs. fake videos that are easy for the human eye to detect.

Objective and Research Method

Objective

- To detect changes made to video media using perturbations. Detection is accomplished by training a deep learning convolutional neural network model to detect when a deep fake face swap has been performed on media. Videos are classified as fake (media has been altered) or real.

Research Method

- We have preprocessed the same dataset in two different ways. In the first, we have extracted the faces from both the testing and training sets as .jpg files and resized them to be 128x128. In the second we resized the videos to be 256x256 and took a 224x224 random crop.
- For the extracted images of faces, we used resnet18, and on the resized videos we used resnet 3D 18. Both were pretrained models from torchvision. Code was executed in the Newton GPU Visualization Cluster at the UCF ARCC.

Contents of the Dataset

The Facebook DFDC dataset (1) contains **a test set, a train set, and a validation set.** The data set also applies two different types of additional perturbations besides face swap.

The first type is called an ***augmenter*** and applies **geometric and color transforms, frame rate changes, audio removal, adding noise, altering resolution,** etc. This has been applied to about **70%** of all videos at randomly chosen videos.

The second type of additional permutation is called a ***distractor*** and **overlays** various kinds of objects (including images, shapes, and text) onto about **30%** of all videos.

The **test** set contains **10,000** video clips. Half of these videos (5,000) are fake. Half are "organic content from the internet" and the other half are previously unseen source videos. 79% of all videos in the test set have additional augmentations (besides face swap) and the facial filters are applied here for the first time (meaning they aren't seen in any of the other sets). Some labels indicate audio removal from videos, but in this paper, we are focusing on video only.

The **train** set contains **119,154** video clips with 486 unique subjects and no additional augmentations besides the Deepfake face swap. Approximately 83.9% (100,000) are fake videos. Methods used to create the fakes include DFAE, MM/NN face swap, NTH, and FS-GAN. These videos have been labeled either fake or real.

The **validation** set contains **4,000** videos with 214 unique subjects. Half (2,000) are fake. The same methods to create the fakes from the training set have been used along with an additional unseen method, StyleGAN. These videos have been labeled either fake or real.

References

[1] Brian Dolhansky, Joanna Bitton, B.P. J. L. R. H. M. W. C. C. F.The deep-fake detection challenge (dfdc) dataset.Facebook AI(2020).

[2] Yisroel Mirsky, W. L.The creation anddetection of deepfakes: A survey.ACMComputing Surveys 54, 7 (2020), 1–41

Conclusion

- The model using extracted faces is more accurate but more uncertain, this could be due to the removal of other information in the images as a side-effect of the face extraction. This model may improve in accuracy if there is a way to reduce the loss.
- While the cropped videos show a lower accuracy, the model is less uncertain. This is possibly due to the additional information in the frames that would typically be cropped out in the face extraction. If there would be an easier way to increase accuracy here than to decrease loss in the face extraction model, then this model may be more desirable.

Results

Training on **frames with faces extracted** with resnet 18 using 4 epochs on 8700 videos: Average loss on test of 4.31 and an accuracy on the test set of approximately 56.39%

Training on **cropped and resized videos** with resnet 3D 18 using 4 epochs on 8700 videos: Average loss on test of 1.08 and an accuracy on test of 51.42%.

Example of a deepfake difficult for the human eye to detect.

Example of a frame from a video with additional perturbations.

UCF Center for Research in Computer Vision
UCF The Center for Initiatives in STEM
Duke Energy