AIT Deep Learning Final Report: Deepfake Detection

Christina Wang CWANG5@SWARTHMORE.EDU
Isabel Grondin ILG1@WILLIAMS.EDU

1. Introduction

As technology is advancing, deepfakes, modified pieces of media that imitate someone's likeness using digital modification, are becoming increasingly realistic. This has serious implications as deepfakes can be used to spread misinformation, as it is challenging for a viewer to discern real media from fake media. Our goal with this project was to detect videos of celebrities that are deepfakes.

2. Previous Solutions

One previous solution that we paid particular attention to was the model depicted in "High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction" (Li et AL. 2021).They created a CNN based solution that utilized data augmentation, target-specific region extraction and other deep learning techniques. They were able to achieve an AUC of 0.978, with a F1-score of 0.9628, on the Celeb-DF v2. They did not, however, utilize an LSTM. We want to see what results we could produce utilizing sequencing, in addition to a base CNN model. Unfortunately, there is not any data regarding human prediction versus machine prediction on this dataset.

3. Dataset

For our project, we used the Celebrity Deepfake dataset, which we found through Papers With Code. This dataset was released in 2019 with the goal of refining the basic deepfake generation algorithm to improve upon the resolution, color matching, and face markers in deepfake videos. The Celebrity Deepfake dataset is heavily skewed towards fake videos, with 5,639 deepfakes and only 590 real videos. All of the original videos were taken from publicly available YouTube videos, taking into consideration the proportions of celebrities from different genders and races. We chose to use a data set with celebrities, so that a deepfake would be more easily detectable by an average viewer. A data set with a random person would make it more challenging for a person to visually detect a deepfake.

In our preprocessing phase, we balanced the dataset so that our training was not biased towards real or fake videos. We organized our data by grouping the first 9 individual frames from each of the videos together, and mapping them to one binary result (0 if fake, 1 if real). Unfortunately due to limitations in computational resources, this was the maximum number of frames that we use from each video. Next we cropped all the frames to 299 by 299, the input frame size for the InceptionV3 model, and standardized the mode of the videos to RGB.

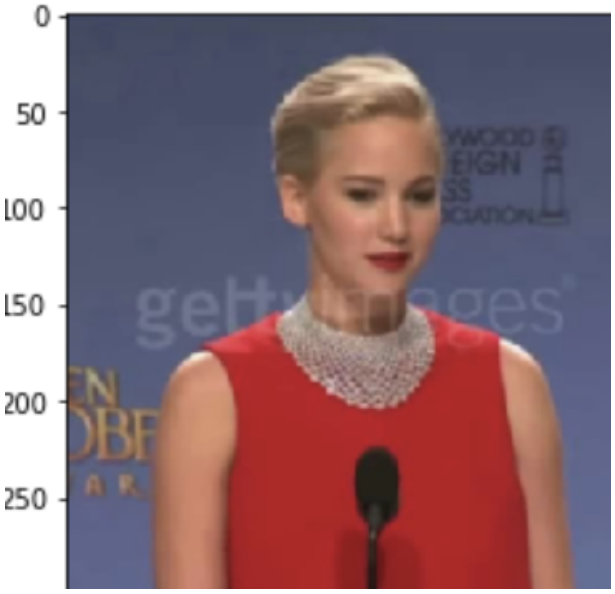

Figure 1. Final Preprocessed Frame

4. Methods

Our final model implemented a CNN LSTM architecture, which uses InceptionV3, a pre-trained CNN for extracting features in images, combined with an LSTM to allow us to use sequence prediction on multiple frames in a video. We froze the InceptionV3 base model to speed up our model, and in turn freeing up computational power for us to use more frames per video. Our output layer contains one neuron, with a sigmoid activation, that gives the binary classification, real or fake, for each of the videos. Additionally,

we fit our model using early stopping to avoid overfitting.

```
Model: "model_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 9, 299, 299, 3)]  0

 time_distributed_3 (TimeDis  (None, 9, 2048)          21802784
 tributed)

 lstm_3 (LSTM)               (None, 128)               1114624

 dense_6 (Dense)             (None, 128)               16512

 dropout_3 (Dropout)         (None, 128)               0

 dense_7 (Dense)             (None, 1)                 129

=================================================================
Total params: 22,934,049
Trainable params: 1,131,265
Non-trainable params: 21,802,784
_________________________________________________________________
```

Figure 2. Model Summary

5. Results

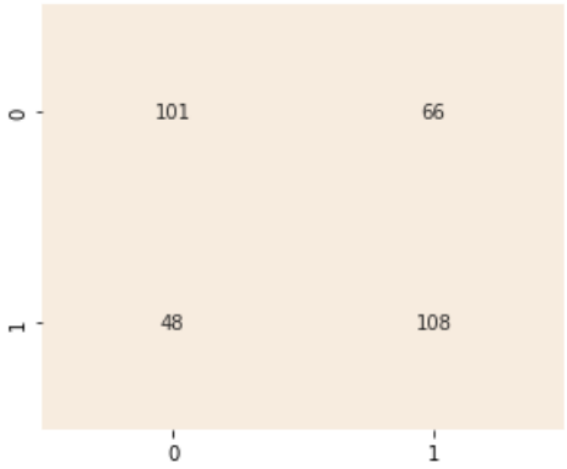

Figure 3. Final Confusion Matrix

Our model obtained an accuracy of 0.6471 and loss of 0.6005. Considering our dataset was split down the middle evenly between real and fake videos, we obtained a good accuracy. Our model was correctly evaluating the videos, not just making random guesses. The low loss value indicates that our model is working and making predictions well. Additionally, the precision was 0.6206, the recall was 0.6923, and the F1 score was 0.6545. In our model, false positives are more detrimental than false negatives, so it is more important to evaluate the precision rather than recall or F1. Since our precision score is above 0.5, this means that the model did well in minimizing the false positive rate, or rate of classifying the video as real when it is actually fake.

6. Discussion

Deepfake detection is becoming increasingly important as disinformation becomes more prevalent in society. As deepfake creation algorithms are becoming increasingly powerful, such doctored videos are more and more challenging to recognize with the human eye. As a result, those who unknowingly consume fake media are more susceptible to believing a false narrative, leading to the further spread of misinformation. Consequently, we consider a false positive, when we classify the video as real when it is in fact fake, as more detrimental than a false negative. Our model did however, have more false positives than negatives and this is something that we would hope to improve in future modifications. Furthermore, if not limited by computing capacity, we believe that we could have improved our accuracy through using the entire videos as opposed to just 9 frames.

References

[1] Tran, V.-N., Lee, S.-H., Le, H.-S., amp; Kwon, K.-R. (2021). High performance deepfake video detection on CNN-based with attention target-specific regions and manual distillation extraction. Applied Sciences, 11(16), 7678. https://doi.org/10.3390/app11167678

[2] Celeb-DF: A large-scale challenging dataset for deepfake ... (n.d.). Retrieved December 12, 2021, from :openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf