Deepfake Detection

Angel Chu, Yanqiao Wang, Zihong Shi
Electrical & Computer Engineering, Carnegie Mellon University

Abstract

In our project, we present a method to detect Deepfakes, existing image or video replaced with someone else's likeness, and identify manipulated media based on depthwise separable convolutions in neural computer vision architectures with residual connections.

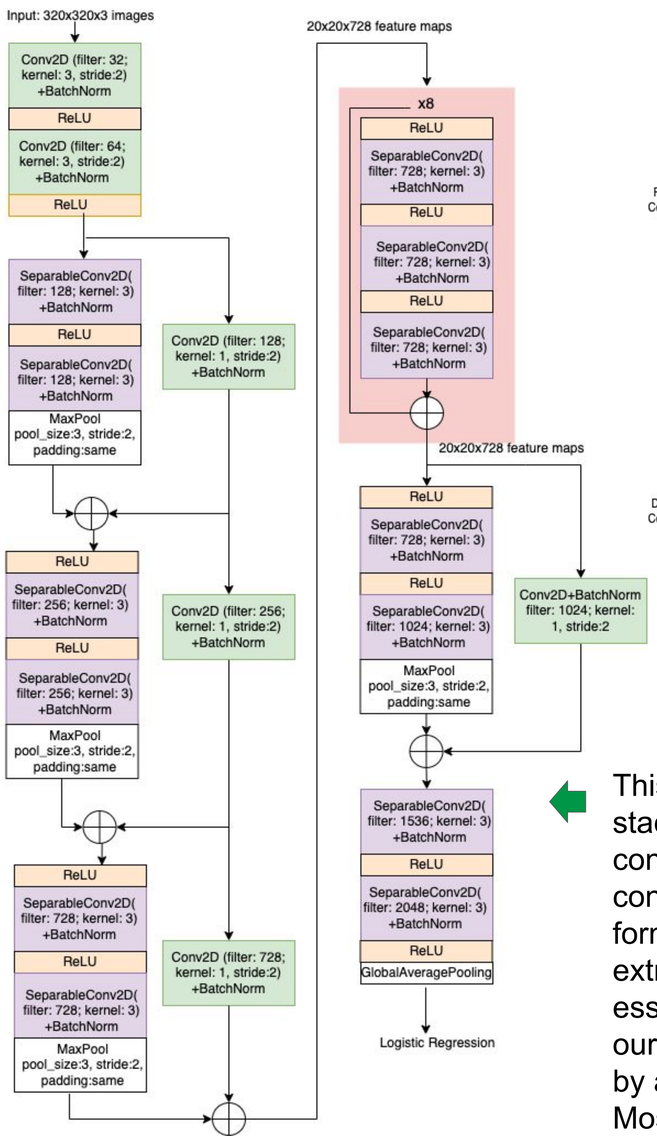
Motivation

With the rapid progress in synthetic image generation and manipulation, digital content that appears in videos and photos can be used to defame persons, maliciously as a source of misinformation and persuasion [1]. Given the negative effects of Deepfake technology, our team aims to implement a method to detect Deepfakes and identify manipulated media within the subset from Kaggle Deepfake Detection Challenge.

References

[1] Advantages and Disadvantages of Deepfake Technology | by Mehmet Emin Masca | Geek Culture | Medium." https://medium.com/geekculture/advantages-and-disadvantages-of-deepfake-technology-ccfa7c12b1ae (accessed Sep. 15, 2022).

Proposed Methods

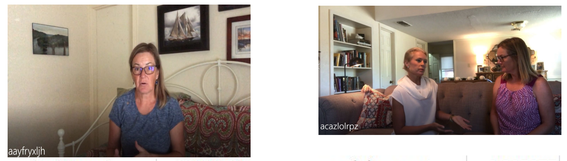


Modified depthwise separable convolution.

This architecture is based on a stack of depthwise separable convolution layers. It overall consists of 36 convolutional layers forming the repeated feature extraction. Since the problem is essentially image classification, our convolutional base is followed by a logistic regression layer. Most modules have linear residual connections around them.

Results

Fig. 1: Real vs. Fake Video w/ Prediction Value



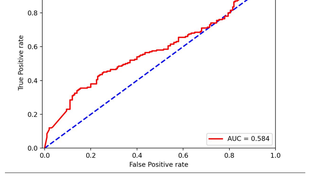
aayfryxljh.mp4 0.756343 Real

acazlolrpz.mp4 0.995077 Fake

The model generates an average prediction value for each test video, and we set a threshold to 0.9. If the predicted value is greater than 0.9, the video is judged to be fake, and if it is less than 0.9, the video is true.

Figure 2 is the ROC curve and AUC value for the model, generated based on 400 test videos. More than half of the videos could be classified correctly.

Fig. 2: ROC Curve for Validation Data



Conclusion & Future Work

- Upon data that even human evaluators have a hard time to distinguish fake videos from real ones, our model performs better than the random chance line.
- With the efficient use of model parameters, our project can help detecting and preventing manipulated media. With linear stack-of-layer architecture, it is easy to modify and be useful for further research.
- Further work could exploit various number of independent channel space segments used for performing spatial convolutions or ensembling with other architecture to seek for higher accuracy rate.