DETECTING DEEPFAKES: LAYERED ARCHITECTURES AND GENERATED ARTIFACTS

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Nathan Williams

April 28th, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Briana B. Morrison, Department of Computer Science

Detecting Deepfakes: Layered Architectures and Generated Artifacts

CS4991 Capstone Report, 2023

Nathan Williams
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
Naw8rc@virginia.edu

ABSTRACT
Financing in the visual effects and film industry has made deepfake technology extremely advanced and accessible in recent years. With these advancements come the risks of deepfaked videos being used by malicious actors, especially for spreading misinformation. One way to mitigate these risks is to have reliable deepfake detection software that uses machine learning to classify whether certain media has been edited or not. At the cutting edge of this technology, researchers are implementing two advanced concepts in order to improve the accuracy of these classifiers. First, recent studies show that layering unique neural network architectures improves accuracy of classifiers depending on the type and order of architectures used. Second, using separate machine learning models to generate training artifacts for the main classifier to analyze also shows promise in improving accuracy. Combining these two concepts could provide accessible and reliable deepfake detection software in the future that would effectively combat the risks of deepfake technology.

1. INTRODUCTION

In 2019, a video of United States Speaker of the House, Nancy Pelosi, stumbling through an interview was shared from unknown origins and spread throughout social media. In this video, Mrs. Pelosi appeared intoxicated and incompetent as she slurred her words and could not focus on the camera. The video was in fact a deepfake, likely intended as a joke, but eventually reposted by President Donald Trump on his Twitter account influencing millions of followers (Mervosh, 2019).

While deepfakes have been a flagship technology for advancements in A.I., there is not nearly as much public interest or support for its twin and counter-technology: deepfake detection classifiers or DDCs. In 2020 Meta along with other tech industry giants started the Deepfake Detection Challenge, inviting experts as well as amateurs to create their own DDCs, offering a prize of a million dollars to the winning team. The highest scoring model only had an accuracy of 65.18%, which while impressive for the task, is not nearly at the point for the model to be acceptable for commercial use (“Deepfake Detection Challenge Results: An open initiative to advance AI”, 2020).

As deepfake technology continues to be improved upon, its potential damage to individuals, organizations, and entire societies only grows. Without a reliable way to separate the fake from real, it is imperative to find new and innovative ways for deepfake detection classifiers to achieve acceptable accuracies.

2. RELATED WORKS

Guera and Delp (2018) were some of the first to publish information on machine learning generated artifacts, specifically

temporal artifacts. My proposal draws heavily on their work since most effective generated artifacts are temporal; however, I recommend the addition that these artifacts are supplemented with layered architectures and transfer-learning.

Another work that influenced this proposal was that of Rana, et al. (2022). This systematic review of deepfake detection literature was incredibly helpful in finding relevant studies, as well as some overarching observations on deepfake detection strategies. While the purpose of Rana, et al.'s (2022) literature review was to compare multiple detection methods such as statistical analysis and blockchain techniques, my focus was on improving deep learning models, since they have been shown to have much higher accuracy than other methods. Also Rana and Sung (2020) showed the efficacy of stacking architectures with their DeepfakeStack. While they achieved high accuracy with only intra-frame data, I propose that incorporating temporal artifacts into stacked networks would lead to even greater accuracy.

3. PROPOSED DESIGN

After researching state-of-the-art deepfake detection classifiers, I found temporal artifacts and stacked neural networks to be some of the most beneficial techniques in improving the model accuracy. My model aims to combine the two without unreasonably increasing the time resources required for training such a model.

3.1 Overview of Model

The proposed model would utilize multiple neural networks consisting of three models that focus on inter-frame image data and 2-3 models that focus on cross-frame data. There would then be a final neural net to determine whether a given video is deepfaked or not (as in Figure 1 below).

3.2 Use of Randomized Weighted Ensemble

The heart of the model is within the randomized weighted ensemble (RWE), which would take the video and run its frame data through different sub-models. It then takes the predictions of each sub-model and weighs them in a final neural net, optimizing the weights based on previous performances. Then it would deliver a final prediction based on the result. Since the whole model is based on the predictions of multiple sub-models, the ability to weigh which models are more effective would not only improve accuracy but help to choose which sub-models to use in the RWE.

3.3 Models and Artifacts

The choice of which sub-models to use is extremely important. First are the three convolutional neural network (CNN) image classifier models. These models are easiest to implement as many pre-trained models already exist online such as ResNet, ImageNet, etc.. The CNN models would be fed the individual frames of a video as pictures and then make predictions based on

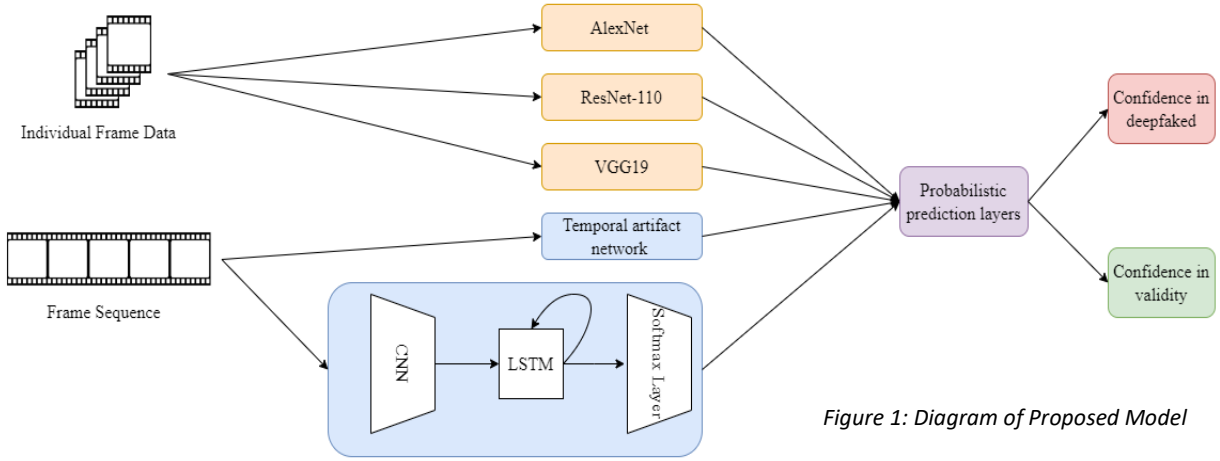

Figure 1: Diagram of Proposed Model

inconsistencies of color and shapes within the frame. Depending on the test data given, different CNN models may be more effective, adding to the benefits of an RWE. The remaining 2-3 models would be comprised of temporal artifact generation models. These models would use Recurrent Neural Networks such as Long Short Term Memory or Hierarchical Memory Networks which excel in recognizing patterns and dependencies over large data sequences (Guera & Delp, 2018). The models would take in multiple frames of a video at once and compare how colors and shapes change over time.

The most common deepfake detection classifiers rely on intra-frame data to detect inconsistencies in the distance between certain facial features and colors within individual frames. However, a model utilizing inter-frame data could detect temporal inconsistencies, such as how often the target blinks or how the edges of the face change over time. While significantly slower to train than static CNN image classifiers, they offer a much more informed prediction.

4. ANTICIPATED RESULTS

When considering the potential results of the proposed model there are a few main concerns. First, will the proposed model improve overall accuracy? Second, will it be applicable to a wide range of deepfake software and videos? Finally, what are the tradeoffs of implementing such a model.

4.1 Effect on Accuracy

Studies on deepfake detection using stacked CNN architectures such as Montserrat, et al. (2020) and Zhang, et al. (2022) were able to achieve a classification test accuracy in the mid-to-high 90^{th} percentile. Compared to a traditional deepfake detection classifier, this is a significant improvement. To add to this, Guera and Delp (2018) were able to break 97% accuracy using temporal training artifacts. By stacking architectures that use temporal artifacts with standard CNNs, I predict that the accuracy of the proposed model could break into the 98-99% accuracy range.

4.2 Generalization

A point of interest when predicting the accuracy of the proposed model is that the experiments considered in making the prediction used frames from the same videos for the training, validation, and test datasets. This means that the model is extremely accurate with predicting deepfakes created with a certain deepfake software, and with specific faces; but there is no data on the accuracy when given a completely new face deepfaked with a different software. However, the usage of an RWE helps to generalize the model by weighing the predictions of each sub-model to allow the best-suited sub-model more influence when changing the subject of the video (Zhang et al., 2022).

4.2 Drawbacks

One issue in implementing the proposed model would be data preprocessing. While the use of multiple sub-models would likely increase accuracy, the use of transfer learning means each sub-model would need to have its own image preprocessing. When dealing with a video comprised of potentially tens of thousands of frames, this image preprocessing could result in significant computational costs and time; however, the time saved by using pre-trained sub-models would almost certainly outweigh the time cost of image pre-processing.

5. CONCLUSION

As deepfake technology continues to advance and become more accessible, it is imperative that new methods for identifying

them are investigated. The state-of-the-art methods of neural network stacking and generating temporal artifacts have been shown to significantly increase detection accuracy and are extremely promising. While research will certainly continue into these techniques, it is also important to explore how the benefits of these methods can be combined to produce even greater accuracy.

6. FUTURE WORK

The goal of this project was to research and analyze state-of-the-art deepfake detection classifiers in order to consider ways to further the development of the technology. As a result, implementing a functional model was beyond the scope of this work. Moving forward with this project, one would need to consider data collection, data preprocessing, and resources required for training.

6.1 Data Collection

The first step in realizing the proposed model would be to start collecting a wide array of both deepfaked and pristine video data. While pristine video data is not hard to come by, as of now the only real source of easily accessible deepfake data comes from the 2020 Deepfake Detection Challenge which has over 100,000 videos. While this is a useful starting point, only a handful of unique deepfake programs were used to make this data set, limiting the generalization of any model trained solely on this data.

6.2 Data Preprocessing

Once a substantial and diverse dataset is acquired for training, the data must then be preprocessed to fit the input layers of each sub model. For the transfer learning sub models, it will be necessary to process the data specifically for that model. However, with the temporal artifact generation sub models the preprocessing would be much more dependent on the designer's choice. The number of frames in a given frame sequence, and the size of the input layer are just some examples of preprocessing that can be tweaked when considering overall accuracy or resource consumption.

6.3 Resource Requirements

In order to create a commercially acceptable deepfake detection classifier, an engineering team would likely need a dataset even larger than the Deepfake Detection Challenge Dataset. Also, with new deepfake software becoming more prevalent, it is necessary to include a wider variety of deepfake technologies in the dataset. If such a dataset were to be collected it would be absolutely massive and training five or more sub-models on this hypothetical dataset simultaneously would require immense graphic processing power. While the models could be trained separately before the probabilistic prediction layer to lower the graphic processing requirements, it would also significantly increase the training time.

REFERENCES

Guera, D., Delp, E. (2018). *Deepfake video detection using recurrent neural networks* [Paper presentation]. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, Auckland, New Zealand.

Mervosh, S. (2019, May 24). Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump. *The New York Times*. https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html

Meta. (2020, June). *Deepfake Detection Challenge results: An open initiative to advance AI*. Meta AI. Retrieved March 23, 2023, https://ai.facebook.com/blog/deepfake-

detection-challenge-results-an-open-initiative-to-advance-ai/

Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., & Delp, E. J. (2020). Deepfakes detection with automatic face weighting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw50498.2020.00342

Rana, S., Nobi, M., Murali, B., & Sung, A. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access, 10.* https://ieeexplore.ieee.org/document/9721302/authors#authors

Rana, S., Sung, A. (2020, August 1-3). *DeepfakeStack: A deep ensemble-based learning technique for deepfake detection* [Paper presentation]. 7th IEEE International Conference on Cyber Security and Cloud Computing, New York, NY, United States.

Zhang, J., Cheng, K., Sovernigo, G., & Lin, X. (2022). A heterogeneous feature ensemble learning based Deepfake Detection Method. *ICC 2022 - IEEE International Conference on Communications*. https://doi.org/10.1109/icc45855.2022.9838630