Truth, Trust, and Transparency in Synthetic Media

John K. Lewis[1], Helen Chen[2], Imad Eddine Toubal[3], Vishal Sandesera[4],
Michael Lomnitz[4], Zigfried Hampel-Arias[4], Prasad Calyam[3], Kannappan Palaniappan[3]

[1]Florida Southern College, [2]University of Maryland–College Park, [3]University of Missouri–Columbia, [4]IQT Labs

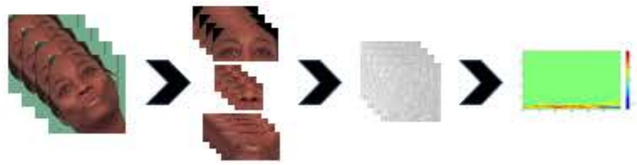
Background

Since the introduction of Generative Adversarial Networks[1] (GANs), synthetic media has become increasingly difficult to identify. Synthetic videos that contain altered faces and/or voices of a person are known as deepfakes, and they threaten the trust and privacy in digital media. Often times, this technology has nefarious intentions, stretching from political intervention to pornography, the latter the most prevalent[2]. The existence of deepfakes also poses a threat to the credibility of authentic videos. As models that create deepfakes improve, the accuracy of human detection decreases[3]. Consequently, it is important to have automated systems that accurately and efficiently classify the validity of digital content. To classify this digital content, we propose a hybrid deep learning model that uses spatial, spectral, and temporal content of an input video.

Methods

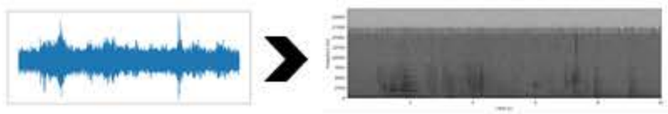
In order for our model to perform efficiently, a series of preprocessing must occur before each video is evaluated. The necessary processes involve both audio processing and image processing and are summarized as follows:

- Image Processing
  - Face Detection
  - Landmark Extraction
  - 1D Discrete Cosine Transform (DCT)

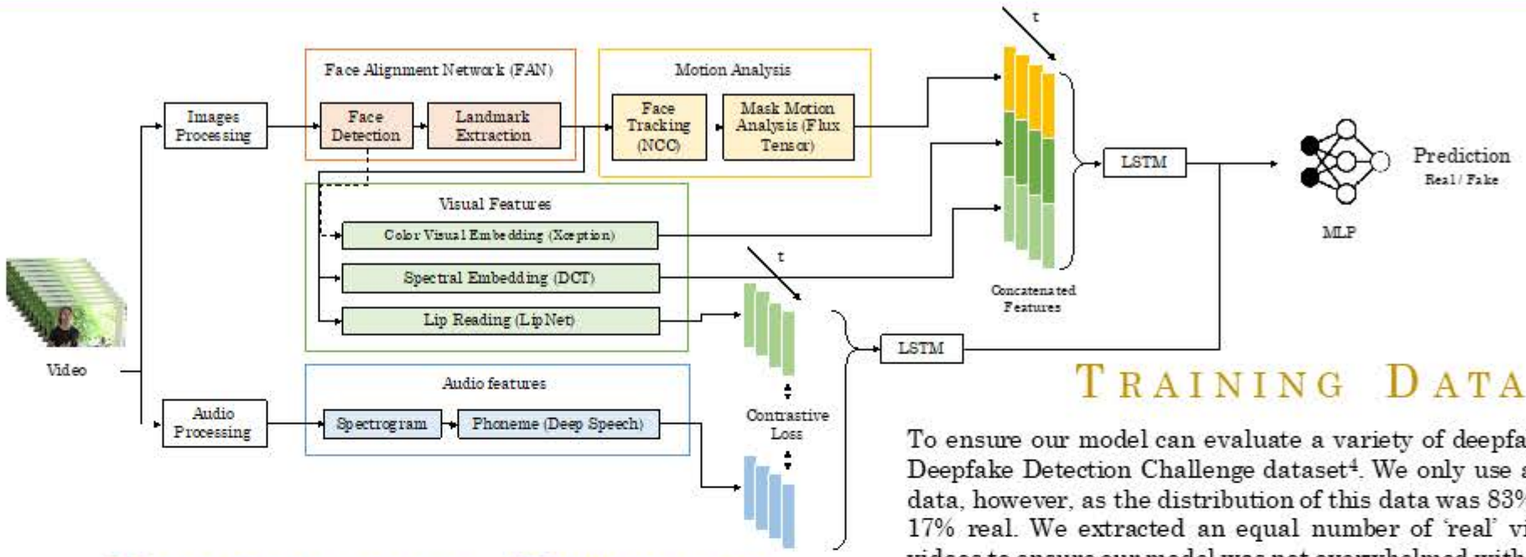


For each video, we extract the faces and save them as JPEG files. We also extract both eyes, the nose, and the mouth, and for each of these landmarks we perform the discrete cosine transform and take the antidiagonal average of the upper left 128x128 pixels and save these in sequential batches.

- Audio Processing (Spectrogram)



For input into the automatic speech recognition model, each video's audio stream must be converted into a spectrogram



Multimodal Network

We worked with a team from IQT Labs that had been working on this project for the last year as part of the Facebook Deepfake Detection Challenge. Our model expanded some of the complexity of their model but resembled many of the core features. For visual features, we first disregard everything in the frame expect the region containing the face. As mentioned in the METHODS section, we perform a series of operations on these faces. We feed each face into XceptionNet, a pretrained Convolutional Neural Network. We extract the feature vectors from this model in the last layer before classification. We concatenate these XceptionNet features and the 1D cosine transform features in batch sizes of 24, which are the input into a Long Short-Term Memory Network (LSTM). Additionally, we feed an affine transformed mouth and feed it through a pretrained lip reading model called LipNet, also in window sizes of 24, from which we extract the feature vectors. These features are combined with the features output from our Automatic Speech Recognition (ASR) model and are sent through a contrastive loss function, the output of which is sent through an LSTM. The outputs of both LSTMs are then concatenated and sent through one final Multilayer Perceptron for a final classification of real of fake. A pipeline of our proposed model is illustrated above.

Electrical Engineering & Computer Science
University of Missouri

Training Data

To ensure our model can evaluate a variety of deepfakes, we use the Deepfake Detection Challenge dataset[4]. We only use a sample of this data, however, as the distribution of this data was 83% Fake and only 17% real. We extracted an equal number of 'real' videos and 'fake' videos to ensure our model was not overwhelmed with any one label.

Conclusion

This research is still a work in progress, and we are continuing to test and evaluate the success of our model. Our model introduces novelty in the following ways:

- Contrastive loss of visual interpretation of speech and ASR
- Spectral Embedding with DCT
- Temporal and Spatial Concatenation
- Audio focused deepfake detection

As new methods are introduced, we hope to continue to expand our model. After research is complete, we hope to build our network to be cloud ready for use in-the-wild.

References

[1] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.
[2] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The State of Deepfakes: Landscape, Threats, and Impact," Sep. 2019.
[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1–11.
[4] "Deepfake Detection Challenge." https://kaggle.com/c/deepfake-detection-challenge.

Acknowledgements

This material is based upon work funded by the National Science Foundation under Award Number: CNS–1950873. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do no necessarily reflect the views of the National Science Foundation