Advanced Machine Learning Project Proposal

Topic: Classification of Deepfake/Real Video Model
Author: Yunting Chiu
Date: Oct.20, 2021

Motivation

Videos may fool people nowadays and they are causing trouble. This is due to a technology called DeepFake. Deepfake is a technique that makes computer-created artificial videos in which images are combined to create new footage. Recently, Deefake technique has been widely discussed in Taiwan since a famous Taiwanese YouTuber was discovered to be responsible for producing, selling, and circulating Deepfake porn videos of women, mostly public figures[1]. Based on this, techniques for solving this kind of problem have been in high demand because more and more relevant issues about misuse of Deepfake technique will be extensively expanded in the future. I have relevant experience in building a machine learning model to determine whether the input video is fake or real. Moreover, I am working on a computer vision project that analyzes deepfake videos as part of my capstone project. As a result, I feel the need to apply my pertinent academic background to this project so that this model can be tested and even improved. The main goal of the project is to assist people in determining whether a given video was generated using the Deepfake technique or a real video by using the model that I have created.

Method

There are diverse elements that are included in a video: frames, images, texts, and audios. The data in my previous experience in building a Deepfake classification machine learning model were silent videos. They are from FaceForensics++[2]. So my previous model is limited in that it can only caprated the frames from those silent videos, missing certain elements mentioned previously. In short, there are no diverse features that I can use in my previous model for prediction. With an intention to improve on the previous model, in this project, I find videos with voices so that I could capture frames using a cv2 package, and collect audios and spectograms using a soundfile package. Then, in the classification model, add all of the features and use a classification in which a real video & audio is labelled as 1 and Deepfake video & audio is labeled 2, 3, 4.

We know that human beings have eyes to watch videos, ears to listen to music, and cognition to determine whether certain information makes sense. That's exactly why I want to add more features in the machine learning model - in order to mimic people's behaviors. The model may find more underlying information in the given features by analyzing frames, texts, voices, and even spectrograms, which can outperform human beings' judgement to tell whether a video is generated using Deepfake techniques or is a real video.

Dataset

FakeAVCeleb[3]: In this dataset, the authors propose a novel Audio-Video Deepfake dataset (FakeAVCeleb) that includes not only Deepfake videos but also respective synthesized lip-synced fake audios. The authors downloaded the original videos from YouTube with different races, including Caucasian, Black, East Asian, and South Asian, to reduce the racial bias in the dataset. There are four types of videos in the data: (1) real audio & real video, (2) real video & fake audio, (3) fake video & real audio, (4) fake video & fake audio.

Intended experiments

Data Preprocessing

1. *Download the vocal videos as data.*
   I download the FakeAVCeleb as a dataset. The strength of FakeAVCeleb is that the videos include different races and texts. Moreover, there are existing text files which are included in the real dataset, but there are no text files in the fake dataset.

2. *Determine how many different types of videos we should categorize.*
   Let's assume R is real and F is Deepfake. The FakeAVCeleb provides 4 types of video: (1) real audio & real video, (2) real video & fake audio, (3) fake video & real audio, (4) fake video & fake audio.

3. *X_1 feature - images*
   I put the frames (images) in X_1. To be more specific, I randomly select 10 frames from a video using the cv2 package that captures different facial expressions in the same video. Then, in all selected images, I use an MTCNN facial detection model[4] to extract face features. Finally, I resize these facial images to 320 x 320 x 3 and flatten each pixel into one-dimensional arrays.

4. *X_2 feature - audio*
   I put the audio in X_2. I use a soundfile package to read each audio as an array. Then, I flatten audio data from two-dimensional arrays to one-dimensional arrays.

5. *X_3 feature - spectrums*
   I put the spectrograms in X_3. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Spectrograms can be used to analyze the frequency content of a waveform to distinguish different types of vibration. It is because the computer might detect the difference between R or F videos through a spectrogram domain. That's why I want to add this feature to the model. For sure, the final shapes of spectrograms are 1D arrays.

6. *X_4 feature - texts (optional)*
   The downside of FakeAVCeleb is that I could only find text files in the "real video & real audio" subset, meaning that the other three types of videos (RF, FR, FF) lack text files that can be directly analyzed to the model. Despite the fact that speech_recognition is a good package for automatically generating texts from a given video, we cannot be certain that the texts are 100 % correct. Plus, whether the video is Deepfake or real, both can display 100% correct grammar, so that texts may not be a good feature to classify the input video as real or fake. If I have time, I try to vectorize the texts by using the count

vectorization tool to build the matrix and include them as a feature 4. But I still need to prioritize the project's deadline, so I won't move forward with the NLP (natural language processing) part right away.

7. *Y - label the observations*

   Straightforwardly, this is supervised learning so I label each observation.
   - Labeled 1: real video & real audio (RR)
   - Labeled 2: real video & fake audio (RF)
   - Labeled 3: fake video & real audio (FR)
   - Labeled 4: fake video & fake video (FF)

Data Modeling

I consider using SVM and random forest to build the classification model and evaluate its performance metrics such as accuracy, precision, recall, and F1 score. If the model does not perform well, there are many remediations we can take to improve model performance, such as normalizing or decomposing the model before refitting it.

References

[1] Hioe, B. (2021, October 21). *Arrest calls attention to issue of "Deepfake" videos in Taiwan. New Bloom Magazine.* Retrieved October 24, 2021, from https://newbloommag.net/2021/10/19/deepfake-arrest/

[2] Khalid, H., Tariq, S., & Woo, S. S. (2021). FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. *arXiv preprint arXiv:2108.05080.* https://arxiv.org/abs/2108.05080

[3] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1-11). https://arxiv.org/abs/1901.08971

[4] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503. https://arxiv.org/abs/1604.02878