AI Gone Rogue: Exterminating Deep Fakes Before They Cause Menace

Vijay Thaware Niranjan Agnihotri

Symantec Software India Private Ltd.

Abstract

The face: A crucial aspect of identity. But what if this crucial part of one's identity is stolen from an individual? Yes, this now happens and is termed "Deep Fake." Deep fake technology is an artificial intelligence (AI)-based human image blending method used in different ways, such as when creating revenge pornography and fake celebrity pornographic videos, or even in cyber propaganda. Videos are altered using Generative Adversarial Networks (GANs) in which the face of the speaker is manipulated by a network by swapping it with someone else's face. These videos can sometimes be identified as fake by the human eye; however, as neural networks get rigorously trained on more resources, it will become difficult for the human eye to identify fake videos. Such videos can cause chaos and bring economic and emotional damages to one's reputation. Videos targeted against politicians in the form of cyber propaganda can prove to be catastrophic to a country's government. In the bigger picture, it has the potential to undermine democracies.

In this study, we have discussed the many tentacles of Deep Fake and the extent of damage it can cause. Additionally, we share our progress on our proposed solution: to identify complex Deep fake videos using deep learning. We used a pre-trained Facenet model. The model was trained on the image data of people of importance or concern. After training, the output of the final layer was stored in a database. A set of sampled images from a video was passed through the same model and the output of the final layer from the neural network was compared with values stored in the database. A simple Euclidean distance measure between the encodings can recognize the actor in the video, which can in turn help us in finding out if the video is fake.

In 2018, we believe that Deep Fakes have already progressed to a different level. We also discuss other defensive measures against Deep Fakes.

Introduction

Each individual has a unique identity that is in large part defined by his or her face. The face is not only a distinct, defining feature on a person's head but also a crucial aspect of one's identity.

As we see most commonly among children, people love to make faces in the mirror and imitate facial expressions as seen in magazines, newspapers, billboards, the TV, tablet, smartphone, or virtually any device that projects images. This most natural of acts is at best playful and harmless—but what if this human act were taught to a machine? This is not unexpected in the 21st century, in which Artificial Intelligence (AI) has so progressed to the extent that, when trained properly with sufficient sources, it can generate or create fake and deceptive videos, which could fool the human eye and trick people into believing they are seeing what the creator of the fake videos wants them to see.

These fake videos are becoming exceedingly sensible, convincing, and hard-to-recognize portrayals of genuine individuals doing or saying things they never actually stated or performed. These types of videos are called "Deep Fakes". "Deep," as taken from "deep learning," was strung with "fake" to form a portmanteau. As the technology advances, the capacity to create false yet believable videos and audio targeting celebrities, politicians, and governments in general will certainly increase. The implications of Deep Fakes can be detrimental to society or individual reputations and identities. Before we delve into the hazards, let's discuss the background of Deep Fakes.

Deep Fake: What makes it so different?

This is not the first time fabricated videos or audios were generated to create content. This practice is believed to be common in international film industries like Hollywood. One of the finest modern examples of this is the "appearance" of 1970s-vintage Peter Cushing and Carrie Fisher in the 2016 film *Rogue One: A Star Wars Story.*

In the past, editing or fabricating videos was an expensive task which required an extensive amount of manpower, time, and money. However, today all it takes is a gaming laptop, an internet connection, and a rudimentary knowledge of neutral networks. To ease this process, there are existing applications such as FakeApp which offer face-swapping in videos with a single click. Deep Fake videos have been democratized such that anyone with zero or little knowledge could generate such videos. The next generation of Deep Fakes are not only

hyperrealistic but also highly impossible to detect through the human eye. With the advancements in deep learning, incredibly realistic fake videos can be generated.

Deep Fakes: The many-faced evil

Imagine a scenario where a Deep Fake video is created to show a politico speaking disrespectful and obscene statements just a few hours before voting starts. Such a video can be easily circulated over social media, which could potentially alter the outcome of the election. This is of high concern as it can instigate tensions not only within the country but also can have ripple effects globally.

Politicians, pundits, the media, independent news outfits, bloggers, and the public have all accused one another over so-called fake news. Deep Fake news springing from fake news is not too far a possibility.

Deep Fakes are rapidly progressing, and this can lead to trust issues. Even genuine video content online can invite suspicion and questions as to its validity. Politicians can flat-out deny or cast doubt on genuine videos that may be embarrassing to themselves, claiming that the video is fake, edited, or completely fabricated.

Deep Fakes circulating on social media could fuel disinformation campaigns, which could lead to severe violence and social unrest. Similar incidents have already taken place where people have lost their lives.

Revenge-porn Deep Fake videos can scar victims for life. Such videos can result in emotional distress, trauma, depression, and even ruin the victim’s reputation and social standing.

Deep Fakes became popular when fabricated porn videos of celebrities started making the rounds online. This obviously violates not only the rules of consent but the victims’ very right to privacy. Creating Deep Fakes using applications without a person’s consent is a form of abuse.

Curbing the mess

Pornographic Deep Fakes can traumatize an individual mentally and ruin them socially. Now that Deep Fake technology has been democratized, creating such videos will only proliferate. As such, there must be laws and/or legal protections introduced for individuals who have been victimized by such videos.

As well, technologies designed to block Deep Fake content on social media need to be developed. These can be in the form of filters or a first line of detection in identifying potential harmful fake videos. Imagine a hypothetical scenario in which a large volume of politically themed Deep Fake videos are uploaded and circulated on social media platforms. How well is the human eye and average user equipped to identify, flag, and report such videos? How much time and effort will it take to block or remove Deep Fakes? Social media platforms will need to be proactive in preventing, disallowing, and removing Deep Fakes. To a limited extent, social media platforms such as Twitter, Reddit, and Facebook have already taken steps to curb Deep Fakes by blocking or taking down user accounts or posts that spread fake content .

Our Solution

Our proposed solution can identify face tampering among face-swapped videos by means of a simple single-sided test. We have the following assumptions for the sake of simplicity of implementation and to compensate for limited resources.

Assumptions:

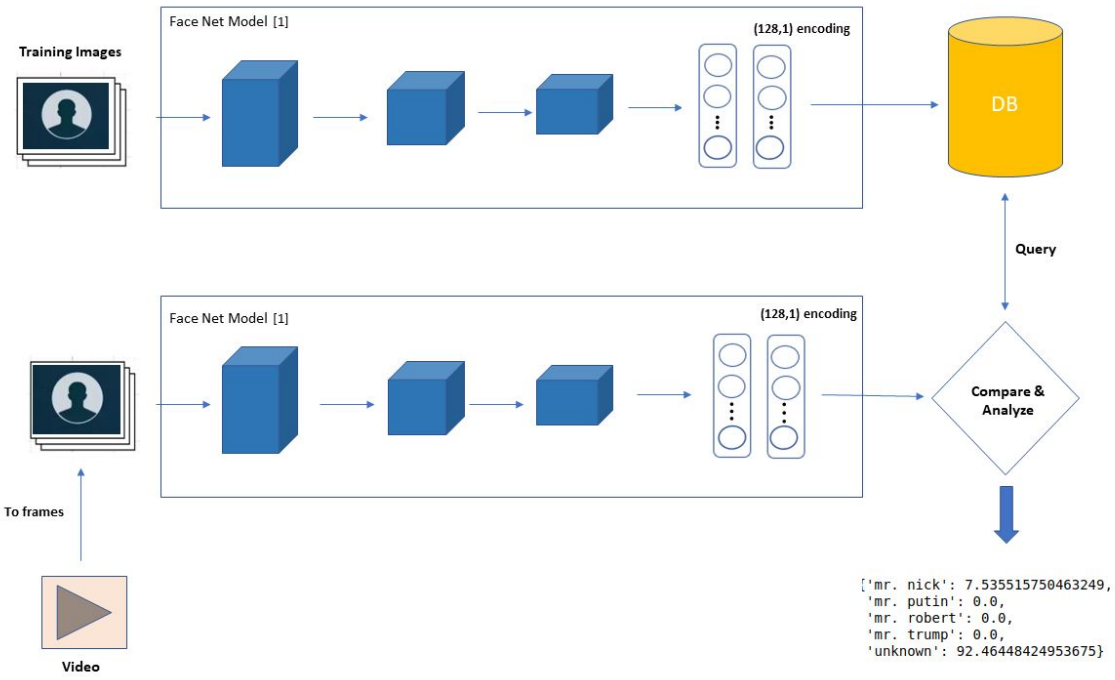
1. The video given at the input is a portrait video: a portrait video is a video which has only one actor.
2. We have all the necessary face encodings stored into our database: for the demo purpose, we worked on face encodings of only four people. (Those whose fake videos are available online.) We represent rest of the class as “unknown,” abbreviated as “unk.”

Data:

For our demo, we used 26 Deep Fake videos that were freely available on the internet. Additionally, we collected 30 genuine videos of the same actors (who appeared in the Deep Fake videos) that were freely available on the internet.

Design:

-



Design of the proposed solution

Methodology:

About Face Net[1]: FaceNet[1] is one of the outstanding neural network architectures designed by researchers at Google for face verification and recognition. The network can be trained on the faces of people and can in turn be used for their face verification and recognition.

The model has a stunning 99.63% of accuracy on the Labelled Faces in the Wild [LFW] dataset. The model uses an optimization function known as triplet loss[1], which minimizes the distance between the anchor and the positive image, and maximizes the distance between the anchor and the negative image. Here, the anchor and the positive image contain the same person, while the negative image contains a different person. The FaceNet model can be trained on the images of people and can be used to verify their identity later. The model produces 128-D embedding, which is representative of the encoding of the image.

- **About our design:**
    - We used the open-source implementation of the FaceNet[1] available on the internet. Firstly, we train the model on the genuine images of people. As a quick and dirty implementation, we trained the model on the genuine images of four people. The network generates 128-D encodings for each person's face. We store this encoding along with the person's name into a database. We call the person's name as label for simplicity.
    - In the next phase of implementation, we split a video into frames using ffmpeg software[2] at the rate of 27 frames per second. We process the video on a frame by frame basis. Each frame is fed as the input to the trained model above.
    - The model generates a 128-D encoding for each frame. We find the simple Euclidean distance between this encoding and all the encodings stored in the database. The label of the encoding pertaining to minimum Euclidean distance s assigned to the frame.
    - After processing all the frames from the video, we get a probability score for all the labels from our database. For the simplicity of our implementation, we consider a video genuine if we get more than 50% of score for a particular label, i.e. person. Going back to our assumptions, if a video gets less than 50% score corresponding to any actor and simultaneously gets more than 50% score for the label "unk," we say that the video might be fake.

Results:

In the observations, during our experimentation we found that our model was good at confirming genuine videos. Out of 30 genuine videos in our dataset, it was able to correctly confirm the authenticity for all 30 videos by providing high confidence to one of our labels. On the other hand, the model also does well on fake videos produced by the face-swapping technique. From our data set, we were able to correctly classify 18[*] fake videos out of 25. This translates to a detection rate of 72%.

Scope for improvement:

There is no one solution that fits all scenes when it comes to Deep Fakes. Therefore, our technique is not foolproof. We are working on the following pointers to develop more effective techniques to identify Deep Fake videos:

- We believe that there could be huge scope for improving our current solution by augmenting our technique with the following mechanisms
    - **Face liveliness detection** – state-of-the art mechanisms have been developed to improve the performance of face detection and recognition task in real time. We believe that this technique if modified appropriately to our problem domain, could give us more insights about the genuineness of the video.
    - **Texture investigation** – To determine patterns in spoofed and real images, we could exploit the fact that tiny texture parts of input images are probed.
    - **User interaction** – Telling the User to perform actions such as head rotation, smiling, blinking so that the machine can detect whether the action is performed in a natural way which resembles human interaction.
    - **Contextual intel mechanisms** – Advanced research in the field of Visual Scene Understanding can help us by giving insights about the surroundings. This could be helpful in deciding the authenticity of the video.

- Our current solution is effective for Deep Fake videos which are created using the face-swapping technique. Modern advancements in face reenactment technologies have the capability of creating fake videos without doing a face swap. They can directly transfer the expressions from a source video to a target video without tampering the face of the target actor. We are working on developing a more effective approach to tackle this new advancement.

- Currently, almost all the Deep Fake videos that are available in the public domain contain a single actor in the video. Rapid advancements in deep learning and computer vision techniques can create Deep Fake videos containing multiple actors. Our current approach works on videos with only one actor: portrait videos.

- As a quick implementation, we have worked on a small data set. But, we believe that our solution is scalable with minor modifications. Deep neural networks like FaceNet are already deployed at scale by social network giants such as Facebook. We believe that same networks could be used to confirm the authenticity of videos on social media.

Mitigating Deep Fakes:

Researchers have used the average human eye blinking rate to detect Deep Fake videos. In the paper In Ictu Oculi [3], Siwei Lyu, *et al* found that in the Deep Fake videos, the actors exhibit an abnormal eye blinking behavior. They used a Long term Recurrent Network (LRCN) model to study the eye blink behavior of the actor in the video based on which they confirmed the authenticity of the video. This technique has shown impressive performance.

- As mentioned earlier, there is no single solution that fits all the cases of Deep Fake videos. With sufficient amount of data and other resources, it is possible to create Deep Fake videos with a natural eye-blinking behavior.

- This technique would also be less effective for Deep Fake videos of short duration, as there the eye blink rates vary highly from person to person within short time spans (<1 minute).

- The technique's effectiveness is also limited to the quality of the video and the distance of the person from the camera. If the person stands far enough from the camera or wears shades, then it would be difficult to detect the eye blinks.

Video watermarking can be implemented to curb the overall widespread of Deep Fakes.

Checking the credibility of the source while accessing information on social media and using personal discretion is the best method to cope up with Deep Fakes.

References

In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking

https://gizmodo.com/deepfake-videos-are-getting-impossibly-good-1826759848

https://www.theverge.com/2018/2/9/16986602/deepfakes-banned-reddit-ai-faceswap-porn

https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

https://motherboard.vice.com/en_us/article/xwvz9a/watch-an-algorithm-turn-winter-into-summer-in-any-video-image-to-image-translation

https://motherboard.vice.com/en_us/article/a3kmpb/facial-recognition-for-porn-stars-is-a-privacy-nightmare-waiting-to-happen

https://motherboard.vice.com/en_us/article/jpgkxp/after-20-minutes-of-listening-new-adobe-tool-can-make-you-say-anything

https://www.hollywoodreporter.com/behind-screen/how-furious-7-brought-late-845763

https://www.cmu.edu/news/stories/archives/2018/september/deep-fakes-video-content.html

https://web.stanford.edu/~zollhoef/papers/SG2018_DeepVideo/page.html

https://web.stanford.edu/~zollhoef/papers/SG2018_DeepVideo/paper.pdf

https://gizmodo.com/insanely-accurate-lip-synching-tech-could-turn-fake-new-1796843610#_ga=2.94590111.22112046.1528730509-3674648492.1521482270

https://www.cs.cmu.edu/news/beyond-deep-fakes

https://www.bloomberg.com/news/articles/2018-09-10/how-faking-videos-became-easy-and-why-that-s-so-scary-quicktake?utm_campaign=socialflow-organic&utm_source=facebook&utm_content=markets&utm_medium=social&cmpid=socialflow-facebook-markets

https://medium.com/mit-technology-review/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery-75256531f667

https://www.theverge.com/2018/2/8/16990076/ai-fakes-deepfakes-video-media

https://www.youtube.com/watch?v=7XchCsYtYMQ

https://www.defenseone.com/threats/2018/10/how-deep-fakes-threaten-democracies/152093/?oref=d-river

https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin

https://www.youtube.com/watch?v=o2DDU4g0PRo

https://www.youtube.com/watch?v=dMF2i3A9Lzw

http://www.niemanlab.org/2018/07/the-universe-of-people-trying-to-deceive-journalists-keeps-expanding-and-newsrooms-arent-ready/

https://www.cfr.org/report/deep-fake-disinformation-steroids?utm_medium=social_earned&utm_term=deep-fake-disinformation-steroids&utm_source=li&utm_content=101718

https://www.economist.com/leaders/2018/05/24/a-faked-video-of-donald-trump-points-to-a-worrying-future

https://www.apnews.com/21fa207a1254401197fd1e0d7ecd14cb

https://arxiv.org/pdf/1805.11714.pdf

https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/

https://www.seeker.com/artificial-intelligence/this-new-ai-system-can-create-convincing-deep-fake-videos

https://www.youtube.com/watch?v=8LhI-e2B8Lg

https://lyrebird.ai

https://www.alanzucconi.com/2018/03/14/an-introduction-to-autoencoders/

https://medium.freecodecamp.org/making-your-own-face-recognition-system-29a8e728107c

https://medium.com/swlh/anti-spoofing-mechanisms-in-face-recognition-based-on-dnn-586011ccc416