Deepfakes: How a pervert shook the world

Ronit Chawla
Student, Modern School, New Delhi, Delhi

ABSTRACT

Recently a machine learning based open source software (i.e. a free to use the software) tool has made it easy to create hyper-realistic face swaps in videos that leave little to no traces of manipulation, in what is known as “deepfake” videos. Scenarios, where these AI manipulated/generated videos, are used for political distress, blackmail or even terrorism are easily envisioned as a near dystopia. This paper explores the various aspects of deepfake videos including its consequences and newly developed innovations in detecting deepfakes.

Keywords — Deepfakes, Deep, Fakes, Machine Learning, Artificial Intelligence, Obama Fake, Deepfake Videos, Face Swap, Face app, Fake app

1. HISTORY OF DEEPFAKE

The first known attempt at face swapping can be found in one of the iconic portraits of U.S. President Abraham Lincoln, circa 1865. The lithography in the picture mixes Lincoln’s head with the body of John Calhoun. Recent advances have radicalised visual-data manipulation. Modern tools such as Tensorflow with the accessibility of the technical literature and access to computing infrastructure have endorsed this paradigm shift. Autoencoders have made tampering images and videos, which used to be reserved to highly-trained professionals, an extremely accessible task, within the reach of any individual with a computer, foul-intentioned or otherwise. Open source software like FaceApp and FakeApp are built upon this phenomenon. Modern tools like FakeApp, FaceApp and other softwares have made it easy for anyone to produce “deepfakes”, such as the one swapping the face of late-night TV hosts Jimmy Fallon with that of John Oliver. FakeApp is a desktop application (and now an open source) that allows the user to create what is now known as deepfakes; these are A.I. manipulated videoclips which were first created by anonymous Redditior, Deepfake, he (or she, as deepfake’s identity remains unknown) used TensorFlow, social media, image search engines and videos to insert someone else’s face onto preexisting videos frame by frame! The realistic appearance of deepfakes also makes them a target for the production of fake scandals, fake news, fake surveillance videos, and malicious hoaxes. These fake videos have already been used to create political tensions and they are being taken into account by governmental entities. As presented in the Malicious AI report, researchers in artificial intelligence should always reflect on the duality of uses of their work, allowing misuse considerations to influence research priorities and norms.

2. DEEPFAKE VIDEOS EXPOSED

Due to the way that FakeApp generates the manipulated deepfake video, intra-frame inconsistencies and temporal inconsistencies between frames are created. These flaws are used to check whether a video is a manipulation.

3. HOW ARE DEEPFAKES MADE?

It is well known that deep learning techniques have been successfully used to enhance the performance of image compression. Especially, autoencoders have been applied for dimensionality reduction, compact representations of images, and generative models learning. Thus, autoencoders are able to extract more compressed representations of images with a minimized loss function and are expected to achieve better compression performance than existing image compression standards. The compressed representations or latent vectors that current convolutional autoencoders learn are the first cornerstones behind the faceswapping capabilities of. The second insight is the use of two sets of encoder-decoders with shared weights for the encoder networks. What makes deepfakes possible is finding a way to force both latent faces to be encoded on the same features. This is solved by having two networks sharing the same encoder, yet using two different decoders (top). When we want to do a new faceswapp, we encode the input face and decode it using the target face decoder (bottom).

Training Two sets of training images are required. The first set only has samples of the original face that will be replaced, which can be extracted from the target video that will be manipulated. This first set of images can be further extended with images from other sources for more realistic results. The second set of images contains the desired face that will be swapped in the target video. To ease the training process of the autoencoders, the easiest face swap would have both the original face and target face under similar viewing and illumination conditions.
However, this is usually not the case. Various viewing angles, lightning conditions or different video codecs makes it difficult for autoencoders to produce realistic faces under all conditions. This usually leads to swapped faces that are visually inconsistent with the rest of the scene. This frame-level scene inconsistency will be the first feature that we will exploit with our approach. It is also important to note that if we train two autoencoders separately, they will be incompatible with each other. If two autoencoders are trained separately on different sets of faces, their latent spaces and representations will be different. This means that each decoder is only able to decode a single kind of latent representations which it has learnt during the training phase. This can be overcome by forcing the two sets of autoencoders to share the weights for the encoder networks, yet using two different decoders. In this fashion, during the training phase, these two networks are treated separately and each decoder is only trained with faces from one of the subjects. However, all latent faces are produced by the same encoder which forces the encoder itself to identify common features in both faces. This can be easily accomplished due to the natural set of shared traits of all human faces (e.g. number and position of eyes, nose etc.) Video Generation When the training process is complete, we can pass a latent representation of a face generated from the original subject present in the video to the decoder network trained on faces of the subject we want to insert in the video. As shown in the Figure, the decoder will try to reconstruct a face from the new subject, from the information relative to the original subject face present in the video. It is important to point out that for doing this frame-level operation, first a face detector is used to extract only the face region that will be passed to the trained autoencoder. This is usually a second source of scene inconsistency between the swapped face and the rest of the scene. Because the encoder is not aware of the skin or other scene information it is very common to have boundary effects due to a seamed fusion between the new face and the rest of the frame. The third major weakness that we exploit is inherent to the generation process of the final video itself. Because the autoencoder is used frame-by-frame, it is completely unaware of any previously generated face that it may have created. This lack of temporal awareness is the source of multiple anomalies. The most prominent is an inconsistent choice of illuminants between scenes with frames, with leads to a flickering phenomenon in the face region common to the majority of fake videos. The phenomenon of incorrect color constancy in CNN-generated videos is a well known and still open research problem in the computer vision field. Hence, it is not surprising that an auto encoder trained with very constrained data fails to render illuminants correctly.

The open computer vision or opencv library will allow us to easily recognised faces in images with just a function call. But the way it's doing this in the background is by using a method called histogram of oriented gradients or hog for short it starts by making our image black and white to simplify it. We don't need color data to find faces. Then it looks at every single pixel in the image on one at a time. For every pixel, it looks at the pixels directly surrounding it. The goal is to figure out how dark the current pixel is compared to the pixel directly surrounding it. Then, it draws an arrow showing in which direction the image is getting darker.
Once it does this process for every single pixel in the image, we'll end up with every pixel being replaced by an arrow. These arrows show the flow from light to dark across the entire image. If we analyse the pixels directly, then really dark images and really light images of the same person would have totally different pixel values. But because we're only considering the direction that brightness changes, both types of images end up with the same representation. Making the problem easier to solve. Saving all these directions is too space intensive. So we break the image into smaller squares and count how many different directions there are. Then replace each square with the direction that has the most count this turns the original image into a very simple representation that captures the basic structure of a face.

This creates a HOG face pattern and when another similar HOG face pattern is compared and the patterns are alike by some threshold value then a face is considered detected. Repeating this multiple times by using multiple image search engines we can crop out just the face and have an exclusive face data set for both are characters.

An autoencoder is something that tries to reconstruct the inputted image what's it into lower level dimensional representation. Deepfakes uses one encoder and two decoders. During trianing, it actually trains two networks both the networks to share and encoder but have different decoders. An encoder transforms an image into a base “vector” this is a set of numbers which indetifies the important features of a person's face the Decoder transforms that vector back to an image. There is an error function which measures how good the transformation was and lowers the overall error during training. The first network is only trained on image A and second, only on image B. Encoder learn how to convert an image into a face vector decoder B learn how to convert a base factor into the image a and decoder B learn how to convert a base vector into image B.
So during training, we are feeding both images to the same encoder but using two different decoders for each. After the network is done training we can feed it a video. Which is simply a collection of images as frames. By repeating this process for every frame of a video, and thereafter concatenating it, we get a “deep faked video”

4. DETECTING DEEPFAKES
There are still defects in new algorithms for “deepdetection”. One of them has to do with how well the recreated faces – or don’t. People blink every 2 to 10 seconds, and a single blink takes between one-tenth and four-tenths of a second. That is what would be found in a suspected video. In any case, it’s not what is seen in the majority of deepfakes.

At the point when a deepfake algorithm is prepared on face images, it's subject to the photographs that are accessible on the web that can be utilized in preparing data. Notwithstanding for individuals who are captured frequently, few images are accessible internet demonstrating their eyes shut.

Without preparing images of the individual blinking, deepfake algorithms are more averse to make faces that flicker regularly. When we compute the general rate of blinking and contrasts that and the normal range, we find that characters in deepfake recordings flicker significantly less. This exploration uses AI itself to analyze recordings.

This gives us hope to someday identify deepfake recordings. To be increasingly explicit, it checks each frame of a video being analysed, recognizes the faces in it and afterward finds the eyes. It, at that point, uses another profound neural system to decide whether the recognized eye is open or close, utilizing the eyes' appearance and highlighted areas on the face (i.e. HOG mechanisms).

This isn't the bottom line on identifying deepfakes, obviously. The innovation is improving quickly, and the challenge among producing and recognizing counterfeit recordings is analogous to a chess game: blinking can be added to deepfakes by incorporating face images with shut eyes or utilizing video successions for preparing it. Deepfakers will keep showing signs of improvement at making deepfakes – and people in the innovation network should keep on discovering approaches to recognize them.

5. RESTRICTIONS
While the outcomes are fascinating, there are clear restrictions on what we can be accomplished by this

- The decoding mechanism only works if there are surplus amounts of pictures of the objective: to put an individual into a video, on an average, about 300–2000 pictures of their face are required so the system can figure out how to reproduce it. The number relies upon how the person's face is oriented, and on numerous external factors: lighting, facial feature, age etc.
- This works fine for celebrities or anybody with bunches of their photographs on the web. In any case, this won't let you manufacture an item that can swap anyone's face.
- You additionally need preparing data that is illustrative of the goal: the algorithm isn't good at producing profile shots of Oliver, basically on the grounds that it didn't have numerous instances of Oliver looking to the side. When all is said and done, preparing pictures of your objective need to estimate the direction, outward appearances, and lighting in the recordings you need to stick them into.

So in case one is fabricating a face swap for an individual, given that the most photographs of them will be forward looking (e.g., selfies on Instagram), he must limit face swaps to for the most part front aligned recordings. In case he is working with a celeb it's simpler to get a different arrangement of pictures. What's more, if your objective (the person whose deepfake is to be created ) is helping him make data, he must include a variety of pictures so he can embed them into anything.

6. POTENTIAL APPLICATIONS
6.1 Video Content Production:
Hollywood now has this innovation readily available. This it could open up new avenues: for example, making motion pictures with obscure on-screen characters, and afterward superimposing celebrities onto their faces. Thus creating a deepfake. This could work for YouTube recordings or even news channels.

In increasingly out-there situations, studios could change entertainers depending on their objective market (more Indian actors in Hollywood movies for the Indians), or Netflix could enable watchers to pick on-screen actors before hitting play. Almost certainly, this technology could produce income for long dead on-screen characters by breathing life into them back.
6.2 Authorizing Celebrity Faces

Envision, if Target could have a big name grandstand their garments for a month, just by paying her specialist an expense, buying some current headshots, and clicking a catch. This would make another income stream for famous people, web based life influencers and other "famous faces”. What's more, it would give organizations another instrument to advance brands and drive revenue. It likewise brings up legal issues about responsibility for, and plan of action on ways to parcel and value rights to utilize them.

6.3 Customized Advertising

Imagine an existence where the advertisements you see as you surf the web incorporate you, your companions, and your family. While this may appear to be shocking today, is it so fantastical to imagine this won't be the standard in a couple of years?

All things considered, we being visual beings, and companies have been attempting to inspire reactions from us for quite a long time, for example, Coke might put your friends' faces in a hip music video, or Allstate may pull at your feelings of trepidation by demonstrating your family in a protection advertisement. Or on the other hand, the methodology might be more straightforward: the banana republic could superimpose your face on a body type that matches yours, and persuade you that it suits their new calfskin coats.

7. CONCLUSION

- Whoever Deepfakes (the aforementioned reddit user) is, he/she has opened the Pandora's box regarding how phony videos will influence society. It is , however, logical to assume that akin to how we have come to acknowledge that pictures can be photo shopped, we will adjust to video vulnerability as well, however not every person shares this expectation.
- What Deepfakes did is simply shine a light on how fascinating deepfakes really are. Profound generative models like the autoencoder that Deepfakes utilizes, enable us to make engineered yet reasonable looking data. This implies once these algorithms are transformed into commercial commodities (which is presumably inevitable ), normal people will acquire inventiveness, ideally towards constructive ends.
- There have been some fascinating utilizations of this strategy, similar to style move applications that make your photographs look like acclaimed artistic creations, however, given the high volume and fascinating nature of the exploration that is being distributed in this space, there's obviously much more to come. All we can say is that somehow, somewhere a pervert, a really smart pervert, has changed the course of humanity

8. REFERENCES

[9] Schwartz, Oscar (12 November 2018). “You thought fake news was bad? Deep fakes are where the truth goes to die”. The Guardian.