DEEP FAKES
One man's tool is another man's weapon

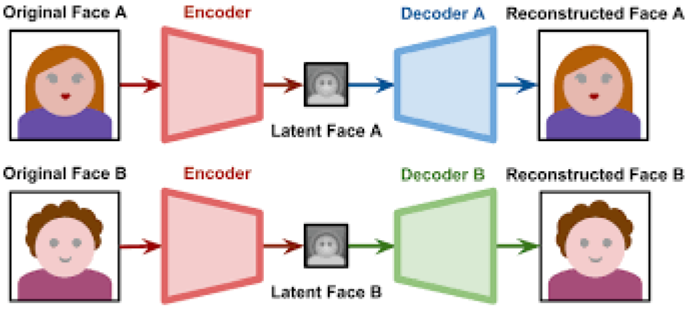
Priyansu Dey
2nd Year Computer Science Department

Abstract

We all have heard about machine learning and artificial intelligence by now. Deep learning, being a greater and true depth of what a machine can do has been a boon or a curse for us. On one end, it helps to solve numerous complex problems such as big data analytics and on the other it is becoming the most hyped up, memed and an upcoming threat to the security of the famed as well as for the common people. Its algorithms create fake images and videos which normal people cannot differentiate from the genuine one.

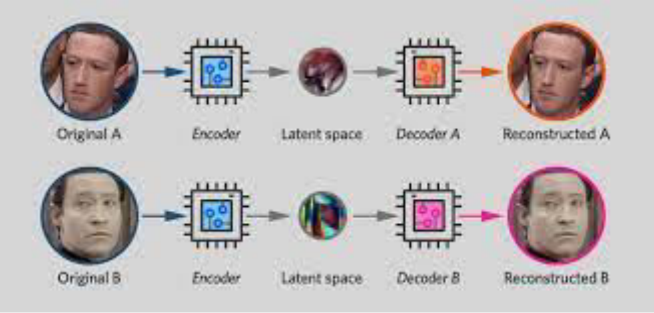
Introduction

In a short definition, deep fakes (originating from “deep learning” and “fake”) are created by certain methods and algorithms which superimpose face images of a certain person onto a video of a source person to make a video of the target person doing or saying things the source person does. This includes the most popular product of deep fake which is face swap. In a more enhanced definition, deep fakes are artificial intelligence based content which can be subdivided into two cases, that is lip-syncing and puppet-mastery. Lip-syncing deep fakes refer to videos that are modified to make the mouth movements consistent with the help of an audio recording. Puppet-mastery deep fakes include videos of a target person also a puppet who is animated following the expressions of face, eye and head movements of another person who is the master sitting in front of a camera.



Is it safe or harmful?

While some deepfakes are often created by traditional visual effects or computer-graphics approaches, the recent common underlying mechanism for deepfake creation is deep learning models like autoencoders and generative adversarial networks, which are applied widely within the computer vision domain . These models are wont to examine facial expressions and movements of an individual and synthesize facial images of another person making analogous expressions and movements . Deepfake methods normally require a large amount of image and video data to coach models to create photo-realistic images and videos. As public figures like celebrities and politicians may have a large number of videos and pictures available online, they are initial targets of deepfakes. Deepfakes were wont to swap faces of celebrities or politicians to bodies in porn images and videos. A primary deepfake video emerged in 2017 where the face of a star was swapped to the face of a porn actor. It's threatening to world security when deepfake methods are often employed to make videos of world leaders with fake speeches for falsification purposes. Deepfakes therefore are often abused to cause political or religious tensions between countries, to fool the public and affect election campaigns, or create chaos in financial markets by creating fake news. They can even lead to generating fake satellite images of the world to contain objects that don't really exist to confuse military analysts, e.g., creating a fake bridge across a river although there's no such a bridge in reality. This will mislead a troop who is guided to cross the bridge during a battle .As the democratization of making realistic digital humans has positive implications, there's also positive use of deepfakes like their applications in visual effects, digital avatars, snapchat filters, creating voices of those who have lost theirs or updating episodes of flicks without reshooting them . However, the amount of malicious uses of deepfakes largely dominates that of the positive ones. The event of advanced deep neural networks and therefore the availability of huge amounts of data have made the cast images and videos almost indistinguishable to humans and even to stylish computer algorithms. The method of making those manipulated images and videos is additionally much simpler today as it needs as little as an identity photo or a brief video of a target individual. Less and fewer effort is required to produce stunningly convincing tempered footage. Recent advances can even create a deepfake with just a still image . Deepfakes therefore are often a threat affecting not only public figures but also ordinary people. For example, a voice deepfake was wont to scam a CEO out of $243,000 . A recent release of a software called DeepNude shows more disturbing threats because it can transform an individual to non-consensual porn . Likewise, the Chinese app Zao has gone viral lately as less-skilled users can swap their faces onto bodies of movie stars and insert themselves into well-known movies and television clips. These sorts of falsification create an enormous threat to violation of privacy and identity, and affect many aspects of human lives. Finding reality in the digital domain therefore has become increasingly critical. it's even tougher when handling deepfakes as they're majorly used to serve malicious purposes and almost anyone can create deepfakes lately using existing deepfake tools. Thus far, there are numerous methods proposed to detect deepfakes . Most of them support deep learning, and thus a battle between malicious and positive uses of deep learning methods has been arising. In order to deal with the threat of face-swapping technology or deepfakes, the United States Defense Advanced Research Projects Agency (DARPA) initiated a research scheme in media forensics (named Media Forensics or MediFor) to accelerate the use of fake digital visual media detection methods . Recently, Facebook Inc. teaming up with Microsoft Corp and the Partnership on AI coalition have launched the Deepfake Detection Challenge to catalyse more research and development in detecting and preventing deepfakes from getting used to mislead viewers.

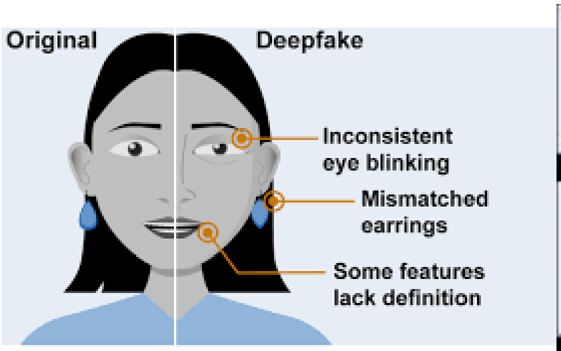


Summary of Notable Deepfake Tools

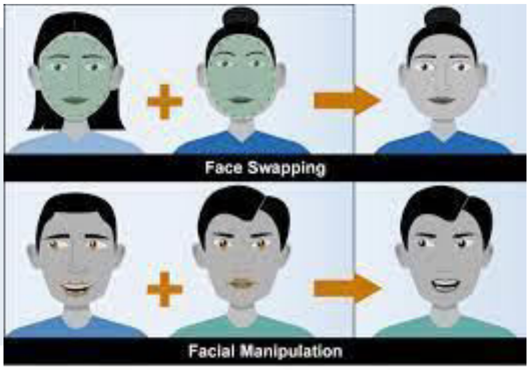
TOOLS	**KEY FEATURES**
Faceswap	Using two encoder-decoder pairs.
Faceswap-GAN	Adversarial loss and perceptual loss (VGGface) are

	added to an auto-encoder architecture.
Few-Shot Face Translation	Use a pre-trained face recognition model to extract latent embeddings for GAN processing
DeepFaceLab	Expand from the Faceswap method with new models
DFaker	DSSIM loss function is used to reconstruct face
DeepFake tf	Similar to DFaker but implemented based on tensorflow
AvatarMe	Reconstruct 3D faces from arbitrary “in-the-wild” images.
MarioNETte	A few-shot face reenactment framework that preserves the target identity
DiscoFaceGAN	Generate face images of virtual people with independent latent variables of identity, expression, pose, and illumination.
StyleRig	Create portrait images of faces with a rig-like control over a pretrained and fixed StyleGAN via 3D morphable face models
FaceShifter	Face swapping in high-fidelity by exploiting and integrating the target attributes
FSGAN	A face swapping and reenactment model that can be applied to pairs of faces without requiring training on those faces.
Neural Voice Puppetry	A method for audio-driven facial video synthesis.

FakeApp, developed by a Reddit user using autoencoderdecoder pairing structure, in which the method, the autoencoder extracts latent features of face images and the decoder is employed to reconstruct the face images. To swap faces between source images and target images, there is a requirement of two encoder-decoder pairs where each pair is employed to coach on a picture set, and the encoder’s parameters are shared between two network pairs. In other words, two pairs have an equivalent encoder network. This strategy enables the common encoder to find and learn the similarity between two sets of face images, which are relatively unchallenging because faces normally have similar features like eyes, nose, mouth positions.



Source: GAO; conceived from DARPA image at https://www.darpa.mil/news-events/2019-09-03a. | GAO-20-379SP



Source: GAO | GAO-20-379SP

How to detect Deep Fakes

Deepfakes are increasingly detrimental to privacy, society security and democracy. Methods for detecting deepfakes are proposed as soon as this threat was introduced. Early attempts were supported handcrafted features obtained from artifacts and inconsistencies of the fake video synthesis process. Recent methods, on the other hand, applied deep learning to automatically extract salient and discriminative features to detect deepfakes . Deepfake detection is generally deemed a binary classification problem where classifiers are wont to classify between authentic videos and tampered ones. This type of method requires an outsized database of real and faux videos to coach classification models. The amount of faux videos is increasingly available, but it's still limited in terms of setting a benchmark for validating various detection methods.

Some measures to detect Deep Fakes

Methods	**Key Features**
Eye blinking	Use LRCN to learn the temporal patterns of eye blinking.
Intra-frame and temporal inconsistencies	CNN is employed to extract frame-level features, which are distributed to LSTM to construct sequence descriptor useful for classification.

Using face warping artifacts	Artifacts are discovered using CNN models based on resolution inconsistency between the warped face area and the surrounding context
MesoNet	Two deep networks, i.e. Meso-4 and MesoInception-4 are introduced to examine deepfake videos at the mesoscopic analysis level.
Capsuleforensics	- Latent features extracted by VGG-19 network are fed into the capsule network for classification.
Head poses	Features are extracted using 68 landmarks of the face region. - Use SVM to classify using the extracted features.
Emotion audiovisual affective cues	Modality and emotion embedding vectors for the face and speech are extracted for deepfake detection.
Using appearance and behaviour	Temporal, behavioral biometric based on facial expressions and head movements are learned using ResNet-101 while static facial biometric is obtained using VGG
FakeCatcher	Extract biological signals in portrait videos and use them as an implicit descriptor of authenticity because they are not spatially and temporally well-preserved in deepfakes
Pairwise learning	Two-phase procedure: feature extraction using CFFN based on the Siamese network architecture and classification using CNN.

Conclusion

Deep fakes have begun to erode trust of individuals in media contents as seeing them is not any longer commensurate with believing in them. they might cause distress and negative effects to those targeted, heighten disinformation and hate speech, and even could stimulate political tension, inflame the general public, violence or war. This is especially critical nowadays because the technologies for creating deep fakes are increasingly approachable and social media platforms can spread those fake contents quickly. Sometimes deepfakes don't have to be spread to a massive audience to cause detrimental effects. People who create deepfakes with malicious purposes only have to deliver them to focus on audiences as a part of their sabotage strategy without using social media. For example, this approach is often utilized by intelligence services trying to influence decisions made by important people like politicians, resulting in national and international security threats. Catching the deep fake alarming problem, the research community has focused on developing deep fake detection algorithms and various results are reported.

References

1) Lyu, S. (2018, August 29). Detecting ‘deepfake’ videos in the blink of an eye. Available at http://theconversation.com/detecting-deepfake-videos-inthe-blink-of-an-eye-101072
2) Bloomberg (2018, September 11). How faking videos became easy and why that’s so scary. Available at https://fortune.com/2018/09/11/deep-fakes-obama-video/
3) The Guardian (2019, September 2). Chinese deepfake app Zao sparks privacy row after going viral. Available at https://www.theguardian.com/technology/2019/sep/02/chineseface-swap-app-zao-triggers-privacy-fears-viral
4) Lyu, S. (2020, July). Deepfake detection: current challenges and next steps. In IEEE International Conference on Multimedia and Expo Workshops (ICMEW) (pp. 1-6). IEEE.
5) Faceswap: Deepfakes software for all. Available at https://github.com/deepfakes/faceswap