House Committee on Science, Space, and Technology
Subcommittee on Investigations and Oversight
Online Imposters and Disinformation, 09/26/2019

Siwei Lyu, Ph.D.
University at Albany, State University of New York

Backgrounds

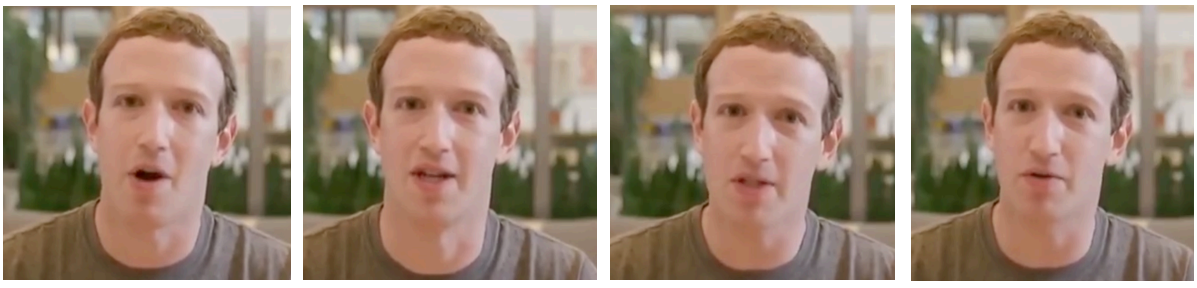
Deepfakes are the most recent twist to the disconcerting problem of online disinformation. The term *deepfake* first emerged in late 2017 as the name of a Reddit account that began posting synthetic pornographic videos generated using an AI-based face-swapping algorithm. The term has subsequently become synonymous with three types of AI-generated impersonation videos.

(1) Head puppetry entails synthesizing a video of a target person's head using a video of a source person's head, so the synthesized target appears to behave the same way as the source.

source (real) **target (fake)** **source (real)** **target (fake)**

(2) Face swapping involves generating a video of the target with the faces replaced by synthesized faces of the source while keeping the same facial expressions.

(3) Lip synching is to create a falsified video by only manipulating the lip region so that the target appears to speak something that s/he does not speak in reality.



[Source: Bill Posters and Daniel Howe, *The Spectre Project*]

Photos and videos have been doctored since their nascence. But there are three reasons why the current concerns over deepfakes and other AI-driven audio and visual media manipulations are justified. First, deepfakes can be made more easily, quickly, and with better quality — thanks to the rapid advancement of computer hardware and software technology, in particular those related to AI. Second, the capability to make deepfakes has been democratized through software tools that can be downloaded freely from online code sharing platforms.[1] Third, anyone with an online media presence is a potential target of a deepfake attack. A fake video showing a politician engaged in an inappropriate activity may be enough to sway an election if released close to voting day. A fake video of a falsified recording of a high-level executive commenting on his/her company's financial situation could potentially send the stock market awry. A fake video made by falsely implanting a woman's face in a pornographic video and shared on social-media platforms could tremendously traumatize the victim. The stakes are too high to ignore.

How are deepfakes made?

Deepfakes are created with a type of AI technology commonly known as deep neural networks.[2] A deep neural network model learns to synthesize realistic faces through *training*, which involves exposing the model to a large number of face images of different people with varying expression, head poses, and lighting conditions. Once the model is properly trained, it is ready to be used to generate deepfakes. Specifically, a face detection method is first run on the input video to locate the target's faces. Then facial landmarks corresponding to distinct locations such as the tips of eyes, eyebrows, nose, mouth, and contour of the face are extracted. Using these landmarks, the detected faces are warped to the same size and in a standard configuration. The standardized faces are fed to the deep neural network model to synthesize a new set of faces of different identity, which are then warped back to match the target's head orientations in the input.

Current computer hardware and AI technology has made it much easier to create deepfakes: a computer that is used to run the generation algorithm with a special computing hardware known as graphical computing unit (GPU) can be readily purchased for an affordable price on Amazon.[3] The training videos for the targets can be easily downloaded from social-media platforms such as YouTube, Instagram, and Facebook in large volume and high quality. Convenient software tools have made the whole process automated barring the choice of a few parameters. As a result, a few good-quality, minute-long videos, a commodity computer with a GPU, and several hours of training are sufficient to generate deepfakes with decent visual quality.

---

[1] e.g., FakeAPP (used to be on Reddit but now defunct), DeepfaceLab (https://github.com/iperov/DeepFaceLab), faceswap-GAN (https://github.com/shaoanlu/faceswap-GA), faceswap (https://github.com/deepfakes/faceswa), and more recently ZAO (https://apkproz.com/app/zao).

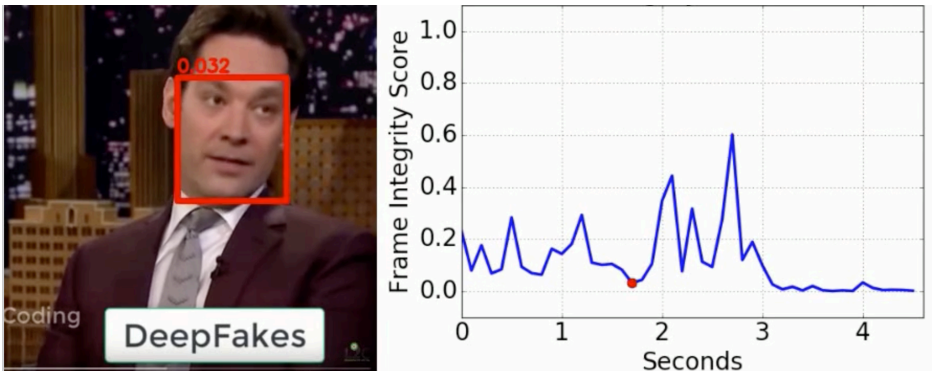
[2] GANs are only one type of deep neural network model for creating deepfakes. For example, the deepfake created for today's hearing did not use a GAN, but instead a different model known as the variational auto-encoders. This is important because any legislative or rule making effort to address deepfakes should not rely on a single tool. Instead Congress should attempt to future-proof regulation to cover the act instead of the tool.

[3] An example of computer configuration for this purpose includes an HP-Z800 workstation (~$1,000) equipped with an Nvidia 2080Ti GPU (~$1,200) and other necessary peripherals. Cost effective and large-scale production can also be conducted using cloud platforms such as Amazon AWS or Google Cloud Platform.

How to combat deepfakes with technology development

While sophisticated deepfakes still take time and skill to produce, rough-and-ready fake videos may still cause harm. It is thus important to have effective technologies to identify, contain, and obstruct deepfakes before they can inflict damage. This should be done by focusing on improving forensic capabilities and making it harder to train deepfakes using online videos.

Effective deepfake forensic detection methods look for traces of the synthesis process to differentiate deepfakes from real videos. For instance, synthesized faces are warped and processed to fit the target's head orientation, such operations leave traces that can be exploited to detect deepfakes. Another type of detection techniques involves examining physiological inconsistencies such as the lack of realistic eye blinking and heart beating. A third approach is to "use AI to fight AI", using another deep neural networks to detect deepfakes. State-of-the-art detection methods have shown promising accuracy on benchmark datasets, but their actual performance on real life deepfakes have yet to be tested.[4] Also, due to the complex nature of deepfakes, no single type of technology or specific method will be the *silver bullet*, and an effective solution may come as a combination of all these approaches.



Detection results of a state-of-the-art deepfake detection method over a fake video on youtube.com. The lower integrity score (range in [0,1]) suggests a video frame more likely to be generated using deepfake algorithms.

In addition to deepfake forensics, there are also technologies to prevent the re-use of online images and videos as training data for the deep neural network generating deepfakes. This would involve inserting imperceptible "adversarial noise" into images and videos before they are uploaded to online social-media platforms. The adversarial noise correspond to subtle perturbations that human eyes cannot see nonetheless can disrupt a face detection algorithm and make it difficult to automate the training process. A dedicated adversary could overcome adversarial noise by painstakingly selecting the target's face in every frame of a training video, but that requires 1,500 hand-marked selections for each 60 second training video.[5]

---

[4] One notable effort towards this goal is the upcoming *Deepfake Detection Challenge* (https://deepfakedetectionchallenge.ai) sponsored by Facebook, Microsoft and Partnership on AI, to advance the state-of-the-art deepfake detection capacities.

[5] This is calculated based on a target video quality of 25 frames per second, which is the lowest frame rate for uploaded YouTube videos. YouTube videos. The highest quality YouTube videos are uploaded at 60 frames per second, which would more than double the number of hand-marked selections for a 60 second video and the work to hand select faces.

Perspectives

As the underlying technology continues to develop, the current barriers to making deepfakes will fall and their quality will keep improving. What is also evolving is the quintessential cat-and-mouse game experienced by all attacker-defender relationships, and malicious attackers seem to have an upper-hand — they can adjust the generation algorithm whenever a new detection method is made public. Currently, the majority of research on combating deepfakes is sponsored under DARPA.[6] But it is important that the federal government also fund more civilian research through NSF. One reason this has not yet happened is because the grant-making capabilities of NSF are focused around existing directorates that are not well equipped to support research into cross-functional emerging technologies. It may be wise to fund the establishment of a new *Emerging Technologies Directorate* at NSF, which can function as a catchall until either an existing directorate's mission is expanded or a new directorate is created. This would create a research home not just for deepfake forensics but also other emerging technologies.

The open-source model of disseminating research code is an enabling factor of the current deepfake problem and requires more scrutiny. The availability of easy-to-use and easy-to-access software tools has significantly lowered the technical threshold for an ordinary user to create deepfakes. A nation state with more manpower and computing resources can build upon them refined and customized versions to make more crafted deepfakes with higher level of realism and use them in a disinformation campaign. It may thus be wise to consider requiring NSF to conduct an ethics review of proposed grants around dual-use technology like deepfakes with mandatory controls on the release of the underlying technology into the proverbial wild.

Last but not least, education on responsible research should be an intrinsic part to the current AI research. Deepfakes add just one more item to the long list of various ethical issues of AI algorithms, such as built-in biases and prejudice, violations of individual privacy and safety, and the lack of accountability and transparency. As academic or industrial researchers working in these areas, we should recognize the potential impact of our research on society, and take them seriously as part of our due responsibilities. We should also provide training to students and post-docs on such issues. These practices could, again, be enforced through requirements from NSF on funded AI research that make such training and compliance mandatory.

Conclusions

It is not an exaggeration to say that we are on the cusp of deepfakes being cheap, easy to produce, indistinguishable from real videos, and ready to cause real damages. We therefore need a comprehensive and robust solution to this problem. The situation calls for continuous investment and perhaps an escalated funding level from the federal government to this strategically important research area. The situation surrounding deepfakes may not turn out to be as severe as we are predicting now. But it is better safe than sorry.

---

[6] Most notably, the DARPA Media Forensics (MediFor) program (https://www.darpa.mil/program/media-forensics).