Engineering 6 (2020) 723–724

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

Media Enhanced by Artificial Intelligence: Can We Believe Anything Anymore?

Ramin Skibba

Senior Technology Writer

While messages, letters, or photographs could be altered to change both content and intent, people generally assumed they were not because it was not easily done. But this is no longer true in today's digital world. Tools for altering photos and other media quickly became available with the advent of computers, the internet, and, most recently, smartphones and social media. Now, artificial intelligence (AI) is further transforming digital media, with far more sophisticated programs that can be used to almost seamlessly manipulate video, photographs, audio, and text for a variety of purposes.

"Manipulating photographs is as old as photography itself," said Siwei Lyu, professor of computer science and director of the Computer Vision and Machine Learning Laboratory at the State University of New York at Albany. "The recent twist is the application of AI, where you can scale up the manipulation. It used to be something that required a lot of time, effort, special training and equipment." With a powerful computer and enough knowledge to run the algorithms, Lyu said, manipulating video can now be done on a much greater scale.

Computer engineers are also working to perfect AI systems for "natural language processing" that can generate text and speech that closely approximates human language. For example, in early 2019 the San Francisco-based research laboratory OpenAI announced they had developed a state-of-the-art text generator called GPT-2, which could write coherent sentences in English and even short stories and poetry with just a few prompts. The researchers initially avoided releasing the full model because they feared the software was good enough that it could be used for malicious purposes such as for generating "fake news" [1]. But they relented in November 2019 after seeing "no strong evidence of misuse" [2]. Nonetheless, in this and other media, the old proverb "seeing is believing" appears itself to be becoming fake news.

Software like Photoshop for modifying photographs has existed for some time (Fig. 1), and now video sequences can be manipulated with similar ease. The most common manipulations, known as "deepfakes," typically involve swapping the face of a person (the target) with that of someone else (the donor). Another type of deepfake, a "lip-sync," involves modifying the source video so that the mouth movements of a speaker are changed to be consistent with a different audio recording. Done well, the resulting video looks realistic to the viewer, with the effect of them seeming to say something they actually have not. Such deceptive videos could be—and have been—used to manipulate public opinion, commit fraud, and wrongly discredit people [3].

In practice, deepfake generation depends on feeding data—a significant number of images or text—into machine learning tools known as generative adversarial networks (GANs). In the simplest version, two such neural networks are trained to develop and improve a model for turning input data into new images or video. Early algorithms were trained with massive datasets, derived from

Fig. 1. Adobe Photoshop software was used to create this fanciful but realistic-looking landscape from 16 different photographs. Software powered by AI algorithms now provides tools to create realistic but manipulated and/or simulated video, text, and speech with perhaps even greater ease. Credit: Wikimedia Commons (CC BY-SA 3.0).

https://doi.org/10.1016/j.eng.2020.05.011

easily accessible images of people like politicians and celebrities. While the process used to require some degree of supervision by programmers, the latest programs are almost entirely automated.

"You do not need huge amounts of training data. Even just a ten-second clip can be enough," said Subbarao Kambhampati, professor of computer science and engineering at Arizona State University in Tempe and an expert on human-aware AI. But training the model with longer clips, with source video that has at least 1000 high-quality frames, will generate a better final product. For each frame in a video, current algorithms can map out "landmarks" on a person's head, as well as their head pose, eye gaze, in addition to the shape of more detailed features, including the eyebrows, eye blinks, eye lids, upper and lower lips, cheeks, chin, and dimples [4].

The motion in the resulting video can look fluid, like human vision expects. But if not done carefully, the output video can contain "tells" that could make a perceptive viewer suspect it has been modified. "Sometimes you can see odd things, like a stretching or distortion of facial features, which do not match exactly," said Doug Goodwin, Fletcher Jones Scholar in computation and a visiting professor in media studies at Scripps College in Claremont, CA, USA. For example, if the training data have insufficient resolution, the output video might have blurry areas, with white strips in the mouth rather than individual teeth, or facial hair that does not move the way it should. The algorithms work better when they are trained with a diversity of facial expressions and words, Goodwin said.

The advances in manipulation have prompted computer scientists and engineers to develop AI algorithms—forensic software—for detecting altered video and audio [5]. "Forensic tools can detect synthesized media and tell whether it is generated by a machine or a human. But if these tools are not kept secret, media can always be crafted to bypass them," said Paarth Neekhara, a doctoral student in computer science at the University of California, San Diego, who studies audio and video deepfakes.

The back and forth between manipulation and detection resembles the computer security arms race of viruses and antivirus software, wherein fixes stop hackers and hackers find ways to overcome the fixes [6]. Experts find a flaw, allowing them to spot manipulated media, and then the makers adapt to generate even more realistic deepfakes. For example, when first generation deepfakes showed faces that did not blink periodically, making them easy to detect, the next generation of deepfake software fixed that problem. In another example, a video of then US President Barack Obama was manipulated to make it look like he said something he had not, but his eyebrow movement did not match his lip movement, Kambhampati said. But then in subsequent deepfakes his eyebrows moved as expected. Because AI can be trained to detect and fix such discrepancies, the latest generation deepfakes have very few faults.

Many negative applications have arisen [3,7], but there are many positive applications as well, and these have also driven advances in the technology. Examples include improving a video or audio recording of someone with a speech impediment, adding more realistic dubbing of language in movies, and even recreating a character in a movie played by an actor who has died, such as Princess Leia, played by the late Carrie Fisher, in the Star Wars movie Rogue One [8]. The application to virtual reality for games or other entertainment seems especially promising and likely [9].

As exemplified by the OpenAI software mentioned above, computer scientists are also using AI programs to generate credible text and speech [1]. Like modified video, this modeling also now uses GANs to produce realistic sentences [10]. Google Translate, for example, now runs on such AI algorithms [11]. The algorithms are sophisticated enough to generate text in the style of a particular person, to produce, for example, a new story seemingly written by the long-deceased author Jane Austen [12]. Programmers have created chatbots, such as on social media platforms, that read and sound real enough that potential customers interact with them as if they were actual people. And in perhaps the most widely used commercial applications of AI-powered communication, Amazon's Alexa and Apple's Siri cloud-based voice services are programmed to mimic real conversations with customers. Alexa and Siri may not be real people, but they do seem to give truthful answers to questions.

To date, programmers have made more headway with realistic video and still images, Goodwin said. But if current trends continue, he said, it may soon be possible to set up AI algorithms to write and digitally create completely new and credible speech and then meld it with simulated audio and video, in a mostly automated process. This prospect, and its potential use for deception, has prompted researchers to develop code to automatically spot deepfakes, and calls for social media sites to identify such media as manipulated [13]. In December 2020, Facebook launched a Deepfake Detection Challenge in a collaboration with Microsoft, Amazon, and academic computer scientists including Lyu, to recruit researchers to submit their own automated detection tools with a chance to win 1 million USD in prizes [14]. Engineers at the US Defense Advanced Research Projects Agency are also working on tools to automatically determine whether a video or photo has been manipulated [15].

References

[1] Schwartz O. For centuries, people dreamed of a machine that could produce language. Then OpenAI made one [Internet]. New York: IEEE Spectrum; 2019 Dec 2 [cited 2020 Apr 18]. Available from: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/for-centuries-people-dreamed-of-a-machine-that-can-produce-language-then-openai-made-one.

[2] OpenAI. GPT-2: 1.5B release [Internet]. OpenAI; 2019 Nov 5 [cited 2020 Apr 18]. Available from: https://openai.com/blog/gpt-2-1-5b-release/.

[3] Verdoliva L. Media forensics and Deepfakes: an overview. 2020. arXiv:2001.06564.

[4] Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting world leaders against deep fakes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 38–45.

[5] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for Deepfake forensics. 2020. arXiv:1909.12962.

[6] Chesney R, Citron DK. Deep fakes: a looming challenge for privacy, democracy, and national security [Internet]. Berkeley: California Law Review; 2019 Dec 17 [cited 2020 Apr 18]. Available from: https://doi.org/10.2139/ssrn.3213954.

[7] Graham J. It's not just phishing emails, now we have to worry about fake calls, too [Internet]. Tysons Corner: USA Today; 2020 Feb 27 [cited 2020 Apr 18]. Available from: https://www.usatoday.com/story/tech/2020/02/27/phishing-deepfake-audio-scams-increasing-fake-calls/4876171002/.

[8] Winick E. How acting as Carrie Fisher's puppet made a career for Rogue One's Princess Leia [Internet]. Cambridge: MIT Technology Review; 2018 Oct 16 [cited 2020 Apr 18]. Available from: https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia/.

[9] National Academies of Sciences, Engineering, and Medicine. Implications of artificial intelligence for cybersecurity: proceedings of a workshop. Washington, DC: The National Academies Press; 2019.

[10] Wang K, Wan X. Automatic generation of sentimental texts via mixture adversarial networks. Artif Intell 2019;275:540–58.

[11] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. arXiv:1609.08144.

[12] Poole S. The rise of robot authors: is the writing on the wall for human novelists? [Internet]. London: The Guardian; 2019 Mar 25 [cited 2020 Apr 18]. Available from: https://www.theguardian.com/books/2019/mar/25/the-rise-of-robot-authors-is-the-writing-on-the-wall-for-human-novelists.

[13] Eggerton J. Hill calls for social media standards from Facebook, Reddit, others on combatting deepfakes [Internet]. Bath: Multichannel News; 2019 Oct 2 [cited 2020 Apr 24]. Available from: https://www.multichannel.com/news/hill-calls-social-media-standards-facebook-reddit-others-combating-deepfakes.

[14] Deepfake detection challenge [Internet]. Menlo Park: Facebook; c2019 [cited 2020 Apr 24]. Available from: https://deepfakedetectionchallenge.ai/.

[15] Turek M. Media Forensics (MediFor) [Internet]. Arlington: DARPA; c2016 [cited 2020 Apr 24]. Available from: https://www.darpa.mil/program/media-forensics.