Engineering 6 (2020) 1339–1340

Contents lists available at ScienceDirect

Engineering

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

News & Highlights

Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge

Ramin Skibba

Senior Technology Writer

The improving power of artificial intelligence (AI) is perhaps most evident in the increasingly realistic manipulation of video and other digital media [1], with the latest generation of AI-altered videos, known as deepfakes [2], prompting a primarily Facebook-sponsored competition to identify them as such. Launched in December 2019, the Deepfake Detection Challenge (DFDC) closed to entries in March 2020 [3]. The results are now in Refs. [3–5]. While somewhat unimpressive, underscoring the difficulty of addressing this growing challenge, they importantly provide a benchmark for automated detection strategies and suggest productive directions for further research.

With little to no help from a human's guiding hand, the advanced computer algorithms used to create today's deepfakes can readily produce manipulated videos and text that are becoming ever more difficult to distinguish from the real thing [1,6,7]. While such technology has many positive applications, computer scientists and digital civil liberties advocates have grown increasingly concerned about its use to inadvertently or deliberately mislead viewers and spread disinformation and misinformation [8].

"These tools are undergoing very fast development," said Siwei Lyu, professor of computer science and director of the Media Forensic Laboratory at the State University of New York in Buffalo, NY, USA. "The trend I am seeing is higher quality, more realistic, and faster, with some algorithms using just somebody's face to generate a video on the fly."

To create the DFDC, Facebook collaborated with Partnership on AI (an AI research and advocacy organization based in San Francisco, CA, USA, that includes Google and Amazon as corporate members), Microsoft, and university scientists in the United States, United Kingdom, Germany, and Italy [3]. "The challenge generated a lot of attention from the research community," said Lyu, who served as an academic advisor for the competition.

The contest provided more than 100 000 newly created 10 s video clips (the DFDC dataset) of face-swap manipulations to train the detection models of the 2114 researchers in academia and industry who submitted entries [4,9]. The contestants' codes were tasked with identifying the deepfakes in the dataset, which included videos altered with a variety of techniques, some of which were likely unfamiliar to existing detection models [3,4]. Their algorithms were then tested against a black box dataset of more than 4000 video clips, including some augmented via advanced methods not used in the training dataset. The results of the competition—and winners of 1 million USD in prize money—were announced in June 2020.

The best models accurately picked out more than 80% of the manipulated videos in the training dataset. With the black box dataset, however, they did not fare as well. In this more realistic scenario, with no training on similarly manipulated data, the most successful code correctly identified only 65% of the deepfakes [4]. The other four winning teams posted results that were close behind. The low success rate "reinforces that building systems that generalize to unseen deepfake generation techniques is still a hard and open research problem," said Kristina Milian, a Facebook company spokesperson.

While "cheapfakes" are easy to make on almost any machine and easy to spot, the best of today's deepfakes are made with complex computer hardware, including a graphics processing unit, said Edward Delp, a professor of computer engineering at Purdue University in West Layfayette, IN, USA. In such altered videos, the lip sync or head tilt might be only slightly and subtly off. The winning code in the DFDC, submitted by machine learning engineer Selim Seferbekov at the mapping firm Mapbox in Minsk, Belarus, used machine learning tools to pick up pixels around a person's head as it moved that were inconsistent with the background. "It was a pretty sophisticated approach," Delp said.

Deepfake code now often includes distracting factors, such as resizing or cropping of the video frames, blurring them a little, or recompressing them, which can introduce artifacts that complicate detection, Delp said. The accuracy of a detection algorithm therefore depends on the diversity and quality of examples in the dataset it was trained on, as shown by the DFDC results.

The key to accurate detection involves correctly spotting inconsistencies, said Matt Turek, a program manager in the Information Innovation Office at the US Defense Advanced Research Projects Agency (DARPA) in Arlington, VA, USA. In addition to digital artifacts, one can examine a video's physical integrity, such as whether the lighting and shadows match correctly, and can look for semantic inconsistencies, such as whether the weather in a video matches what is known independently. One can also analyze the social context of a deepfake's creation and discovery to infer the intent of the person who published it [10]. DARPA has begun

https://doi.org/10.1016/j.eng.2020.10.008

dedicated research in this area in its new semantic forensics program [11].

In all detection efforts, the biggest problem might not be missing a couple manipulated videos but incorrectly flagging many more unaltered ones. "It is the false positives that kill you," said Nasir Memon, a professor of computer science at New York University in New York City, NY, USA. If most of the events are benign, he said, what is known as the "base rate fallacy" always makes detection problematic. For example, it is likely that only a handful of the millions of videos people upload to YouTube every day have been manipulated. Given such numbers, even a detection algorithm with 99% accuracy would flag many thousands of benign videos incorrectly, making it difficult to quickly catch the truly malicious ones. "You cannot respond to all of them," Memon said.

To reduce the impact of false positives, some digital forensic experts are focusing on the opposite side of the problem, which was not incorporated into the DFDC contest. "Instead of chasing down what is fake, I have been working on establishing the provenance of what is not fake," said Shweta Jain, a professor of computer science at John Jay College of Criminal Justice in New York City, NY, USA.

Using blockchain technology, Jain has developed E-Witness, a way to register a unique "hash," or fingerprint, for image or video files that can be recomputed to verify their integrity [12]. The process is similar to using watermarks with photographs but more difficult for someone to tamper with since the original hash will always live in a blockchain, Jain said. The hash can include "meta data" about the file, including information about the device that made the image or video, location data, and data compression algorithm used. DARPA researchers are also working on secure ways to attribute media to a particular source, but these efforts remain in early development, Turek said.

Meanwhile, the ability to create algorithms that produce altered yet convincing media while evading detection continues to improve as well [9]. "You always assume your adversary knows your techniques," Memon said. "Then it becomes a cat and mouse game." In the most recent developments of this game, Microsoft has developed its own deepfake detection tool [13], and TikTok has followed other social media companies, including Facebook and Twitter [14,15], in beginning to take steps to ban deepfakes on its platform [16].

References

[1] Skibba R. Media enhanced by artificial intelligence: can we believe anything anymore? Engineering 2020;6(7):723–4.

[2] Adee S. What are deepfakes and how are they created? [Internet]. New York: IEEE Spectrum; 2020 Apr 29 [cited 2020 Aug 30]. Available from: https://spectrum.ieee.org/tech-talk/computing/software/what-are-deepfakes-how-are-they-created.

[3] Ferrer CC, Dolhansky B, Pflaum B, Bitton J, Pan J, Lu J. Deepfake detection challenge results: an open initiative to advance AI [Internet]. Menlo Park: Facebook AI Blog; 2020 Jun 12 [cited 2020 Sep 15]. Available from: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

[4] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deepfake detection challenge dataset. 2020. arXiv:2006.07397.

[5] Knight W. Deepfakes aren't very good. Nor are the tools to detect them [Internet]. San Francisco: Wired; 2020 Jun 12 [cited 2020 Sep 20]. Available from: https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/.

[6] Manjoo F. How do you know a human wrote this? [Internet]. New York: New York Times; 2020 Jul 29 [cited 2020 Sep 20]. Available from: https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html?smid=em-share.

[7] Brockman G, Murati M, Welinder P; OpenAI. OpenAI API [Internet]. San Francisco: OpenAI; 2020 Jun 11 [cited 2020 Sep 15]. Available from: https://openai.com/blog/openai-api/.

[8] Lyu S. Deepfakes and the new AI-generated fake media creation-detection arms race [Internet]. New York: Scientific American; 2020 Jul 20 [cited 2020 Sep 10]. Available from: https://www.scientificamerican.com/article/detecting-deepfakes1/.

[9] Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The deepfake detection challenge (DFDC) preview dataset. 2019. arXiv:1910.08854v2.

[10] Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S. Deep learning for deepfakes creation and detection: a survey. 2019. arXiv:1909.11573.

[11] Turek M. Semantic forensics (SemaFor) [Internet]. Arlington: DARPA; c2020 [cited 2020 Sep 10]. Available from: https://www.darpa.mil/program/semantic-forensics.

[12] Samanta P, Jain S. E-Witness: preserve and prove forensic soundness of digital evidence. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking; 2018 Oct 29–Nov 2; New Delhi, India; 2018. p. 832–4.

[13] Kelion L. Deepfake detection tool unveiled by Microsoft [Internet]. London: BBC News; 2020 Sep 1 [cited 2020 Sep 25]. Available from: https://www.bbc.com/news/technology-53984114.

[14] Kelly M. Facebook bans deepfake videos ahead of the 2020 election [Internet]. New York: The Verge; 2020 Jan 7 [cited 2020 Sep 15]. Available from: https://www.theverge.com/2020/1/7/21054504/facebook-instagram-deepfake-ban-videos-nancy-pelosi-congress.

[15] Robertson A. Twitter will ban 'deceptive' faked media that could cause 'serious harm' [Internet]. New York: The Verge; 2020 Feb 4 [cited 2020 Sep 15]. Available from: https://www.theverge.com/2020/2/4/21122661/twitter-deepfake-manipulated-media-policy-rollout-date.

[16] Statt N. TikTok is banning deepfakes to better protect against misinformation [Internet]. New York: The Verge; 2020 Aug 5 [cited 2020 Sep 15]. Available from: https://www.theverge.com/2020/8/5/21354829/tiktok-deepfakes-ban-misinformation-us-2020-election-interference.