Deepfakes aren’t very good—nor are the tools to detect them

Enlarge / A comparison of an original and deepfake video of Facebook CEO Mark Zuckerberg.

reader comments

27 with 19 posters participating

Share this story

We’re lucky that deepfake videos aren’t a big problem yet. The best deepfake detector to emerge from a major Facebook-led effort to combat the altered videos would only catch about two-thirds of them.

In September, as speculation about the danger of deepfakes grew, Facebook challenged artificial intelligence wizards to develop techniques for detecting deepfake videos. In January, the company also banned deepfakes used to spread misinformation.

Facebook’s Deepfake Detection Challenge, in collaboration with Microsoft, Amazon Web Services, and the Partnership on AI, was run through Kaggle, a platform for coding contests that is owned by Google. It provided a vast collection of face-swap videos: 100,000 deepfake clips, created by Facebook using paid actors, on which entrants tested their detection algorithms. The project attracted more than 2,000 participants from industry and academia, and it generated more than 35,000 deepfake detection models.

The best model to emerge from the contest detected deepfakes from Facebook's collection just over 82 percent of the time. But when that algorithm was tested against a set of previously unseen deepfakes, its performance dropped to a little over 65 percent.

"It's all fine and good for helping human moderators, but it's obviously not even close to the level of accuracy that you need," says Hany Farid, a professor at UC Berkeley and an authority on digital forensics, who is familiar with the Facebook-led project. "You need to make mistakes on the order of one in a billion, something like that."

Deepfakes use artificial intelligence to digitally graft a person's face onto someone else, making it seem as if that person did and said things they never did. For now, most deepfakes are bizarre and amusing; a few have appeared in clever advertisements.

The worry is that deepfakes might someday become a particularly powerful and potent weapon for political misinformation, hate speech, or harassment, spreading virally on platforms such as Facebook. The bar for making deepfakes is worryingly low, with simple point-and-click programs built on top of AI algorithms already freely available.

"Frustrated"

"I was pretty personally frustrated with how much time and energy smart researchers were putting into making better deepfakes," says Mike Schroepfer, Facebook's chief technology officer. He says the challenge aimed to encourage "broad industry focus on tools and technologies to help us detect these things, so that if they're being used in malicious ways we have scaled approaches to combat them."

Schroepfer considers the results of the challenge impressive, given that entrants had only a few months. Deepfakes aren't yet a big problem, but Schroepfer says it's important to be ready in case they are weaponized. "I want to be really prepared for a lot of bad stuff that never happens rather than the other way around," Schroepfer says.

The top-scoring algorithm from the deepfake challenge was written by Selim Seferbekov, a machine-learning engineer at Mapbox, who is in Minsk, Belarus; he won $500,000. Seferbekov says he isn't particularly worried about deepfakes, for now.

"At the moment their malicious use is quite low, if any," Seferbekov says. But he suspects that improved machine-learning approaches could change this. "They might have some impact in the future the same as the written fake news nowadays." Seferbekov's algorithm will be open sourced, so that others can use it.

Cat and mouse

Catching deepfakes with AI is something of a cat-and-mouse game. A detector algorithm can be trained to spot deepfakes, but then an algorithm that generates fakes can potentially be trained to evade detection. Schroepfer says this caused some concern around releasing the code from the project, but Facebook concluded that it was worth the risk in order to attract more people to the effort.

Facebook already uses technology to automatically detect some deepfakes, according to Schroepfer, but the company declined to say how many deepfake videos have been flagged this way. Part of the problem with automating the detection of deepfakes, Schroepfer says, is that some are simply entertaining while others might do harm. In other words, as will other forms of misinformation, the context is important. And that is hard for a machine to grasp.

Creating a really useful deepfake detector might be even harder than the contest suggests, according to Farid of UC Berkeley, because new techniques are rapidly emerging and a malicious deepfake maker might work hard to outwit a particular detector.

Farid questions the value of such a project when Facebook seems reluctant to police the content that users upload. “When Mark Zuckerberg says we are not the arbiters of truth, why are we doing this?” he asks.

Even if Facebook’s policy were to change, Farid says the social media company has more pressing misinformation challenges. “While deepfakes are an emerging threat, I would encourage us not to get too distracted by them,” says Farid. “We don’t need them yet. The simple stuff works.”

This story originally appeared on wired.com.

https://arstechnica.com/?p=1683607