Demystifying Deepfakes: 3 Truths About AI-Generated Videos

Date : November 14, 2019

This week amidst all the Twitter memes and hashtags, something big was trending.

Del Harvey, vice president of trust and safety at Twitter, announced a draft of how the social-media platform plans to handle "synthetic and manipulated" media that purposely tries to mislead or confuse people. In his blog post, Harvey said, "Twitter may: place a notice next to Tweets that share synthetic or manipulated media; warn people before they share or like Tweets with synthetic or manipulated media; or add a link – for example, to news article or Twitter Moment – so that people can read more about why various sources believe the media is synthetic or manipulated."

One concept that has been getting lots of attention lately helped to inspire Twitter's announcement: deepfakes.

A deepfake is an artificial intelligence-generated video that shows someone saying or doing something that he or she never actually did. These altered videos use neural networks to overlay someone's face – in other words, computing systems learn to perform tasks by considering examples and recognizing patterns. For example, Facebook's Mark Zuckerberg was deepfaked this year when a video of him showed up online preaching about FB's powers – but it wasn't actually Zuckerberg. The video had been altered to look like him.

As Zuckerberg and other deepfaked celebs made news in recent months, the deepfake discussion also escalated at the University of Pennsylvania in Philadelphia, Pa., where Knowledge@Wharton High School is based. During PennApps XX in September – one of the world's largest college hackathons – a team of four local students took home the grand prize, beating out some 250 teams from more than 750 high schools and colleges. Their winning project? DeFake, described as "a browser extension that uses machine learning to assess the likelihood that a given video has been subtly manipulated." It's an app designed to detect deepfakes.

KWHS caught up with Sofiya Lysenko, a senior at Abington Senior High School in Pennsylvania and a leader of the DeFake team, to learn more about deepfake technology. "I think what is so fascinating about deepfakes is how difficult they are to detect with known computational methods," notes Lysenko, who has competed in PennApps for several years and has also won attention for other machine learning projects. When she was 14, for instance, she created a program that could predict the next mutation in the Zika virus. "Our team was stuck in the very beginning about how to resolve [deepfake detection]. We investigated several methods, which ultimately we found to be unsatisfactory, until we tried and were successful with the final methods of the project," adds Lysenko, who created DeFake's machine-learning algorithm that helps determine if a video is fake or real. "Machine learning and computer vision [a field of computer science that enables computers to see, identify and process images in the same way that human vision does] are becoming interesting topics to learn because of all the applications that stem from them, such as deepfakes."

Inspired by Lysenko's deep research into deepfakes, here are a few additional truths to help demystify this infamous technology:

1. Still not quite sure how this works? Michael Kearns, a computer and information science professor at the University of Pennsylvania, recently suggested that the process to create a deepfake is like a "personal trainer" for software. Speaking with *The Christian Science Monitor* in October, Kearns helped them to better understand how a deep-learning application compares one image with another to identify distinguishing characteristics and uses that information to then create a synthetic image. Each time the program successfully identifies the differences between a fake image and a real one, "the next fake it produces becomes more seemingly authentic" – or in better shape, as the personal trainer image suggests. As deepfakes become more and more

indistinguishable from the real thing, Kearns added this warning: “Be ever vigilant.”

2. Growing concerns about deepfakes – and even DeFake’s recent hackathon grand-prize victory – have lots to do with the upcoming race for president in the U.S. “Deepfakes are a threat that needs to be detected due to the possibility that this could be used as a quick and deceptive form of misinformation as we approach the 2020 US Presidential elections,” says Lysenko. In fact, DeFake describes the motivation for its machine-learning project like this: “The upcoming election season is predicted to be drowned out by a mass influx of fake news. Deepfakes are a new method to impersonate famous figures saying fictional things, and could be particularly influential in the outcome of this and future elections. With international misinformation becoming more common, we wanted to develop a level of protection and reality for users.” Alex Wang, a U Penn freshman studying computer and information science in the School of Engineering and Applied Science, provides this context in an opinion piece in the *Penn Political Review* “Much of the concern surrounding deepfakes centers around the 2020 election due to the existence of both large datasets and motivations to target political figures. What would the public reaction be if a doctored video of Elizabeth Warren disparaging Mexican immigrants were to be released?…Would it be legal for Joe Biden’s campaign to create negative deepfakes of opposition candidates?”

3. In the universe of Internet interaction, deepfakes have much broader implications about how we communicate and how we make decisions based on what we believe to be true. Wharton management professor Saikat Chaudhuri, who is also the executive director of Wharton’s Mack Institute for Innovation Management, recently interviewed Sherif Hanna, vice president of ecosystem development at Truepic, a photo and video verification platform, on his SiriusXM radio show, *Mastering Innovation*. Hanna, whose company has developed a solution to verify the source of images, has given a great deal of thought to this issue of misrepresentation, noting that the website thispersondoesnotexist.com presents a completely AI-generated image every time you hit the refresh button. “We as a society and as a world at large depend on photos and videos in almost every aspect of daily life…at the same time there’s a rising tide of threats against those photos and videos that we’ve come to rely on and there’s a decline in trust of what you see,” said Hanna. “The danger for society is losing consensus around the shared sense of reality. That’s kind of a big deal if we can’t all agree on what it is that happened because we can’t agree that we trust the photos and videos that document events. It becomes very difficult to make joint decisions as a society if everyone’s perception of what happened differs substantially.”

Truepic and the PennApps project DeFake are working to restore and preserve truth – at least in what we see. Lysenko, who plans to pursue a career in research by leading her own research group and teaching at a university, as well as developing technologies within startups and companies, calls machine learning “truly a super power when it comes to solving some of the hardest challenges today.”

Skeptics, like Wang, point out that the superpower can also serve to make the challenge even greater, calling deepfake detection a losing battle. “The ongoing battle to detect deepfakes is a perfect mirror for the technology itself: as algorithms to detect deepfakes improve, deepfake creators adapt to changes by generating even more realistic ones.”

Is what you’re viewing real? “Be ever vigilant.”