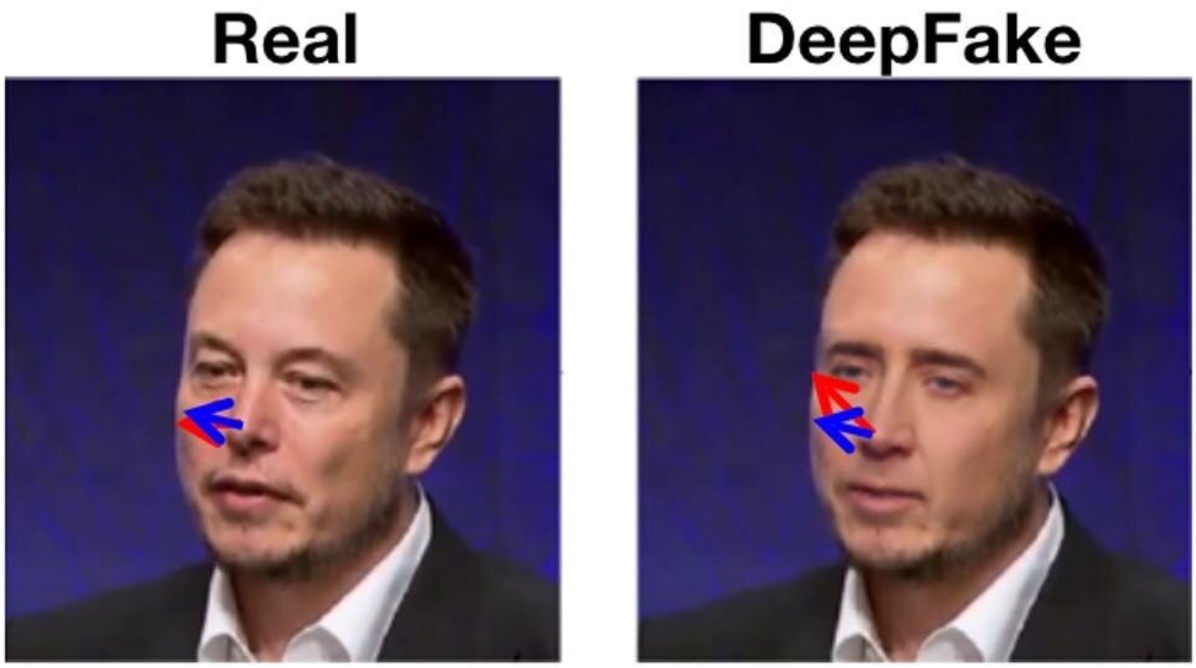
Detecting deepfakes by looking closely reveals a way to protect against them

June 26 2019, by Siwei Lyu



When a computer puts Nicolas Cage's face on Elon Musk's head, it may not line up the face and the head correctly. Credit: Siwei Lyu, CC BY-ND

Deepfake videos are hard for untrained eyes to detect because they can be quite realistic. Whether used as personal weapons of revenge, to manipulate financial markets or to destabilize international relations, videos depicting people doing and saying things they never did or said are a fundamental threat to the longstanding idea that "seeing is

believing." Not anymore.

Most deepfakes are made by showing a computer algorithm many images of a person, and then having it use what it saw to generate new face images. At the same time, their voice is synthesized, so it both looks and sounds like the person has said something new.

Some of my research group's earlier work allowed us to detect deepfake videos that did not include a person's normal amount of eye blinking—but the latest generation of deepfakes has adapted, so our research has continued to advance.

Now, our research can identify the manipulation of a video by looking closely at the pixels of specific frames. Taking one step further, we also developed an active measure to protect individuals from becoming victims of deepfakes.

Finding flaws

In two recent research papers, we described ways to detect deepfakes with flaws that can't be fixed easily by the fakers.

When a deepfake video synthesis algorithm generates new facial expressions, the new images don't always match the exact positioning of the person's head, or the lighting conditions, or the distance to the camera. To make the fake faces blend into the surroundings, they have to be geometrically transformed—rotated, resized or otherwise distorted. This process leaves digital artifacts in the resulting image.

You may have noticed some artifacts from particularly severe transformations. These can make a photo look obviously doctored, like blurry borders and artificially smooth skin. More subtle transformations still leave evidence, and we have taught an algorithm to detect it, even

when people can't see the differences.

These artifacts can change if a deepfake video has a person who is not looking directly at the camera. Video that captures a real person shows their face moving in three dimensions, but deepfake algorithms are not yet able to fabricate faces in 3-D. Instead, they generate a regular two-dimensional image of the face and then try to rotate, resize and distort that image to fit the direction the person is meant to be looking.

At left, a face is easily detected in an image before our processing. In the middle, we've added perturbations that cause an algorithm to detect other faces, but not the real one. At right are the changes we added to the image, enhanced 30 times to be visible. Credit: Siwei Lyu, CC BY-ND

They don't yet do this very well, which provides an opportunity for detection. We designed an algorithm that calculates which way the person's nose is pointing in an image. It also measures which way the head is pointing, calculated using the contour of the face. In a real video of an actual person's head, those should all line up quite predictably. In deepfakes, though, they're often misaligned.

Defending against deepfakes

The science of detecting deepfakes is, effectively, an arms race—fakers will get better at making their fictions, and so our research always has to try to keep up, and even get a bit ahead.

If there were a way to influence the algorithms that create deepfakes to be worse at their task, it would make our method better at detecting the fakes. My group has recently found a way to do just that.

Image libraries of faces are assembled by algorithms that process thousands of online photos and videos and use machine learning to detect and extract faces. A computer might look at a class photo and detect the faces of all the students and the teacher, and add just those faces to the library. When the resulting library has lots of high-quality face images, the resulting deepfake is more likely to succeed at deceiving its audience.

We have found a way to add specially designed noise to digital photographs or videos that are not visible to human eyes but can fool the face detection algorithms. It can conceal the pixel patterns that face detectors use to locate a face, and creates decoys that suggest there is a face where there is not one, like in a piece of the background or a square of a person's clothing.

With fewer real faces and more nonfaces polluting the training data, a deepfake algorithm will be worse at generating a fake face. That not only slows down the process of making a deepfake, but also makes the resulting deepfake more flawed and easier to detect.

As we develop this algorithm, we hope to be able to apply it to any images that someone is uploading to social media or another online site. During the upload process, perhaps, they might be asked, "Do you want

to protect the faces in this video or image against being used in deepfakes?" If the user chooses yes, then the algorithm could add the digital noise, letting people online see the faces but effectively hiding them from algorithms that might seek to impersonate them.

Provided by The Conversation