This deep neural network fights deepfakes

July 19 2019

AI will serve to develop a network control system that not only detects and reacts to problems but can also predict and avoid them. Credit: CC0 Public Domain

Seeing was believing until technology reared its mighty head and gave us powerful and inexpensive photo-editing tools. Now, realistic videos that map the facial expressions of one person onto those of another, known as deepfakes, present a formidable political weapon.

But whether it's the benign smoothing of a wrinkle in a portrait, or a video manipulated to make it look like a politician saying something offensive, all photo editing leaves traces for the right tools to discover.

Research led by Amit Roy-Chowdhury's Video Computing Group at the University of California, Riverside has developed a deep neural network architecture that can identify manipulated images at the pixel level with high precision. Roy-Chowdhury is a professor of electrical and computer engineering and the Bourns Family Faculty Fellow in the Marlan and Rosemary Bourns College of Engineering.

A deep neural network is what artificial intelligence researchers call computer systems that have been trained to do specific tasks, in this case, recognize altered images. These networks are organized in connected layers; "architecture" refers to the number of layers and structure of the connections between them.

Objects in images have boundaries and whenever an object is inserted or removed from an image, its boundary will have different qualities than the boundaries of objects in the image naturally. For example, someone with good Photoshop skills will do their best to make the inserted object looks as natural as possible by smoothing these boundaries.

While this might fool the naked eye, when examined pixel by pixel, the boundaries of the inserted object are different. For example, they are often smoother than the natural objects. By detecting boundaries of inserted and removed objects, a computer should be able to identify altered images.

The researchers labeled nonmanipulated images and the relevant pixels in boundary regions of manipulated images in a large dataset of photos. The aim was to teach the neural network general knowledge about the manipulated and natural regions of photos. They tested the neural network with a set of images it had never seen before, and it detected the altered ones most of the time. It even spotted the manipulated region.

"We trained the system to distinguish between manipulated and nonmanipulated images, and now if you give it a new image it is able to provide a probability that that image is manipulated or not, and to localize the region of the image where the manipulation occurred," Roy-Chowdhury said.

The researchers are working on still images for now, but they point out that this can also help them detect deepfake videos.

"If you can understand the characteristics in a still image, in a video it's basically just putting still images together one after another," Roy-Chowdhury said. "The more fundamental challenge is probably figuring out whether a frame in a video is manipulated or not."

Even a single manipulated frame would raise a red flag. But Roy-Chowdhury thinks we still have a long way to go before automated tools can detect deepfake videos in the wild.

"It's a challenging problem," Roy-Chowdhury said. "This is kind of a cat and mouse game. This whole area of cybersecurity is in some ways trying to find better defense mechanisms, but then the attacker also finds better mechanisms."

He said completely automated deepfake detection might not be achievable in the near future.

"If you want to look at everything that's on the internet, a human can't do it on the one hand, and an automated system probably can't do it reliably. So it has to be a mix of the two," Roy-Chowdhury said.

Deep neural network architectures can produce lists of suspicious videos and images for people to review. Automated tools can reduce the amount of data that people—like Facebook content moderators—have to sift through to determine if an image has been manipulated.

For this use, the tools are right around the corner.

"That probably is something that these technologies will contribute to in a very short time frame, probably in a few years," Roy-Chowdhury said.

More information: Jawadul H. Bappy et al. Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries, *IEEE Transactions on Image Processing* (2019). DOI: 10.1109/TIP.2019.2895466

Provided by University of California - Riverside

Citation: This deep neural network fights deepfakes (2019, July 19) retrieved 25 April 2024 from https://techxplore.com/news/2019-07-deep-neural-network-deepfakes.html