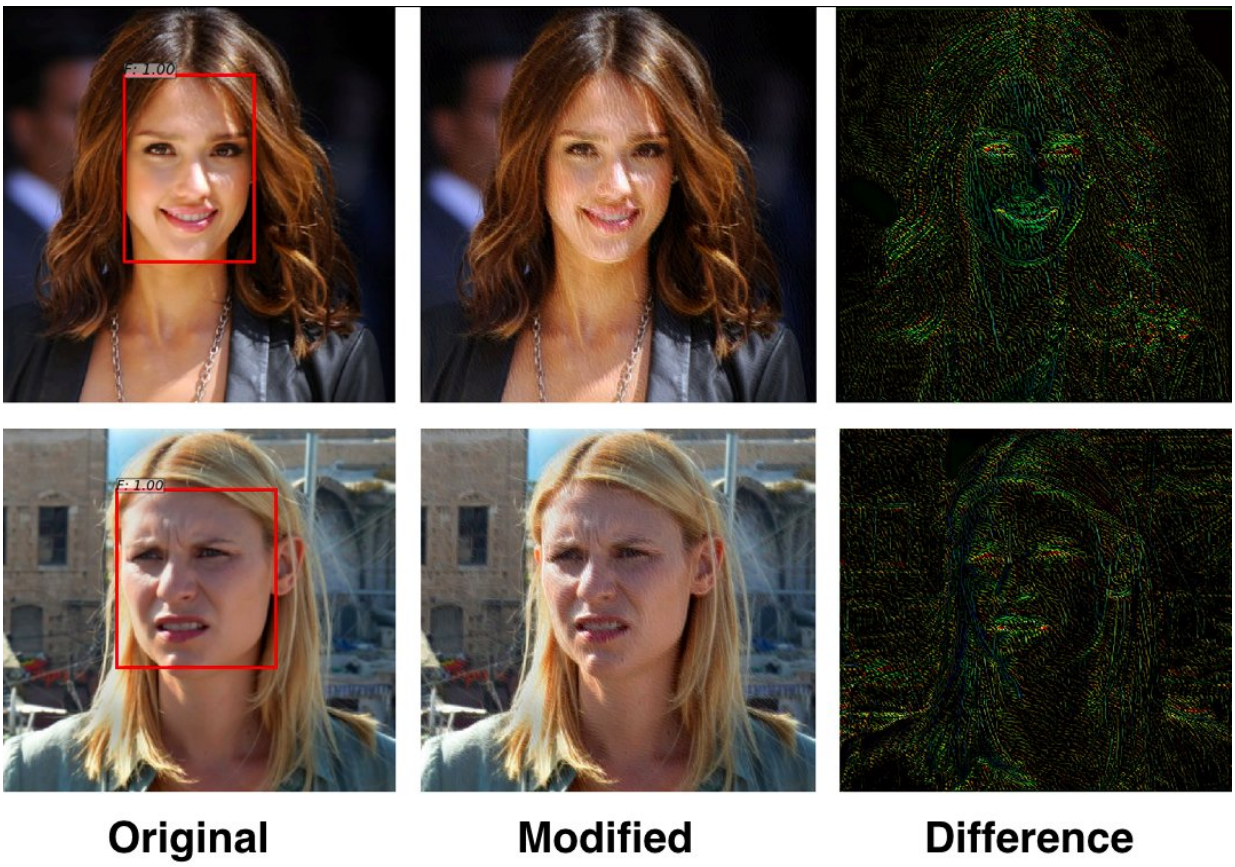
AI researchers design 'privacy filter' for your photos that disables facial recognition systems

May 31 2018



Researchers in U of T Engineering have designed a 'privacy filter' that disrupts facial recognition algorithms. The system relies on two AI-created algorithms: one performing continuous face detection, and another designed to disrupt the first. Credit: Avishek Bose

Each time you upload a photo or video to a social media platform, its facial recognition systems learn a little more about you. These algorithms ingest data about who you are, your location and people you know — and they're constantly improving.

As concerns over privacy and data security on social networks grow, U of T Engineering researchers led by Professor Parham Aarabi and graduate student Avishek Bose have created an algorithm to dynamically disrupt facial recognition systems.

"Personal privacy is a real issue as facial recognition becomes better and better," says Aarabi. "This is one way in which beneficial anti-facial-recognition systems can combat that ability."

Their solution leverages a deep learning technique called adversarial training, which pits two artificial intelligence algorithms against each other. Aarabi and Bose designed a set of two neural networks: the first working to identify faces, and the second working to disrupt the facial recognition task of the first. The two are constantly battling and learning from each other, setting up an ongoing AI arms race.

The result is an Instagram-like filter that can be applied to photos to protect privacy. Their algorithm alters very specific pixels in the image, making changes that are almost imperceptible to the human eye.

"The disruptive AI can 'attack' what the neural net for the face detection is looking for," says Bose. "If the detection AI is looking for the corner of the eyes, for example, it adjusts the corner of the eyes so they're less noticeable. It creates very subtle disturbances in the photo, but to the detector they're significant enough to fool the system."

Aarabi and Bose tested their system on the 300-W face dataset, an industry standard pool of more than 600 faces that includes a wide range

of ethnicities, lighting conditions and environments. They showed that their system could reduce the proportion of faces that were originally detectable from nearly 100 per cent down to 0.5 per cent.

"The key here was to train the two neural networks against each other—with one creating an increasingly robust facial detection system, and the other creating an ever stronger tool to disable facial detection," says Bose, the lead author on the project. The team's study will be published and presented at the 2018 IEEE International Workshop on Multimedia Signal Processing later this summer.

In addition to disabling facial recognition, the new technology also disrupts image-based search, feature identification, emotion and ethnicity estimation, and all other face-based attributes that could be extracted automatically.

Next, the team hopes to make the privacy filter publicly available, either via an app or a website.

"Ten years ago these algorithms would have to be human defined, but now neural nets learn by themselves — you don't need to supply them anything except training data," says Aarabi. "In the end they can do some really amazing things. It's a fascinating time in the field, there's enormous potential."

Provided by University of Toronto

Citation: AI researchers design 'privacy filter' for your photos that disables facial recognition systems (2018, May 31) retrieved 18 April 2024 from https://techxplore.com/news/2018-05-ai-privacy-filter-photos-disables.html