Defending against Deepfake Creations

1 min

For deepfake technology to operate successfully, it generally requires a base training dataset. While some aspects, such as audio generation, may only need a small snippet of data, typically, the more data a model has, the more effectively it will generate its required content. Because of this, many researchers have begun investigating ways to tamper or manipulate potential input data in such a way that it will negatively affect the AI models.

A great example of this can be seen when referring to deepfake AI-generated images. Researchers have begun to play with the concept of adding "noise" to image files. This concept is simple in theory but quite powerful in practice. By manipulating an image, it is possible to add additional data and information to an image file that, while unnoticeable by humans, is observed and processed by AI models. By doing this, some models begin to include this noise data in their generation models. While the original noise is not expressed visibly, the model integrates this noise to better detect and defend against deep fake images.

Similar protections are also being introduced by tools like "NightShade" and "Gaze". These tools have been developed by researchers for artists to help protect their intellectual property from various image generation models. By injecting invisible pixels into these images, models produce unpredictable and undesirable content.

Of course, researchers need to continue to develop these systems. As models improve, these defense measures will also need to improve. In many ways, we have created a new cat-and-mouse game where AI developers and developers attempting to thwart certain aspects of AI are trying to perform one another out.

Image by Spamrakuen, CC0, via Wikimedia Commons