Research Against Mitigation of Generative Adversarial Network

Author: Jhonatan Leonardo Zuluaga Torres

DeepFakes Is reality in danger? Is it possible to protect yourself from manipulation with artificial intelligence?

Currently with AI (artificial intelligence) it is possible to generate multiple applications in the area of information, medicine, engineering, architecture, astronomy etc.

However, this time we are going to talk about something not so positive that arose as a result of this progress. We refer to the famous Deep Fakes that have given what to talk about in recent months due to the concern they have caused not only in those who are engaged in artificial intelligence research but also in governments.

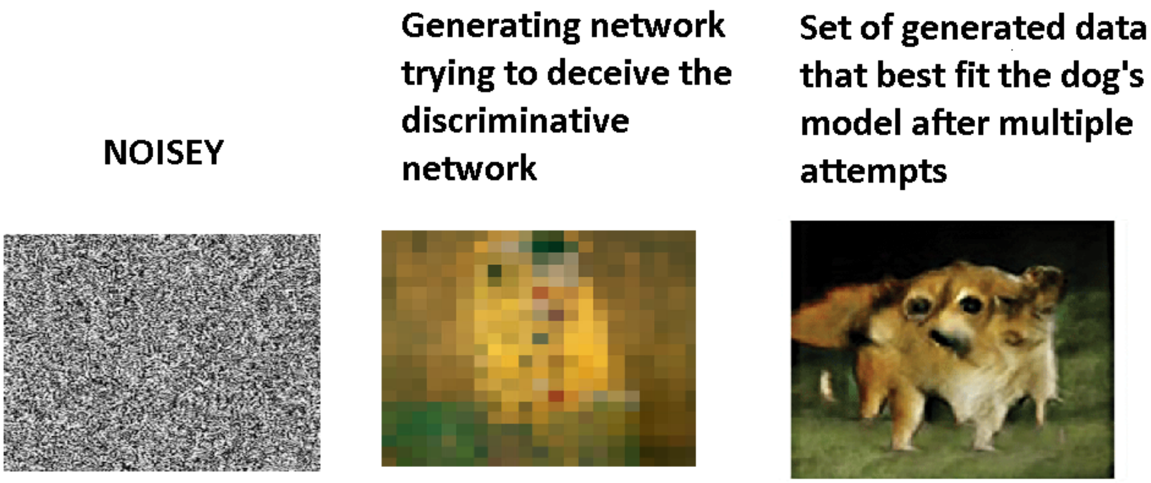
And to explain it, we first have to understand what the GANs (Antagonistic Generative Networks) are.

What are the GANs?

As the name implies, GANs are nothing more than placing a discriminative (antagonistic) neural network to confront or compete against a regular generative neural network. Remember that neural networks are a machine learnig system that artificially seeks to recreate a biological neuron using nodes and links that can be adjusted by passing them through a "training" process.

Basically, the discriminative network enters the game with a previously trained model, while the generative network enters the game trying to match the images or audio that fit those already set by the discriminant model. The generative network will move its first file sending an image of junk bytes known as “noise”, said principle will be known to be incorrect but has the sole purpose of being immediately discarded by the discriminative network, since the generating network expects its Forcibly opposing the discard by throwing a signal on the errors that justify discarding that image.

In this way the generative network will make small corrections and new attempts continuously rejected until it can fit with the most accurate model possible. But the game is a simple "cold" and "hot" to try to hit a model, the interesting thing is that the generating network tries to deceive the discriminating network trying to get false data from the image as if they were real using ambiguities, which will force the discriminating network to perfect the verification of its model so as not to continue being deceived. It is precisely this continuous process of attempted deception and verification that generates such a successful associative structure capable of creating new images through pre-established concepts.


NOISEY

Generating network trying to deceive the discriminative network

Set of generated data that best fit the dog's model after multiple attempts

Generating neural network trying to meet the conceptual expectations of a "dog" for a previously trained desicrimination network

And what is the threat?

The problem arises because this technology can also be used for various purposes of deception and manipulation. The most famous are the use of superimposing faces of influential people in different bodies or even conserving the image and movement so that the person says or performs actions that I never perform but being indifferent from a real video since they are based on real mechanics rigorously learned . So far it has only been used in a didactic way but the destructive potential for society and media warfare is more than obvious. It would not be difficult with this type of technology to cause widespread outrage events that can trigger even armed wars.

Is there any way to counter it?

Although there is already financing of programs by different private and governmental entities to mitigate the impact of this new technology, the truth is that I am one of those who think that a point of no return has been reached, since the artificial intelligence triggered has a impact capacity and evolution superior to mitigation capacity by laws and research.

However, I have proposed a possible system to contain this potential threat in the future:

The only real way to do this is to focus directly on what is not necessarily within reach of being manipulated by artificial intelligence, in other words, the point of origin, time, hardware and infrastructure.

Ej:

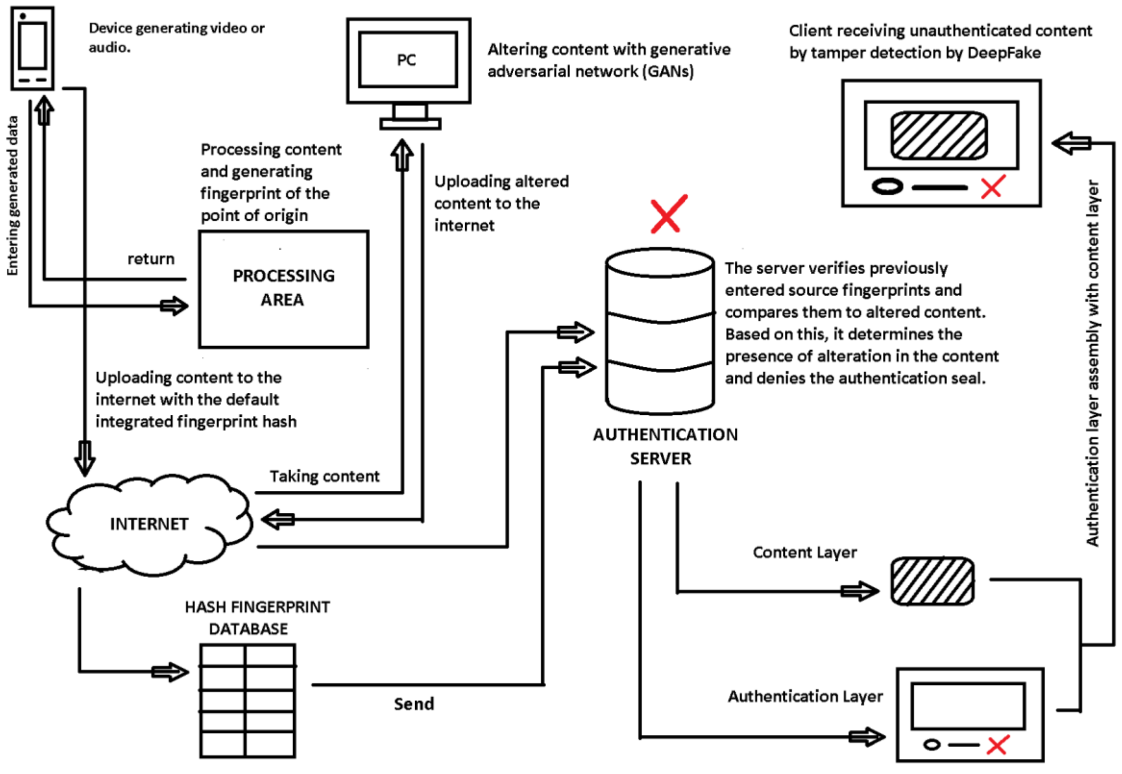
1. Each originally recorded video and audio must have a fingerprint.

2. Taking into account that the video compositions are made with fragments of other videos and audios, each device must generate this fingerprint at the same time as the one that is generating the material from the device as the point of origin.

3. The sum of these authenticated footprints within an edited video must generate an authentication seal, said seal will be processed and generated in a separate layer or in an alternate network that is isolated from the internet and will be dedicated solely to this task.

4. This stamp must be placed on each video uploaded to the network and on a development layer equally separated from the audiovisual material.

In this way, through an effective and visible authentication, the population can easily identify a reliable video as information material of another that is not. Naturally this will require rethinking the manufacturing of the devices and developing new network protocols focused on this new task.


Device generating video or audio.
PC
Altering content with generative adversarial network (GANs)
Client receiving unauthenticated content by tamper detection by DeepFake
Entering generated data
Processing content and generating fingerprint of the point of origin
return
PROCESSING AREA
Uploading altered content to the internet
The server verifies previously entered source fingerprints and compares them to altered content. Based on this, it determines the presence of alteration in the content and denies the authentication seal.
Authentication layer assembly with content layer
Uploading content to the internet with the default integrated fingerprint hash
AUTHENTICATION SERVER
Taking content
INTERNET
Content Layer
HASH FINGERPRINT DATABASE
Send
Authentication Layer

GAN DETECTION AND AUTHENTICATION SYSTEM (GDAS)

Author: Jhonatan Leonardo Zuluaga Torres