C2 Houston, We Have a Problem: Deepfake Is the Word!

Sebastiano Battiato, PhD, Università di Catania, Catania 95125, ITALY; Oliver Giudice, PhD*, Rome 00044, ITALY; Luca Guarnera, MS*, Università di Catania, Catania 95125, ITALY*

Learning Overview: After attending this presentation, attendees will be familiar with various techniques of Deepfake creation and forensic detection. The main techniques are currently based on statistical analysis of Deepfakes coming from specific architectures.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by attempting to present all the existing solutions as of today with a simple but schematic view. All the techniques will be presented organized in various categories and the corresponding pros and cons will be exposed, considering practical application scenarios.

Multimedia contents created by means of Artificial Intelligence (AI) techniques are introducing several big problems in the multimedia forensics field with the birth of the Deepfake phenomenon. Images and videos can be created through models like Generative Adversarial Network (GAN) where it is possible to create or swap a person's face from input source to target data.[1] Various state-of-the-art approaches are available to create AI-synthetic data, including film, photography, video game, and virtual reality productions. Politicians, actors, and famous people are the main targets of this new phenomenon: a famous example was the video of the former Italian premier, Matteo Renzi, created by a satirical newscast in September 2019, in which Renzi talked about his colleagues in a "not very respectful" way. Another famous example was the video of Mark Zuckerberg in which he claimed a series of statements about the ability of his platform (Facebook®) to steal data of its users. However, GANs are the new weapons as they can be used to create pornographic Deepfakes, raising a serious security threat. It is a fact that 96% of Deepfakes on the internet are pornographic. This threat needs to be mitigated with new forensic detection techniques capable of identifying whether multimedia data is deepfake, (i.e., created/altered through/by GANs). This isn't a new problem; in fact, image manipulation was born almost immediately after the advent of photography, and currently, through the use of image editing tools (e.g., Adobe® Photoshop®, After Effects® Pro and GNU Image Manipulation Program [GIMP]) it is very easy to perform a forgery operation. Using conventional signal processing tools, images can be easily edited to achieve realistic results that can deceive even a careful observer. Various studies show that a digital image or video (without any kind of manipulation) contains within them different characteristics that allow the acquisition device to be uniquely identified. To detect Deepfakes, different methods have already been presented, many of them based on complex neural network solutions, and claim to obtain excellent classification results, but unfortunately it has been shown that they are not generalizable.[2] It was recently discovered that all GANs leave a fingerprint in the generated images that characterize the generative deep architecture. By exploiting these unique traces, it has been shown that it is possible to distinguish not only if the data is a Deepfake, but also reconstruct, with high precision, the type of architecture employed.

Reference(s):

1. Goodfellow, Ian et al. Generative adversarial nets. *Advances in neural information processing systems*. 2014.
2. Hulzebosch, Nils, Sarah Ibrahimi, and Marcel Worring. Detecting CNN-Generated Facial Images in Real-World Scenarios. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020.

Deepfake, Multimedia Forensics, Generative Adversarial Network

*Presenting Author