Wrestling with the deepfakes: Detection and beyond

openaccessgovernment.org/article/wrestling-with-the-deepfakes-detection-and-beyond/177705

3 June 2024

Siwei Lyu, SUNY Empire Innovation Professor from the University at Buffalo, State University of New York, delves into detection and beyond in the realm of DeepFakes, starting with a look at what they are

Since 2017, the term “DeepFake” has become widely known, often appearing in news and media. It combines “deep learning” (a type of artificial intelligence [AI] model) and “fake” to describe synthetic media – like text, audio, images, and videos – created with advanced AI technologies. The concept gained notoriety when a Reddit user began sharing AI-generated pornographic videos that superimposed celebrity faces onto other figures.

DeepFakes represent a new chapter in the manipulation of digital media, raising significant safety concerns stemming from the easy access to powerful AI tools that allow almost anyone to create highly realistic fake content. Previously, creating such sophisticated fakes often required extensive technical knowledge, but now, tools with user-friendly web interfaces make this technology accessible to a broader audience.

Recent developments in generative AI, such as Midjourney and Stable Diffusion for images, Elevenlab for audio, and Pika and OpenAI’s Sora for videos, illustrate how advanced these generative AI models have become. They can create realistic media content by learning from vast amounts of data available on the Internet and social media.

DeepFake detection

While DeepFakes can be used positively, they also pose a real threat to the truthfulness of digital content and can cause significant harm if used maliciously. The rapid increase in DeepFakes has led to the development of technologies aimed at detecting them. Typically, these detectors analyze whether a given image, audio, or video is real or generated by AI.

Here is how it works: a DeepFake detector utilizes machine learning – a type of AI – to analyze the incoming data (like a video). It categorizes the data into “real” or “AI-generated” by identifying specific characteristics or patterns. DeepFake detection can be grouped into three main types based on the techniques they use:

- Physical and physiological inconsistencies: Some detectors look for errors in the physical or biological details that DeepFakes often get wrong.
- Signal-level artifacts: Others detect anomalies introduced by the AI tools that create the DeepFakes.

- Data-driven methods: The most common detectors use deep neural networks, which are trained to spot DeepFakes by learning from large sets of real and fake data.

An example of such technology is the DeepFake-o-meter, developed by my team. It is a flexible and open-source platform designed to stay up-to-date with the latest advancements in DeepFake detection. This tool integrates various detection modules from researchers around the world into a single, easy-to-use system. It not only helps identify DeepFakes but also provides a user-friendly interface that makes cutting-edge detection methods accessible to everyone.

While current methods for detecting DeepFakes show promising results on benchmark test datasets, relying solely on detection to combat DeepFakes is not enough due to several key limitations.

First, these methods can be less effective when the DeepFake has undergone post-processing operations like compression or resizing. They can also be biased by the data they were trained on or tricked by targeted adversarial attacks meant to evade detection.

Second, most detection techniques only provide a “yes” or “no” answer, which often lacks a detailed explanation. This makes it difficult for non-experts to understand why a piece of media has been flagged as a DeepFake. In addition, these methods tend to act too late, identifying DeepFakes only after they have already been circulated and potentially caused harm.

Although the tools we use to detect DeepFakes aren’t perfect, they still have valuable uses. For instance, online platforms can use these tools to help identify automated accounts and flag them for further investigation.

The continuous improvement of synthesis models put DeepFake making and forensics in a cat-and-mouse game. As such, we must continue developing DeepFake Forensic methods that are more effective, efficient, robust, and explainable, focusing on identifying inconsistencies or inaccuracies, understanding their impact, and uncovering the intent behind DeepFakes. New methods also need to resolve zero-day attacks when a new approach, form, or model of media synthesis is not strongly related to previously known cases.

As the detection of DeepFakes gains attention, it is also essential to tread carefully. Rapid solutions to this intricate issue are often not as straightforward or effective as they may appear. Policymakers and society should be cautious about claims of easy fixes and commit to ongoing research in this evolving field. It is equally important to recognize that detection tools could harbor biases that disproportionately affect marginalized communities.

Beyond DeepFake detection

As distinguishing DeepFakes from authentic content is becoming increasingly challenging, and these detection tools are susceptible to manipulation by malicious entities or might simply fall short. As generative AI and synthetic media become more prevalent –for example, in applications such as high-quality video compression and creating realistic avatars for virtual environments – the line between real and synthetic might blur, making traditional DeepFake detection methods less relevant. Instead, a more pertinent issue could be ensuring that generative AI models are used responsibly and ethically.

Identifying a DeepFake is just the first step in addressing its challenges. To effectively mitigate their harm, it is vital to understand the origin, methods, and intentions behind these fabrications. In addition, proactive measures are necessary to shield users from potential DeepFake attacks. These measures might include techniques that prevent the creation of DeepFakes by disrupting their training processes or by embedding subtle “traces” in training data. These traces, invisible to humans, can later be extracted from DeepFakes as concrete evidence of their fabricated nature.

Furthermore, verifying the authenticity of genuine media is crucial. This can be achieved through digital watermarking, where invisible signals are embedded into authentic media for future verification. An alternative method involves controlled capture, where unique statistical features (signatures) of real media are extracted and securely stored, possibly using blockchain technology. These signatures can later be used to verify the authenticity of media against copies or modifications.

Such a comprehensive approach tackles the technical aspects of combating DeepFakes and considers the broader ethical and legal implications of synthetic media. This multifaceted strategy is essential for effectively countering DeepFakes’ negative impacts.

Please Note: This is a Commercial Profile