Lancashire Cyber Foundry

An Introduction to Deepfakes

READ TIME: 2 MINS

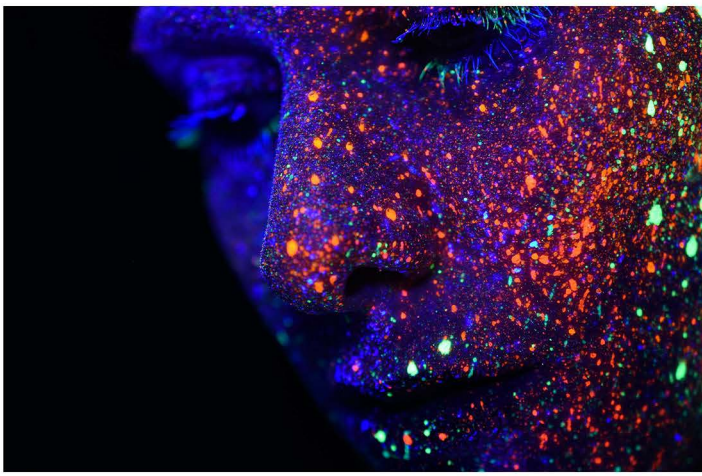
AUDIENCE: BUSINESS & TECHNOLOGY

Deepfakes are a form of synthetic media generated by using artificial intelligence / machine learning algorithms.

They can manipulate original content, typically a person's face or body, with the intention of giving the impression that the person may have said or done something they haven't. Hence, it has the potential to be used maliciously to damage a person's reputation, for blackmail, or to spread false information.

The technology combines computer vision and machine learning techniques in a Generative Adversarial Network (GAN) architecture. This consists of two neural networks, a generator and a discriminator. The two neural networks are trained simultaneously in a supervised learning setting where they essentially "compete" against each other in a zero-sum game. Put more specifically, this is where the gain of one model is exactly balanced by the loss of the other. The generator network produces new data samples which are images or videos, while the discriminator evaluates the performance of the samples in terms of how realistic they are.

The aim of the generator is to produce realistic data such that the discriminator is unable to identify it as fake, and the discriminator aims to improve its accuracy for detecting fake images / videos. For example, if the generator creates a sample that is judged as fake by the discriminator, this is a loss for the generator but a win for the discriminator. conversely, the generator wins and the discriminator loses if the generator creates a sample that is judged as real by the discriminator. Therefore, both models improve after each "game" and as



European Union
European Regional Development Fund

NORTHERN POWERHOUSE HM Government

Lancaster University

a result, high-quality data samples resembling of the real world are produced.

The purpose of GANs is to produce synthetic data in applications where real data is limited. This may include image or video synthesis aka "deepfakes", text-to-image synthesis, and artwork or music in the style of existing artists or to a set of training images or songs. However, such "deepfakes" when unregulated and undetected by high performing discriminators can be highly impactful to the lives of individuals within the population, celebrities, large companies or political parties. Further examples include:

- Political propaganda and disinformation.
- Spreading disinformation in media with the aim to cause public panic or confusion.
- Fake pornographic content featuring celebrities or other individuals without their consent.
- Blackmail and extortion, where a deepfake video may be used to demand payment.

In summary, the application of AI and GANs to produce "deepfakes" raises many ethical and security concerns about its potential impact on society. Methods to prevent malicious use of these must be developed to prevent significant harm. This may include the application of visual or audio analysis, deep learning, and forensic tools. However, it should be noted that both producing and detecting deepfakes requires large amount of training data and computing power, which may not be widely accessible to some people. Consequently, it is critical that such detection methods are developed and perfected before deepfakes become a bigger threat than they already are.

JOSH BAKER - LCF INTERN

ABOUT US

Lancashire Cyber Foundry

The Lancashire Cyber Foundry runs a programme designed to s upport businesses facing cyber challenges in Lancashire. Digital Innovation support is part of this programme but there is also business strategy support available which includes specialised workshops to help businesses innovate and grow.

To find out more visit our website, https://www.lancashirecyberfoundry.co.uk/ or email us at;

cyberfoundry@lancaster.ac.uk