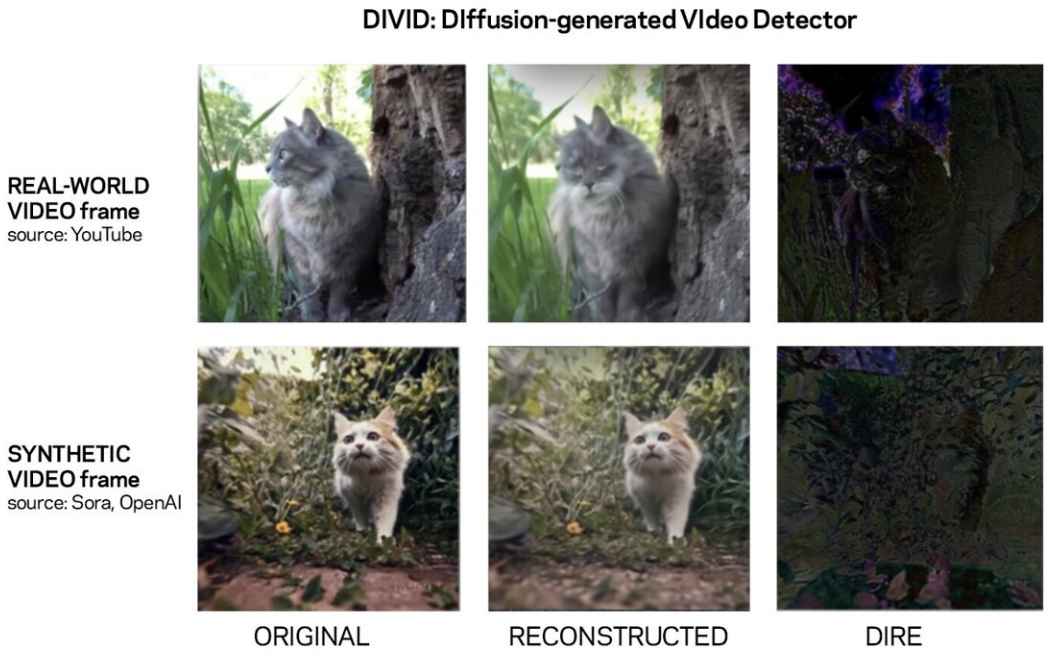
New tool detects AI-generated videos with 93.7% accuracy

June 26 2024, by Bernadette Young



Pictured: first column: Video frames taken from YouTube and fake videos generated from Sora by OpenAI; second column: frames reconstructed by diffusion; third column: the differences between the first and the second columns. As illustrated, the real-world video frames differ more from their diffusion-reconstructed frames than diffusion-generated video, a key insight for DIVID to detect diffusion-generated video. DIRE (DIffusion Reconstruction Error) is a method that measures the difference between an input image and the corresponding output image reconstructed by a pretrained diffusion model. Credit: Software Systems Laboratory/Columbia Engineering

Earlier this year, an employee at a multinational corporation sent fraudsters $25 million. The instructions to transfer the money came—the employee thought—straight from the company's CFO. In reality, the criminals had used an AI program to generate realistic videos of the CFO and several other colleagues in an elaborate scheme.

Videos created by AI have become so realistic that humans (and existing detection systems) struggle to distinguish between real and fake videos. To address this problem, Columbia Engineering researchers, led by Computer Science Professor Junfeng Yang, have developed a new tool to detect AI-generated video called DIVID, short for DIffusion-generated VIdeo Detector. DIVID expands on work the team released earlier this year–Raidar, which detects AI-generated text by analyzing the text itself, without needing to access the inner workings of large language models.

A paper on the new tool appears on the *arXiv* preprint server.

DIVID detects a new generation of generative AI videos

DIVID improves upon earlier existing methods that detect generative videos that effectively identify videos generated by older AI models like generative adversarial networks (GAN). A GAN is an AI system with two neural networks: One creates fake data, and another evaluates it to distinguish between fake and real. Through continuous feedback, both networks improve, resulting in a highly realistic synthetic video. Current AI detection tools look for telltale signs like unusual pixel arrangements, unnatural movements, or inconsistencies between frames that wouldn't typically occur in real videos.

The new generation of generative AI video tools, like Sora by OpenAI,

Runway Gen-2, and Pika, uses a diffusion model to create videos. A diffusion model is an AI technique that creates images and videos by gradually turning random noise into a clear, realistic picture. For videos, it refines each frame individually while ensuring smooth transitions, producing high-quality, lifelike results. This increasing sophistication of AI-generated videos poses a significant challenge in detecting their authenticity.

Yang's group used a technique called DIRE (DIffusion Reconstruction Error) to detect diffusion-generated images. DIRE is a method that measures the difference between an input image and the corresponding output image reconstructed by a pretrained diffusion model.

Expanding Raidar's AI-generated texts to video

Yang, who co-directs the Software Systems Lab, has been exploring how to detect AI-generated text and videos. Earlier this year, with the release of Raidar, Yang and collaborators are enabling a way to detect AI-generated text by analyzing the text itself, without needing to access the inner workings of large language models like chatGPT-4, Gemini, or Llama. Raidar uses a language model to rephrase or alter a given text and then measures how many edits the system makes to the given text. Many edits mean humans likely wrote the text, while fewer modifications mean the text is likely machine-generated.

"The insight in Raidar—that the output from an AI is often considered high-quality by another AI so it will make fewer edits—is really powerful and extends beyond just text," said Yang. "Given that AI-generated video is becoming more and more realistic, we wanted to take the Raidar insight and create a tool that can detect AI-generated videos accurately."

The researchers used the same concept to develop DIVID. This new

generative video detection method can identify video generated by diffusion models. The research paper, which includes open-sourced code and datasets, was presented at the Computer Vision and Pattern Recognition Conference (CVPR) in Seattle on June 18, 2024.

How DIVID works

DIVID works by reconstructing a video and analyzing the newly reconstructed video against the original video.It uses DIRE values to detect diffusion-generated videos since the method operates on the hypothesis that reconstructed images generated by diffusion models should closely resemble each other because they are sampled from the diffusion process distribution. If there are significant alterations, the original video is likely human-generated. If not, it is likely AI-generated.

The framework is based on the idea that AI generation tools create content based on the statistical distribution of large data sets, resulting in more "statistical means" content such as pixel intensity distributions, texture patterns, and noise characteristics in video frames, subtle inconsistencies or artifacts that change unnaturally between frames, or unusual patterns that are more likely in diffusion-generated videos than in real ones.

In contrast, human video creations exhibit individuality and deviate from the statistical norm. DIVID achieved a groundbreaking detection accuracy of up to 93.7% for videos from their benchmark dataset of diffusion-generated videos from Stable Vision Diffusion, Sora, Pika, and Gen-2.

For now, DIVID is a command line tool that analyzes a video and outputs whether it is AI or human-generated and can only be used by developers. The researchers note that their technology has the potential to be integrated as a plugin to Zoom to detect deepfake calls in real time.

The team is also considering developing a website or browser plugin to make DIVID accessible to ordinary users.

"Our framework is a significant leap forward in detecting AI-generated content," said Yun-Yun Tsai, one of the authors of the paper and a Ph.D. student of Yang. "There are way too many scammers who use AI-generated video, and it's critical to stop them and protect society."

What's next?

The researchers are now working to improve the framework of DIVID so it can handle different kinds of synthetic videos from open-source video generation tools. They are also using DIVID to collect videos for the DIVID dataset.

More information: Qingyuan Liu et al, Turns Out I'm Not Real: Towards Robust Detection of AI-Generated Videos, *arXiv* (2024). DOI: 10.48550/arxiv.2406.09601

Provided by Columbia University School of Engineering and Applied Science