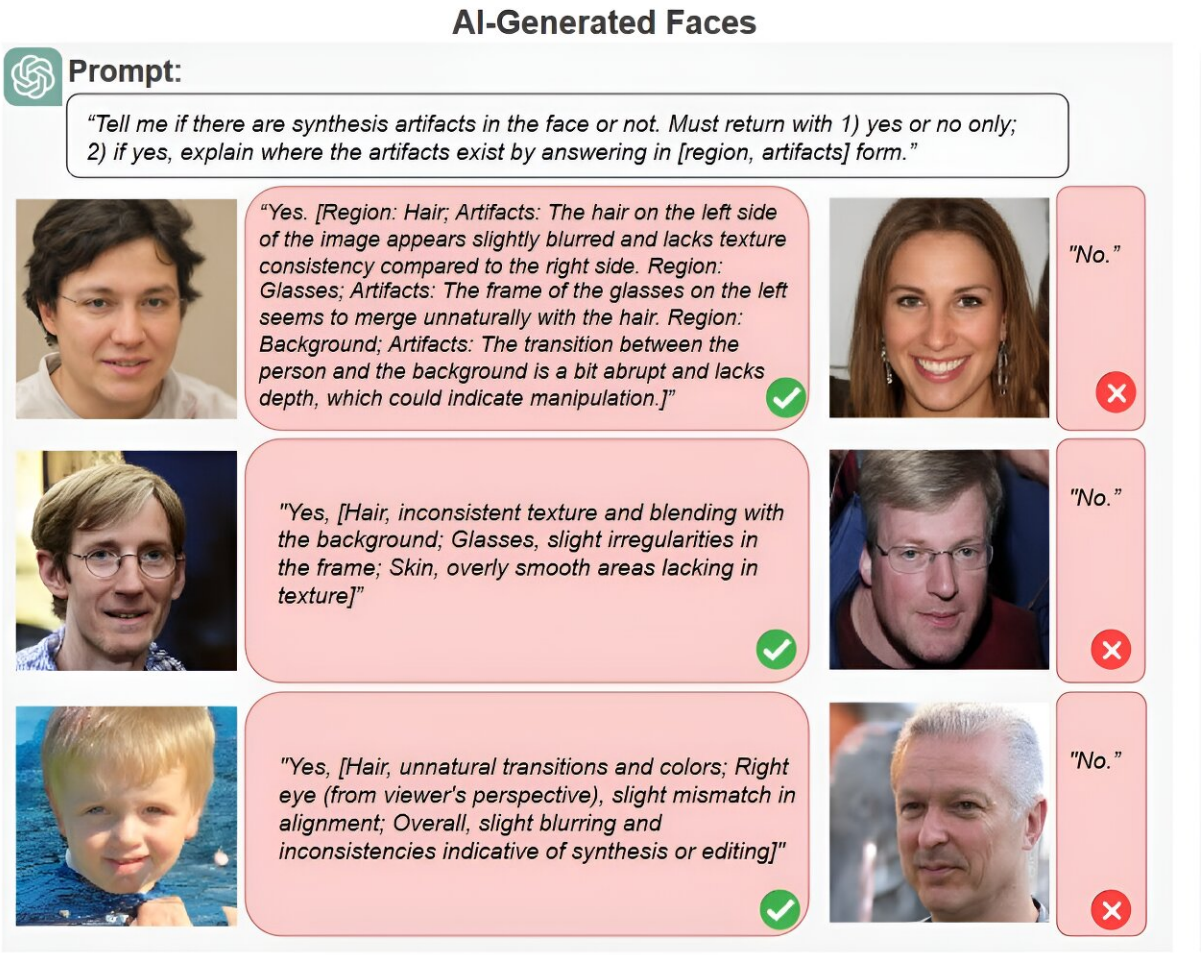
Is ChatGPT the key to stopping deepfakes? Study asks LLMs to spot AI-generated images

June 27 2024, by Tom Dinki



An example of ChatGPT's analysis of deepfake images. The large language model was less accurate than state-of-the-art deepfake detectors, but impressed researchers with its ability to explain its analysis in plain language. Credit:

When most people think of artificial intelligence, they're probably thinking of—and worrying about—ChatGPT and deepfakes. AI-generated text and images dominate our social media feeds and the other websites we visit, sometimes without us knowing it, and are often used to spread unreliable and misleading information.

But what if text-generating models like ChatGPT could actually spot deepfake images?

A University at Buffalo-led research team has applied large language models (LLMs), including OpenAI's ChatGPT and Google's Gemini, toward spotting deepfakes of human faces. Their study, presented last week at the IEEE/CVF Conference on Computer Vision & Pattern Recognition, found that LLMs' performance lagged behind that of state-of-the-art deepfake detection algorithms, but their natural language processing may actually make them the more practical detection tool in the future.

The study is also published on the *arXiv* preprint server.

"What sets LLMs apart from existing detection methods is the ability to explain their findings in a way that's comprehensible to humans, like identifying an incorrect shadow or a mismatched pair of earrings," says the study's lead author, Siwei Lyu, Ph.D., SUNY Empire Innovation Professor in the Department of Computer Science and Engineering, within the UB School of Engineering and Applied Sciences. "LLMs were not designed or trained for deepfake detection, but their semantic knowledge makes them well suited for it, so we expect to see more efforts toward this application."

Collaborators on the study include the University at Albany and the Chinese University of Hong Kong, Shenzhen.

How language models understand images

Trained on much of the available text on the internet—amounting to some 300 billion words—ChatGPT finds statistical patterns and relationships between words in order to generate responses.

The latest versions of ChatGPT and other LLMs can also analyze images. These multimodal LLMs use large databases of captioned photos to find the relationships between words and images.

"Humans do this as well. Whether it be a stop sign or a viral meme, we constantly assign a semantic description to images," says the study's first author, Shan Jai, assistant lab director in the UB Media Forensic Lab. "In this way, images become their own language."

The Media Forensics Lab team decided to test if GPT-4 with vision (GPT-4V) and Gemini 1.0 could tell the difference between real faces and faces generated by AI. They gave it thousands of images of both real and deepfake faces and asked it to identify any potential signs of manipulation, or synthetic artifacts.

ChatGPT advantages

ChatGPT was accurate 79.5% of the time in detecting synthetic artifacts in images generated by latent diffusion, and 77.2% of the time on StyleGAN-generated images.

"This is comparable to earlier deepfake detection methods, so with proper prompt guidance, ChatGPT can do a fairly decent job at

detecting AI-generated images," says Lyu, who is also co-director of UB's Center for Information Integrity.

More crucially, ChatGPT could explain its decision making in plain language. When provided an AI-generated photo of a man with glasses, the model correctly pointed out that "the hair on the left side of the image slightly blurs" and "the transition between the person and the background is a bit abrupt and lacks depth."

"Existing deepfake detection models will tell us the probability of an image being real or fake, but they will very rarely tell us why they came to this conclusion. And even if we look into the model's underlying mechanisms, there will be features that we simply can't understand," Lyu says. "Meanwhile, everything ChatGPT outputs is understandable to humans."

That's because ChatGPT bases its analysis on semantic knowledge alone. Whereas traditional deepfake detection algorithms distinguish real from fake by training on large datasets of images labeled real or fake, LLMs' natural language abilities give them something of a common sense understanding of reality—at least when they're not hallucinating—including the typical symmetry of human faces and the appearance of real photographs.

"Once the vision component of ChatGPT understands an image as a human face, the language component can make the inference that a face will typically have two eyes, and so on," Lyu says. "The language component provides a deeper connection between visual and verbal concepts."

ChatGPT's semantic knowledge and natural language processing make it a more user-friendly deepfake tool for both users and developers, the study concluded.

"Typically, we take insights about detecting deepfakes and convert them into programming language. Now, all this knowledge is present within a single model and we need only use natural language to bring out that knowledge," Lyu says.

ChatGPT drawbacks

ChatGPT's performance was well below the latest deepfake detection algorithms, which have accuracy rates in the mid- to high-90s.

This was partly because LLMs can't catch signal-level statistical differences that are invisible to the human eye but often used by detection algorithms to spot AI-generated images.

"ChatGPT focused only on semantic-level abnormalities," Lyu says. "In this way, the semantic intuitiveness of the ChatGPT's results may actually be a double-edged sword for deepfake detection."

And other LLMs may not be as effective at explaining their analysis. Despite performing comparatively to ChatGPT at guessing the presence of synthetic artifacts, Gemini's supporting evidence was often nonsensical, like pointing out nonexistent moles.

Another drawback is that LLMs often refused to analyze images. When asked directly whether a photo was generated by AI, ChatGPT typically replied with, "Sorry, I can't assist with that request."

"The model is programmed not to answer when it doesn't reach a certain confidence level," Lyu says. "We know that ChatGPT has information relevant to deepfake detection, but again, a human operator is needed to excite that part of its knowledge base. Prompt engineering is effective, but not very efficient, so the next step is going one level down and actually fine tuning LLMs for this task specifically."

More information: Shan Jia et al, Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics, *arXiv* (2024). DOI: 10.48550/arxiv.2403.14077

Provided by University at Buffalo