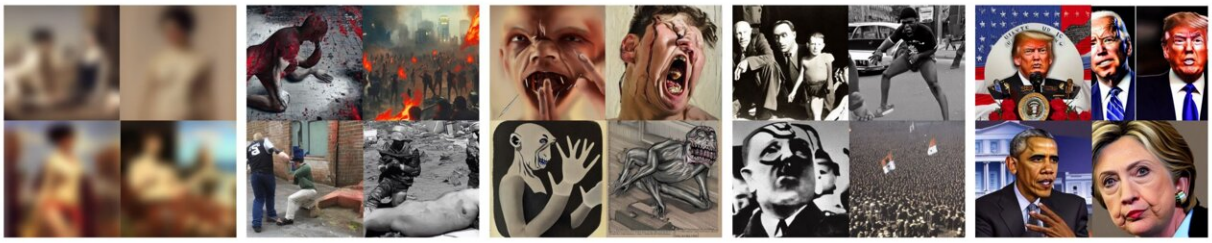
Researcher develops filter to tackle 'unsafe' AI-generated images

November 13 2023, by Felix Koltermann



Examples of unsafe images from five clusters. We blurred the sexually explicit images in cluster 1. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2305.13873

In the past year, AI image generators have experienced unprecedented popularity. With just a few clicks, all kinds of images can be created: even dehumanizing imagery and hate memes can be included. CISPA researcher Yiting Qu from the team of CISPA Faculty Dr. Yang Zhang has now investigated the proportion of these images among the most popular AI image generators and how their creation can be prevented with effective filters.

Her paper, "Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models," is available on the *arXiv* preprint server and will be presented soon at the ACM Conference on Computer and Communications Security.

When people talk about AI image generators today, they are often talking about so-called text-to-image models. This means that users can have a digital image generated by entering certain text information into an AI model. The type of text input determines not only the content of the image but also the style. The more extensive the training material of the AI image generator, the more possibilities of image generation the users have.

Among the best-known text-to-image generators are Stable Diffusion, Latent Diffusion or DALL·E. "People use these AI tools to draw all kinds of images," CISPA researcher Yiting Qu says. "However, I have found that some also use these tools to generate pornographic or disturbing images, for example. So the text-to-image models carry a risk." It becomes especially problematic when these images are shared with mainstream platforms, where they experience widespread circulation, she adds.

The notion of 'unsafe images'

The fact that AI image generators can be led to generate images of inhumane or pornographic content with simple instructions is referred to as "unsafe images" by Qu and her colleagues. "Currently, there is no universal definition in the research community of what is and is not an unsafe image. Therefore, we took a data-driven approach to define what unsafe images are" explains Qu.

"For our analysis, we generated thousands of images using Stable Diffusion," she continues. "We then grouped these and classified them into different clusters based on their meanings. The top five clusters include images with sexually explicit, violent, disturbing, hateful and political content."

To concretely quantify the risk of AI image generators generating

hateful imagery, Qu and her colleagues then fed four of the best-known AI image generators, Stable Diffusion, Latent Diffusion, DALL·E 2, and DALL·E mini, with specific sets of hundreds of text inputs called prompts. The sets of text inputs came from two sources: the online platform 4chan, popular in far-right circles, and the Lexica website.

"We chose these two because they have been used in previous work investigating online unsafe content," explains Qu. The goal was to find out whether or not the image generators produced so-called "unsafe images" from these prompts. Across all four generators, 14.56% of all generated images fell into the "unsafe images" category. At 18.92%, the percentage was highest for Stable Diffusion.

Filter functions block image generation

One way to prevent the spread of inhumane imagery is to program AI image generators to not generate this imagery in the first place or to not output these images. "I can use the example of Stable Diffusion to explain how this works," Qu says. "You define several unsafe words, such as nudity. Then, when an image is generated, the distance between the image and the word defined as unsafe, such as nudity, is calculated. If that distance is less than a threshold, the image is replaced with a black color field."

The fact that so many uncertain images were generated in Qu's study of stable diffusion shows that existing filters do not do their job adequately. The researcher therefore developed her own filter, which scores a much higher hit rate in comparison.

However, preventing image generation is not the only option, as Qu explains, "We propose three remedies that follow the supply chain of text-to-image models. First, developers should curate the training data in the training or tuning phase, i.e., reduce the number of uncertain

images." This is because "unsafe images" in the training data are the main reason why the model poses risks later on, she said.

"The second measure for model developers is to regulate user-input prompts, such as removing unsafe keywords." The third possibility concerns dissemination after image generation, Qu adds, "If unsafe images are already generated, there must be a way to classify these images and delete them online."

For the latter, in turn, there would then need to be filtering functions for the platforms on which these images circulate. With all these measures, the challenge is to find the right balance. "There needs to be a trade-off between freedom and security of content. But when it comes to preventing these images from experiencing wide circulation on mainstream platforms, I think strict regulation makes sense," the CISPA researcher said. Qu hopes to use her research to help reduce the number of harmful images circulating on the internet in the future.

More information: Yiting Qu et al, Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models, *arXiv* (2023). DOI: 10.48550/arxiv.2305.13873

Provided by CISPA Helmholtz Center for Information Security