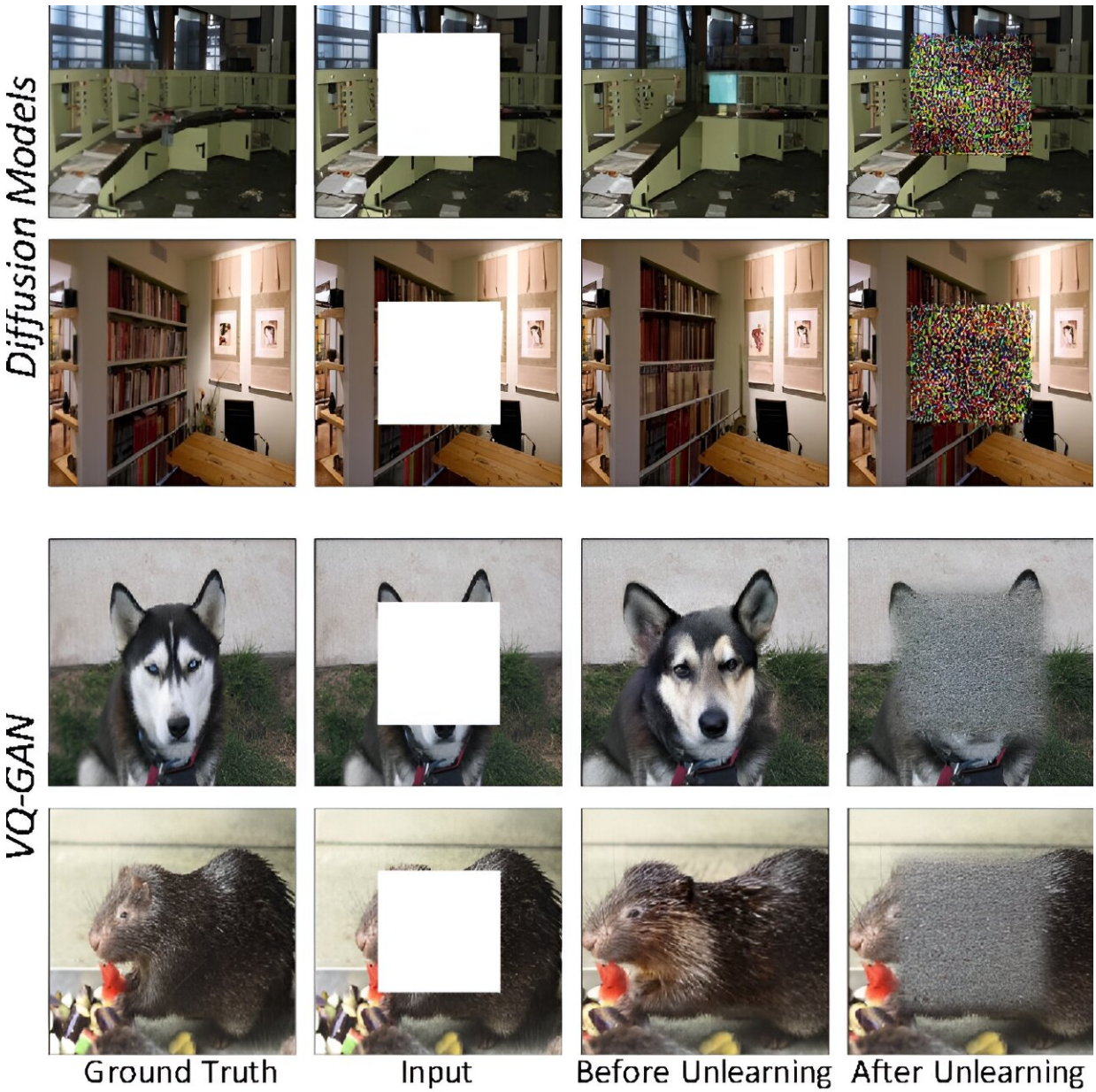
Machine 'unlearning' helps generative AI forget copyright-protected and violent content

March 22 2024



Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.00351

When people learn things they should not know, getting them to forget that information can be tough. This is also true of rapidly growing artificial intelligence programs that are trained to think as we do, and it has become a problem as they run into challenges based on the use of copyright-protected material and privacy issues.

To respond to this challenge, researchers at The University of Texas at Austin have developed what they believe is the first "machine unlearning" method applied to image-based generative AI. This method offers the ability to look under the hood and actively block and remove any violent images or copyrighted works without losing the rest of the information in the model. The study is published on the *arXiv* preprint server.

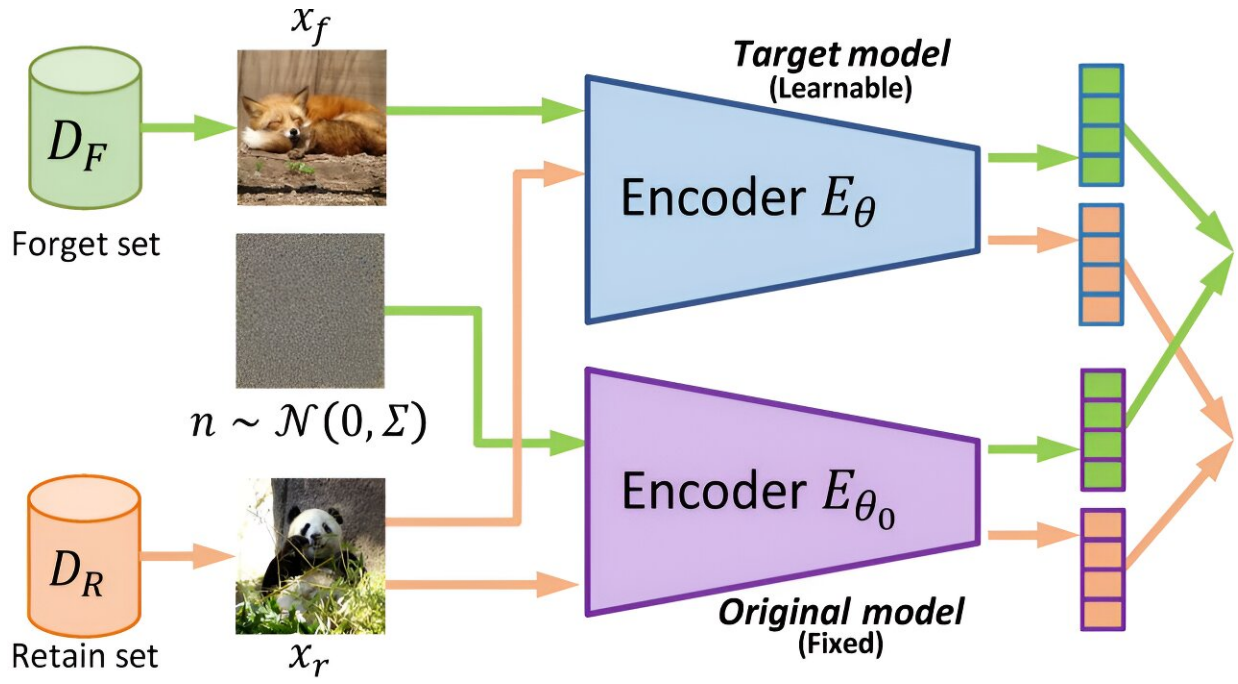
"When you train these models on such massive data sets, you're bound to include some data that is undesirable," said Radu Marculescu, a professor in the Cockrell School of Engineering's Chandra Family Department of Electrical and Computer Engineering and one of the leaders on the project.

"Previously, the only way to remove problematic content was to scrap everything, start anew, manually take out all that data and retrain the model. Our approach offers the opportunity to do this without having to retrain the model from scratch."

Generative AI models are trained primarily with data on the internet because of the unrivaled amount of information it contains. But it also

contains massive amounts of data that is protected by copyright, in addition to personal information and inappropriate content.

Underscoring this issue, The New York Times recently sued OpenAI, maker of ChatGPT, arguing that the AI company illegally used its articles as training data to help its chatbots generate content.



Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.00351

"If we want to make generative AI models useful for commercial purposes, this is a step we need to build in, the ability to ensure that we're not breaking copyright laws or abusing personal information or using harmful content," said Guihong Li, a graduate research assistant in Marculescu's lab who worked on the project as an intern at JPMorgan Chase and finalized it at UT.

Image-to-image models are the primary focus of this research. They take an input image and transform it—such as creating a sketch, changing a particular scene and more—based on a given context or instruction.

This new machine unlearning algorithm provides the ability of a machine learning model to "forget" or remove content if it is flagged for any reason without the need for retraining the model from scratch. Human teams handle the moderation and removal of content, providing an extra check on the model and ability to respond to user feedback.

Machine unlearning is an evolving branch of the field that has been primarily applied to classification models. Those models are trained to sort data into different categories, such as whether an image shows a dog or a cat.

Applying machine unlearning to generative models is "relatively unexplored," the researchers write in the paper, especially when it comes to images.

More information: Guihong Li et al, Machine Unlearning for Image-to-Image Generative Models, *arXiv* (2024). DOI: 10.48550/arxiv.2402.00351

Provided by University of Texas at Austin