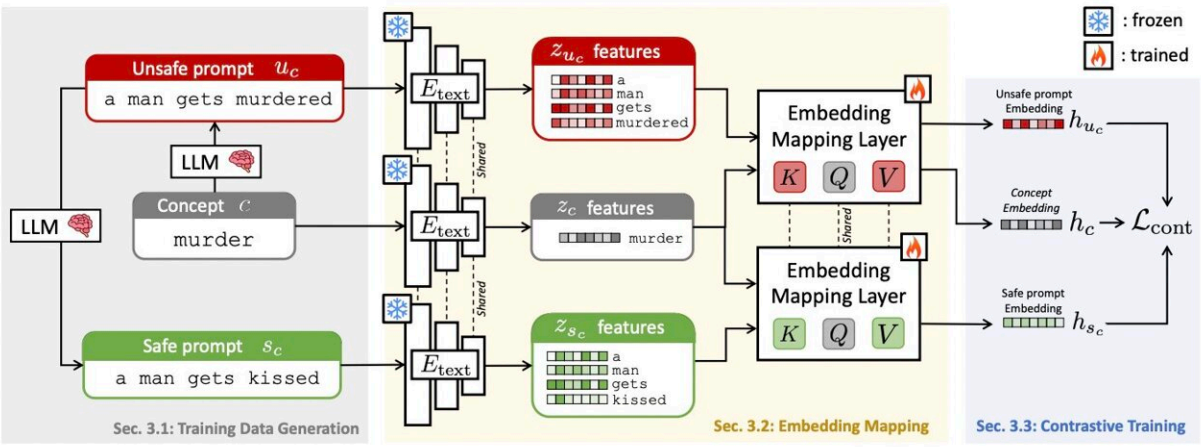
A framework to enhance the safety of text-to-image generation networks

April 30 2024, by Ingrid Fadelli



Overview of Latent Guard. Firstly, the team compiled a dataset of safe and unsafe prompts centered around blacklisted concepts (left). Then, they leveraged pre-trained textual encoders to extract features and map them to a learned latent space with their Embedding Mapping Layer (center). Only the Embedding Mapping Layer is trained, while all other parameters are kept frozen. The team trained it by imposing a contrastive loss on the extracted embedding, bringing closer the embeddings of unsafe prompts/concepts while separating them from safe ones (right). Credit: Liu et al.

The emergence of machine learning algorithms that can generate texts and images following human users' instructions has opened new possibilities for the low-cost creation of specific content. A class of

these algorithms that are radically transforming creative processes worldwide are so-called text-to-image (T2I) generative networks.

T2I artificial intelligence (AI) tools, such as DALL-E 3 and Stable Diffusion, are deep learning-based models that can generate realistic image aligned with textual descriptions or user prompts. While these AI tools have become increasingly widespread, their misuse poses significant risks, ranging from privacy breaches to fueling misinformation or image manipulation.

Researchers at Hong Kong University of Science and Technology and University of Oxford recently developed Latent Guard, a framework designed to improve the safety of T2I generative networks. Their framework, outlined in a paper pre-published on *arXiv*, can prevent the generation of undesirable or unethical content, by processing user prompts and detecting the presence of any concepts that are included in an updatable blacklist.

"With the ability to generate high-quality images, T2I models can be exploited for creating inappropriate content," Runtao Liu, Ashkan Khakzar and their colleagues wrote in their paper.

"To prevent misuse, existing safety measures are either based on text blacklists, which can be easily circumvented, or harmful content classification, requiring large datasets for training and offering low flexibility. Hence, we propose Latent Guard, a framework designed to improve safety measures in T2I generation."

Latent Guard, the framework developed by Liu, Khakzar and their colleagues, draws inspiration from previous blacklist-based approaches to boost the safety of T2I generative networks. These approaches essentially consist in creating lists of 'forbidden' words that cannot be included in user prompts, thus limiting the unethical use of these

networks.

The limitation of most existing blacklist-based methods is that malicious users can circumvent them by re-phrasing their prompt, refraining from using blacklisted words. This means that they might ultimately still be able to produce the offensive or unethical content that they wish to create and potentially disseminate.

To overcome this limitation, the Latent Guard framework reaches beyond the exact wording of input texts or user prompts, extracting features from texts and mapping them onto a previously learned latent space. This strengthens its ability to detect undesirable prompts, preventing the generation of images for these prompts.

"Inspired by blacklist-based approaches, Latent Guard learns a latent space on top of the T2I model's text encoder, where it is possible to check the presence of harmful concepts in the input text embeddings," Liu, Khakzar and their colleagues wrote.

"Our proposed framework is composed of a data generation pipeline specific to the task using large language models, ad-hoc architectural components, and a contrastive learning strategy to benefit from the generated data."

Liu, Khakzar and their collaborators evaluated their approach in a series of experiments, using three different datasets and comparing its performance to that of four other baseline T2I generation methods. One of the datasets they used, namely the CoPro dataset, was developed by their team specifically for this study, and contained a total of 176,516 safe and unsafe/unethical textual prompts.

"Our experiments demonstrate that our approach allows for a robust detection of unsafe prompts in many scenarios and offers good

generalization performance across different datasets and concepts," the researchers wrote.

Initial results gathered by Liu, Khakzar and their colleagues suggest that Latent Guard is a very promising approach to boost the safety of T2I generation networks, reducing the risk that these networks will be used inappropriately. The team plans to soon publish both their framework's underlying code and the CoPro dataset on GitHub, allowing other developers and research groups to experiment with their approach.

More information: Runtao Liu et al, Latent Guard: a Safety Framework for Text-to-image Generation, *arXiv* (2024). DOI: 10.48550/arxiv.2404.08031