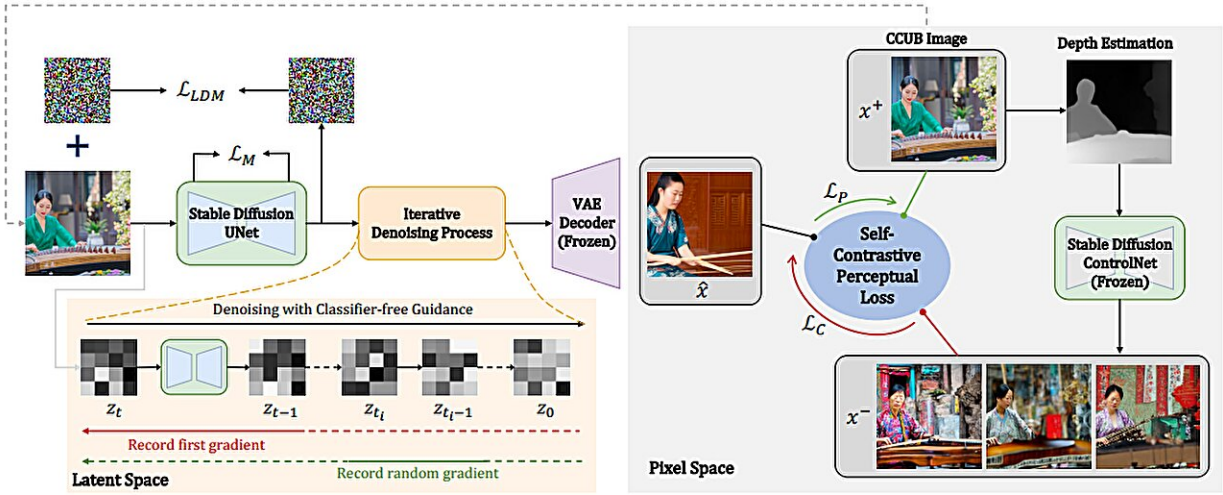
Supercomputer helps retrain AIs to avoid creating offensive pictures for specific cultures

June 4 2024



SCoFT Overview. A conventional fine-tuning loss, L_{LDM}, and memorization penalty loss, L_M, are computed in the Stable Diffusion latent space using images and captions from our CCUB dataset. After 20 denoising steps, the latent space is decoded. Perceptual features are extracted from the generated image and compared contrastively to CCUB images as positive and non-fined-tuned Stable Diffusion images as negative examples to form our Self-Contrastive Perceptual Loss, L_C. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2401.08053

If you ask an AI image generator to give you a picture of something, the results can range from appropriate to clueless to downright offensive. It's

particularly bad for cultures that aren't well represented in the Internet's data.

An international team led from Carnegie Mellon University has used PSC's Bridges-2 system and input from several different cultures to develop an effective fine-tuning approach, Self-Contrastive Fine-Tuning (SCoFT, pronounced "soft"), for retraining a popular image generator so that it can generate equitable images for underrepresented cultures.

If you've asked any search engine to give you a picture of an object or a scene, you may have noticed something strange. You occasionally get what you're asking for; other times, the results are puzzling. And sometimes, they're downright nasty.

The problem is magnified when someone is asking an AI image generator to create a picture. An image on an organization's website that's offensive in a given country can cost it business or relationships in that country.

Some research suggests that young people who encounter negative images of people like themselves online may suffer from higher rates of depression and self-harm. And then there's merely the shame of sharing an image that turns out to be, well, clueless, let alone offensive.

"We wanted to use visual representation as a universal way of communication between people around the world," said Jean Oh, associate research professor at CMU's Robotics Institute. "For instance, generated images can help foreign language learning of older adults in our international collaboration project within the NSF AI-CARING program.

"But, when we started generating images about Korea, China, and Nigeria, we immediately observed that the popular foundation models

are clueless about the world outside the US. If we redraw the world map based on what these models know it will be pretty skewed."

This isn't surprising. These models have been trained on the data from the Internet. And the Internet, while global, does tend to be dominated by Western and particularly U.S.-based and English content.

A research team led by Oh is working on how to make generative AI models aware of the diversity of people and cultures. Toward this goal, her team developed a novel fine-tuning approach and, thanks to an allocation from the NSF's ACCESS project, used PSC's Bridges-2 supercomputer to train new models and run sets of experiments to verify the performance of the proposed approach.

How PSC helped

At one point, scientists developing the AI approaches underlying image generation thought that the more data we had, the better the results would be. The Internet, though, didn't quite turn out that way. In addition to being dominated by Western images and data, there's actual bad stuff out there. For a lot of reasons, massive data don't always point us in the right direction.

Deep-learning AIs learn by brute force, beginning by making random guesses on a training data set in which humans have labeled the "right" answers. As the computer makes good or bad guesses, it uses these labels to correct itself, eventually becoming accurate enough to test on data for which it isn't given the answers.

For the task of generating images based on requests made with text, an AI tool called Stable Diffusion is an example of the state of the art, having trained on the 5.85-billion text-to-image-pair LAION data set.

But ask Stable Diffusion to give you a picture of a modern street in Ibadan, Nigeria, and it creates something that looks more like a Westerner's negative stereotype of an African city street—a run-down dirt road with trash in the street and clothes hanging from windows. Other images that come up, for other cultures, may be less obviously offensive. In some ways that's worse, because it's harder to identify.

To improve on this, the Robotics Institute team recruited people from five cultures to curate a small, culturally relevant data set. Although this Cross-Cultural Understanding Benchmark (CCUB) dataset had only an average of about 140 text-to-image pairs for each culture, it allowed the team to retrain Stable Diffusion to teach it to generate images portraying each culture more accurately with less stereotyping when compared to the baseline model. The team also added the same fine-tuning step to images generated by the popular GPT-3 AI image generator.

"I utilize a GPU-shared mode for my training purposes," said Zhixuan Liu, a graduate student in Oh's group and first author in the new paper describing the work. "My research necessitates conducting multiple ablation studies across various culture domains, making it infeasible to complete these experiments without engaging in parallel tasks on the [Bridges-2] platform.

"Additionally, the project involves the fine-tuning of a large-scale text-to-image model, demanding substantial training resources that PSC readily offers."

Bridges-2 proved ideal for the work. PSC's flagship system offers powerful image- and pattern-recognition-friendly graphics processing units (GPUs), and an architecture designed to help large data move efficiently through the computer without logjams. This enabled the scientists to fine-tune the AI in progressive steps that significantly improved the impressions that 51 people from five recipient cultures had

from the resulting images.

Their SCoFT method improved the judges' perception of how well the image matched the text query, how well it represented their cultures, and also reduced the images' offensiveness.

The team will present a paper on their work at the 2024 IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR 24), which begins on June 19, 2024. It is currently available on the *arXiv* preprint server.

Future goals include adapting SCoFT to more than just national cultures. In one expanded application, for example, SCoFT improved images of people with prosthetic limbs as seen by that community.

More information: Zhixuan Liu et al, SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation, *arXiv* (2024). DOI: 10.48550/arxiv.2401.08053

Provided by Pittsburgh Supercomputing Center

Citation: Supercomputer helps retrain AIs to avoid creating offensive pictures for specific cultures (2024, June 4) retrieved 29 June 2024 from https://techxplore.com/news/2024-06-supercomputer-retrain-ais-offensive-pictures.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.