New tool finds bias in state-of-the-art generative AI model

August 10 2023, by Emily Cerf

Examples of images generated by text prompts imputed to the Stable Diffusion model with and without gender-specific language in the prompt. For example, the upper left group of four images were produced from the prompt "child studying science." Credit: Xin Wang, via Stable Diffusion

Text-to-image (T2I) generative artificial intelligence tools are

increasingly powerful and widespread tools that can create nearly any image based on just a few inputted words. T2I generative AI can create convincingly realistic photos and videos, which are being used more and more for a multitude of purposes, from art to political campaigning.

However, the algorithmic models that power these tools are trained on data from humans, and can replicate human biases in the images they produce, such as biases around gender and skin tone. These biases can harm marginalized populations, reinforcing stereotypes and potentially leading to discrimination.

To address these implicit biases, Assistant Professor of Computer Science and Engineering Xin (Eric) Wang and a team of researchers from Baskin Engineering at UC Santa Cruz created a tool called the Text to Image Association Test, which provides a quantitative measurement of complex human biases embedded in T2I models, evaluating biases across dimensions such as gender, race, career, and religion. They used this tool to identify and quantify bias in the state-of-the-art generative model Stable Diffusion.

The tool is detailed in a paper for the 2023 Association for Computational Linguistics (ACL) conference, and is available for use in a demo version.

"I think both the model owners and users care about this issue," said Jialu Wang, a UCSC computer science and engineering Ph.D. student and the first author on the paper. "If the user is from an unprivileged group, they may not want to see just the privileged group reflected in the images they generate."

To use the tool, a user must tell the model to produce an image for a neutral prompt, for example "child studying science." Next, the user inputs gender specific prompts, such as "girl studying science" and "boy

studying science." Then, the tool calculates the distance between the images generated with the neutral prompt and each of the specific prompts. That difference between those two distances is a quantitative measurement of bias.

Using their tool, the research team found that the state-of-the-art generative model Stable Diffusion both replicates and amplifies human biases in the images it produces. The tool tests the association between two concepts, such as science and arts, to two attributes, such as male and female. It then gives an association score between the concept and the attribute and a value to indicate how confident the tool is in that score.

The team used their tool to test whether the model associates six sets of opposing concepts with positive or negative attributes. The concepts they tested were: flowers and insects, musical instruments and weapons, European American and African American, light skin and dark skin, straight and gay, and Judaism and Christianity. For the most part, the model made associations along stereotypical patterns. However, the model associated dark skin as pleasant and light skin as unpleasant, which surprised researchers as one of the few results in opposition to common stereotypes.

Additionally, they found that the model associated science more closely with males and art more closely with females, and associated careers more closely with males and family more closely with females.

In the past, techniques for evaluating bias in T2I models required researchers to annotate results received from the models when entering a neutral prompt. For example, a researcher might enter a gender neutral prompt such as "child studying science" and label whether the model produces images of boys versus girls. But the labor that goes into this annotation process is costly and could potentially be inaccurate, and is

often constricted to just gender biases.

"We want to get rid of this human annotating process and propose an automatic tool to evaluate those biases, without the tedious laboring," Xin Wang said.

Additionally, unlike others, the UCSC team's bias evaluation tool considers aspects of the background of the image such as the colors and warmth.

The researchers based their tool on the Implicit Association Test, a well-known test in social psychology used for evaluating human biases and stereotypes. This test evaluates how closely people associate concepts such as "doctors" or "family" with attributes such as "men" or "women."

Beyond evaluating and analyzing biases in existing tools such as Stable Diffusion and Midjourney, the team envisions that the tool will allow software engineers to get more accurate measurements of biases in their models while in the development phase and track their efforts to address those biases.

"With a quantitative measurement, people can work on mitigating those biases and use our tool to quantify their progress in doing so," Xin Wang said.

The team said they received plenty of positive feedback from other researchers when presenting this work at the ACL conference.

"Many in the community showed a great interest in this work," Xin Wang said. "Some researchers immediately shared this work within their groups and asked me for the details."

Going forward, the team plans to propose suggested methods to mitigate

these biases, both in training new models from scratch, or to de-bias existing models during fine-tuning.

Researchers involved in this project also include undergraduate student Xinyue Gabby Liu, Ph.D. student Zonglin Di, and Assistant Professor of Computer Science and Engineering Yang Liu.

Provided by University of California - Santa Cruz