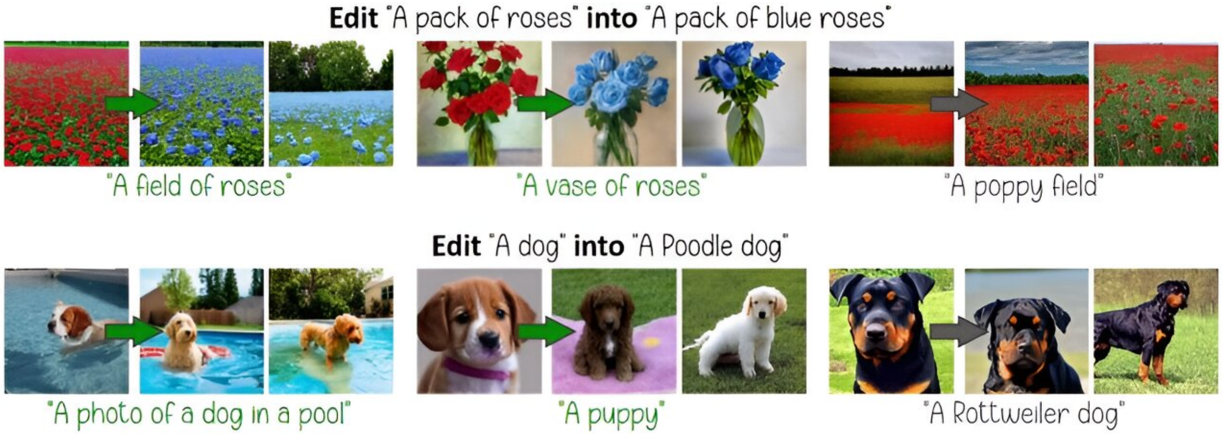
Correcting biases in image generator models

June 24 2024


Edit "A pack of roses" into "A pack of blue roses"
"A field of roses"
"A vase of roses"
"A poppy field"
Edit "A dog" into "A Poodle dog"
"A photo of a dog in a pool"
"A puppy"
"A Rottweiler dog"

Editing a model based on a source and destination prompt. The edit generalizes to related prompts (green), leaving unrelated ones unaffected (gray). Credit: Hadas Orgad et al

Image generator models—systems that produce new images based on textual descriptions—have become a common and well-known phenomenon in the past year. Their continuous improvement, largely relying on developments in the field of artificial intelligence, makes them an important resource in various fields.

To achieve good results, these models are trained on vast amounts of image-text pairs—for example, matching the text "picture of a dog" to a picture of a dog, repeated millions of times. Through this training, the model learns to generate original images of dogs.

However, as noted by Hadas Orgad, a doctoral student from the Henry and Marilyn Taub Faculty of Computer Science, and Bahjat Kawar a graduate of the same Faculty, "since these models are trained on a lot of data from the real world, they acquire and internalize assumptions about the world during the training process.

"Some of these assumptions are useful, for example, 'the sky is blue,' and they allow us to obtain beautiful images even with short and simple descriptions. On the other hand, the model also encodes incorrect or irrelevant assumptions about the world, as well as societal biases. For example, if we ask Stable Diffusion (a very popular image generator) for a picture of a CEO, we will only get pictures of women in 4% of cases."

Another problem these models face is the significant number of changes occurring in the world around us. The models cannot adapt to the changes after the training process.

As Dana Arad, also a doctoral student at the Taub Faculty of Computer Science, explains, "during their training process, models also learn a lot of factual knowledge about the world. For example, models learn the identities of heads of state, presidents, and even actors who portrayed popular characters in TV series.

"Such models are no longer updated after their training process, so if we ask a model today to generate a picture of the President of the United States, we might still reasonably receive a picture of Donald Trump, who of course has not been the president in recent years. We wanted to develop an efficient way to update the information without relying on expensive actions."

The "traditional" solution to these problems is constant data correction by the user, retraining, or fine-tuning. However, these fixes incur high costs financially, in terms of workload, in terms of result quality, and in environmental aspects (due to the longer operation of computer servers). Additionally, implementing these methods does not guarantee control over unwanted assumptions or new assumptions that may arise. "Therefore," they explain, "we would like a precise method to control the assumptions that the model encodes."

The methods developed by the doctoral students under the guidance of Dr. Yonatan Belinkov address this need. The first method, developed by Orgad and Kawar and called TIME (Text-to-Image Model Editing), allows for the quick and efficient correction of biases and assumptions.

The reason for this is that the correction does not require fine-tuning, retraining, or changing the language model and altering the text interpretation tools, but only a partial re-editing of around 1.95% of the model's parameters. Moreover, the same editing process is performed in less than a second.

In ongoing research based on TIME, called UCE, which has been developed in collaboration with Northeastern and MIT universities, they proposed a way to control a variety of undesirable ethical behaviors of the model—such as copyright infringement or social biases—by removing unwanted associations from the model such as offensive content or artistic styles of different artists.

Another method, developed subsequently by Arad and Orgad, is called ReFACT. It offers a different algorithm for parameter editing and achieves more precise results.

ReFACT edits an even smaller percentage of the model's parameters—only 0.25%—and manages to perform a wider variety of edits, even in cases where previous methods failed. It does so while maintaining the quality of the images and the facts and assumptions of the model that we want to preserve.

The methods receive inputs from the user regarding a fact or assumption they want to edit. For example, in cases of implicit assumptions, the method receives a "source" on which the model bases implicit assumptions (e.g., "red roses" by default the model assumes red roses) and a "target" that describes the same circumstances but with the desired features (e.g., "blue roses").

When wanting to use the method for role editing, the method receives an editing request (e.g., "President of the United States") and then a "source" and "target" ("Donald Trump" and "Joe Biden," respectively). The researchers collected about 200 works and assumptions on which they tested the editing methods and showed that these are efficient methods for updating information and correcting biases.

TIME was presented in October 2023 at the ICCV conference, a conference in the field of computer vision and machine learning. UCE was recently presented at the WACV conference.

ReFACT was presented in Mexico at the NAACL conference, a conference in natural language processing research.

More information: Editing Implicit Assumptions in Text-to-Image Diffusion Models

ReFACT: Updating Text-to-Image Models by Editing the Text Encoder

Provided by Technion - Israel Institute of Technology

Citation: Correcting biases in image generator models (2024, June 24) retrieved 28 June 2024 from https://techxplore.com/news/2024-06-biases-image-generator.html