The Weizmann Institute of Science
Faculty of Mathematics and Computer Science

Vision and AI

Room 1, Ziskind Building
on Thursday, May 30, 2024
at 12:15

Hadas Orgad
Technion

will speak on

Editing methods for Text-to-Image Models

Abstract:
Text-to-image generative diffusion models are trained on huge amounts of web-scraped image-caption pairs. As a result, these models encode real-world information and correlations, such as the identity of the President of the United States, or the color of the sky. While this knowledge can be useful, and allows easy and efficient generation of beautiful images from simple prompts, it may also be outdated, reflect assumptions and biases (e.g., doctors are always white male), or violate copyrights (as was demonstrated in recent lawsuits for models imitating artistic styles). However, model providers and creators currently have no efficient means to update models without either retraining them---which is costly in computation and time, and might also require data curation---or requiring explicit prompt engineering from the end user. In this talk, we will discuss three of our recent papers, which aim to offer a fast and practical way to control model behavior post-training. We modify a small, targeted part of the model that is responsible for encoding a certain part in the computation process of the deep network. This is done without training, by editing the model weights using a closed-form solution. The different papers target different parts of the model, as well as various types of information encoded in it: implicit assumptions, factual associations, artistic style, social biases, and harmful content. We will also discuss some of the interpretability aspects and insights that can be gained from these editing methods. Overall, the methods presented in the talk offer a fast and practical means for safe deployment of text-to-image models.