The Weizmann Institute of Science
Faculty of Mathematics and Computer Science

Vision and AI

Room 1, Ziskind Building
on Thursday, Jul 04, 2024
at 12:15

Eliahu Horwitz
HUJI

will speak on

Recovering the Pre-Fine-Tuning Weights of Generative Models

Abstract:
The dominant paradigm in generative modeling consists of two steps: i) pre-training on a large-scale but unsafe dataset, ii) aligning the pre-trained model with human values via fine-tuning. This practice is considered safe, as no current method can recover the unsafe, pre-fine-tuning model weights. In this paper, we demonstrate that this assumption is often false. Concretely, we present Spectral DeTuning, a method that can recover the weights of the pre-fine-tuning model using a few low-rank (LoRA) fine-tuned models. In contrast to previous attacks that attempt to recover pre-fine-tuning capabilities, our method aims to recover the exact pre-fine-tuning weights. Our approach exploits this new vulnerability against large-scale models such as a personalized Stable Diffusion and an aligned Mistral.

Bio:

Eliahu Horwitz is a PhD candidate in Computer Science at the Hebrew University of Jerusalem, working under the supervision of Prof. Yedid Hoshen. His research area is computer vision, with a focus on representation learning and generative models. Currently, his work revolves around reversing the training trajectories of neural networks.

A recipient of the KLA Scholarship for Outstanding Graduate Students and a CIDR (Center for Interdisciplinary Data Science Research) fellow, Eliahu's academic achievements are complemented by his practical experience. Before transitioning to research, he honed his skills as a self-taught software developer, working with diverse technologies across the tech stack at both startups and large-scale companies. His latest research can be found on his website: pages.cs.huji.ac.il/eliahu-horwitz.