The Weizmann Institute of Science
Faculty of Mathematics and Computer Science

Vision and AI

Room 1, Ziskind Building
on Sunday, Oct 06, 2024
at 12:00

*** Please note the unusual day and time***

Yossi Gandelsman
UC Berkeley

will speak on

Reverse Engineering CLIP

Abstract:
In this talk, I reverse engineer CLIP, one of the most commonly used computer vision backbones. I analyze how individual model components affect the final CLIP representation. I show that the image representation can be decomposed as a sum across individual image patches, model layers, neurons, and attention heads, and use CLIP's text representation to interpret the summands.

When interpreting the attention heads, each head role can be characterized by automatically finding text representations that span its output space, which reveals property-specific roles for many heads (e.g. location or shape). Next, interpreting the image patches uncovers an emergent spatial localization within CLIP. Finally, the automatic description of the contributions of individual neurons shows polysemantic behavior - each neuron corresponds to multiple, often unrelated, concepts (e.g. ships and cars).

The gained understanding of different components allows three main applications: First, the discovered head roles enable the removal of spurious features from CLIP. Second, emergent localization is used for a strong zero-shot image segmenter. Finally, the extracted neuron polysemy allows the mass production of "semantic" adversarial examples by generating images with concepts spuriously correlated to the incorrect class. The results indicate that a scalable understanding of transformer models is attainable and can be used to detect model bugs, repair them, and improve them.

BIO:

Yossi is a computer science PhD at UC Berkeley, advised by Alexei Efros, and a visiting researcher at Meta. Before that, he was a member of the perception team at Google Research (now Google-

DeepMind). He completed his M.Sc. at Weizmann Institute, advised by Prof. Michal Irani. His research centers around deep learning, computer vision, and mechanistic interpretability.