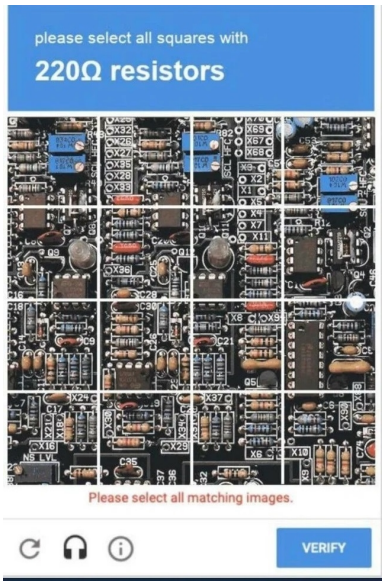
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Can You See What AI Can't?



Recent Vision-Language Models (VLMs) have achieved remarkable performance on image understanding and visual question answering. Nevertheless, even state-of-the-art models still exhibit surprisingly brittle behavior on seemingly simple visual tasks, such as accurately counting objects or reasoning about combinations of their attributes and spatial relations. Humans, on the other hand, can often solve such tasks almost effortlessly. Can these differences be exploited to design a new generation of simple and automatically generated CAPTCHAs? The aim of this thesis is to investigate whether such *compositional visual challenges* can provide a robust separation between humans and modern vision models. In particular, we are interested in designing a flexible synthetic image generator, identifying visual reasoning tasks on which current models remain brittle, and studying whether models can generalize to previously unseen combinations of objects, attributes, and relations.

We are looking for motivated students who are excited to work at the intersection of computer vision, multimodal models, and AI security.

Requirements

Programming skills (Python) and knowledge of machine learning. Prior experience with vision-language models, computer vision, or related machine learning frameworks such as PyTorch is an advantage. We will have weekly meetings to address questions together, discuss progress, and think about future ideas.

Contact

In a few short sentences, please tell us why you are interested in the project and about your coding and machine learning background (i.e., your own projects or courses).

- Giacomo Camposampiero: gcamposampie@ethz.ch, ETZ G97