Tatsu Hashimoto
Assistant Professor of Computer Science
Stanford

Empowering Instruction Following Research with Language Models as Simulators

Abstract: Instruction-following language models have driven remarkable progress in a range of NLP tasks and have been rapidly adopted across the world. However, academic research into these models has lagged behind due to the lack of open, reproducible, and low-cost environments with which to develop and test instruction-following models. In this talk, I will discuss how new, emerging approaches that study an LLM's ability to emulate human annotators and API endpoints hold promise in improving and critiquing LLMs.

To improve instruction-following methods, recent work from our group such as AlpacaFarm shows how an LLM-based simulator can help test scientific hypotheses (e.g. is reinforcement learning helpful?) develop better instruction-following methods, and red-team LLMs in a more open and reproducible way. At the same time, there are major limits to LLMs' ability to simulate annotators — such as in the opinions they reflect or the consistency of their responses — and we will discuss how these gaps raise important open problems in the trustworthiness of existing LLMs.

Yan Liu
Professor of Computer Science
University of Southern California

Deciphering Neural Networks through the Lenses of Feature Interactions

Abstract: Interpreting how neural networks work is a crucial and challenging task in machine learning. In this talk, I will discuss a novel framework, namely neural interaction detector (NID), for interpreting complex neural networks by detecting statistical interactions captured by the neural networks. Furthermore, we can construct a more interpretable generalized additive model that achieves similar prediction performance as the original neural networks. Experiment results on several applications, such as recommender systems, image recognition, and sentiment prediction, demonstrate the effectiveness of NID.

Yao Qin
Assistant Professor of Electrical and Computer Engineering
University of California, Santa Barbara

Effective Robustness against Natural Distribution Shifts for Models with Different Training Data

Abstract: "Effective robustness'' measures the extra out-of-distribution (OOD) robustness beyond what can be predicted from the in-distribution (ID) performance. Existing effective robustness evaluations typically use a single test set such as ImageNet to evaluate ID accuracy. This becomes problematic when evaluating models trained on different data distributions, e.g., comparing models trained on ImageNet vs. zero-shot language-image pre-trained models trained on LAION. In this paper, we propose a new effective robustness evaluation metric to compare the effective robustness of models trained on different data distributions. To do this we control for the accuracy on multiple ID test sets that cover the training distributions for all the evaluated models. Our new evaluation metric provides a better estimate of the effectiveness robustness and explains the surprising effective robustness gains of zero-shot CLIP-like models exhibited when considering only one ID dataset, while the gains diminish under our evaluation.