香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE
計算機科學系

TMLR Young Scientist SEMINAR

2024 SERIES

Trustworthy Machine Learning and Reasoning Group

Mr. Yibo Jiang

PhD student,
University of Chicago.

Date: 09 Oct 2024 (Wednesday)

Time: 19:00 – 20:00 (HKT)

Meeting: https://meeting.tencent.com/dm/RSf3XzcVpP0s

On the origins of linear representations in LLMs

ABSTRACT

Recent works have argued that high-level semantic concepts are encoded "linearly" in the representation space of large language models. In this work, we study the origins of such linear representations. To that end, we introduce a simple latent variable model to abstract and formalize the concept dynamics of the next token prediction. We use this formalism to show that the next token prediction objective (softmax with cross-entropy) and the implicit bias of gradient descent together promote the linear representation of concepts. Experiments show that linear representations emerge when learning from data matching the latent variable model, confirming that this simple structure already suffices to yield linear representations. We additionally confirm some predictions of the theory using the LLaMA-2 large language model, giving evidence that the simplified model yields generalizable insights.

BIOGRAPHY

I'm a PhD student at the University of Chicago, working under the guidance of Prof. Victor Veitch. My research focuses on causality, interpretability, and representation learning. I earned my Master's degree in Computer Science from Columbia University and completed my Bachelor's degrees in Electrical Engineering and Mathematics at the University of Illinois Urbana-Champaign.

ENQUIRY

Email: bhanml@comp.hkbu.edu.hk