Department of Computer Science
香港城市大學
City University of Hong Kong

COMPUTER SCIENCE COLLOQUIUM

Project ALPINE: Unveiling The Planning Capability of Autoregressive Learning in Language Models

SPEAKER **Dr. Wei Chen**

Principal Researcher at Microsoft Research Asia, and the Chair of MSRA Theory Center
Microsoft Research Asia

DATE 21 Nov, 2024 (Thu)
TIME 10:30 AM - 12:00 PM
VENUE CS Seminar Room, Y6405, 6th Floor, Yellow Zone, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong

ABSTRACT

Planning is a crucial element of both human intelligence and contemporary large language models (LLMs). In this talk, I introduce the project ALPINE, which initiates a theoretical investigation into the emergence of planning capabilities in Transformer-based LLMs via their next-word prediction mechanisms. We model planning as a network path-finding task, where the objective is to generate a valid path from a specified source node to a designated target node. Our mathematical characterization shows that Transformer architectures can execute path-finding by embedding the adjacency and reachability matrices within their weights. Furthermore, our theoretical analysis of gradient-based learning dynamics reveals that LLMs can learn both the adjacency and a limited form of the reachability matrices. These theoretical insights are then validated through experiments, which demonstrate that Transformer architectures indeed learn the adjacency and an incomplete reachability matrices, consistent with our theoretical predictions. When applying our methodology to the real-world planning benchmark Blocksworld, our observations remain consistent. Additionally, our analyses uncover a fundamental limitation of current Transformer architectures in path-finding: these architectures cannot identify reachability relationships through transitivity, which leads to failures in generating paths when concatenation is required. These findings provide new insights into how the internal mechanisms of autoregressive learning facilitate intelligent planning and deepen our understanding of how future LLMs might achieve more advanced and general planning-and-reasoning capabilities across diverse applications.

BIOGRAPHY

Wei Chen is a Principal Researcher at Microsoft Research Asia, and the Chair of MSRA Theory Center. He is a guest professor at several universities including Tsinghua University, Shanghai Jiao Tong University, Hong Kong University of Science and Technology – Guangzhou, and Shenzhen University. He is a Fellow of Institute of Electrical and Electronic Engineers (IEEE). He serves as a standing committee member of the Technical Committee on Theoretical Computer Science, Chinese Computer Federation, and a member of the CCF Technical Committee on Big Data. He is recognized by Elsevier as the most cited Chinese researchers (2021-2023), and is ranked as the top 2% scientists worldwide by the Stanford ranking (2020-2024). Wei Chen's main research interests include online learning and optimization, social and information networks, network game theory and economics, distributed computing, and fault tolerance. He has one coauthored monograph in English in 2013 and one sole authored monograph in Chinese in 2020, both on information and influence propagation in social networks. He has won several best paper awards including 2021 ICDM 10-Year Highest-Impact Paper Award, and William C. Carter Award for best paper based on a dissertation research in DSN'2000. He has served as editors, academic conference chairs and program committee members for many academic conferences and journals. Wei Chen has bachelor's and master's degrees from Tsinghua University and a Ph.D. degree in computer science from Cornell University. For more information, you are welcome to visit his home page at http://research.microsoft.com/en-us/people/weic/.

All are welcome!

In case of questions, please contact Prof Jinhang ZUO at jinhang.zuo@cityu.edu.hk, or visit the CS Departmental Seminar Web at https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums.