THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

Research Seminar

Do We Need Attention for Large-Scale Deep Learning Models?

Speaker: Dr. Jing Nathan Yan, Cornell Bowers CIS (Ann Bowers College of Computing and Information Science) and Cornell Tech

Date: January 14, 2024 (Sun)

Time: 14:00 (HKT)

Mixed Mode: Room 308, Chow Yei Ching Building & Zoom (Hybrid)

Abstract

Attention mechanisms play a pivotal role in the realms of language modeling and image generation, despite their quadratic complexity. As we continue to scale models, the interplay between attention-based architectures and their learning capabilities remains a central theme in the development of large-scale deep learning models. While alternative architectures like CNNs and RNNs have been explored, they often fall short in accuracy or necessitate additional attention layers for optimal performance. In this talk, we re-examine seminal results in bidirectional language modeling and image generation, while also introducing a new architecture that eschews traditional attention mechanisms. We propose replacing self-attention layers with a recent approach for long-range sequence modeling, alongside variants of the transformer architecture. Drawing inspiration from recent works such as the Structured State Space Sequence Model (S4), our approach utilizes straightforward routing layers based on State-Space Models (SSM) and a bidirectional architecture incorporating multiplicative gating. We will discuss the outcomes of implementing our Bidirectional Gated SSM (BiGS) and Diffusion State Space Model (DiffuSSM) in the contexts of both language modeling and image generation. Our analysis dives into the properties of these models, highlighting how architectural choices significantly influence downstream performance and introduce a distinct inductive bias. Preliminary results suggest that these new models not only perform effectively but also reduce computational costs, presenting an avenue worth further exploration in the field of deep learning.

About the Speaker:

Jing Nathan Yan is a PhD candidate at Cornell Bowers CIS (Ann Bowers College of Computing and Information Science) and Cornell Tech. He is extremely fortunate to be advised by Prof. Sasha Rush and Prof. Jeff Rzezsotarski, and he is currently based in New York City. Before coming to Cornell, he received his MPhil degree in the Department of Computer Science from HKU. His research focuses on alternative architectures for large-scale deep learning systems and machine learning fairness assessment. In the past, he interned at Google Research/Google DeepMind, Microsoft Research, and Facebook AI Research Labs. When he is not doing research, he can be found making noise with his band friends or climbing walls.

All are welcome!

For enquiries, please call 3917 2180 or email enquiry@cs.hku.hk