Surbhi Goel: Thinking fast with Transformers - Algorithmic Reasoning via Shortcuts

In this new era of deep learning, the emergent algorithmic reasoning capabilities of Transformer models have led to significant advancements in natural language processing, program synthesis, and theorem proving. Despite their widespread success, the underlying reasons for their efficacy and the nature of their internal representations remain elusive. In this talk, we take the lens of learning the dynamics of finite-state machines (automata) as the underlying algorithmic reasoning task and shed light on how shallow, non-recurrent Transformer models emulate these recurrent dynamics. By employing tools from circuit complexity and semigroup theory, we characterize "shortcut" solutions that allow a shallow Transformer to precisely replicate T computational steps of an automaton with only $o(T)$ layers. We show that Transformers are efficiently able to represent these “shortcuts” using their parameter-efficient ability to compute sparse functions and averages. Furthermore, through synthetic experiments, we confirm that standard training successfully discovers these shortcuts. We conclude with highlighting the brittleness of these “shortcuts” in out-of-distribution scenarios. *This talk is based on joint work with Bingbin Liu, Jordan T. Ash, Akshay Krishnamurthy, and Cyril Zhang.*