Chain of Thought Empowers Transformers to Solve Inherently Serial Problems

AgentX Competition

Yalei Yu

April 8, 2025

Tasks

- a theoretical understanding of the power of CoT for decoder-only transformers through the lens of expressiveness

Notations

- Decoder-only transfer: involves adapting decoder-only transformers (e.g., GPT-series) to downstream tasks efficiently, exploiting their generative strengths without introducing separate encoder architectures
- Lens of expressiveness is simply a way of evaluating or analyzing models based on their ability to represent complex patterns or functions.
- Input length: refers to the maximum number of tokens (words or subwords) that an LLM can process simultaneously
- Constant depth of transfer: The main goal is to maintain efficiency—keeping the computational cost constant while transferring representations or adapting the model.
- Finite precision poly(n) embedding size:
  - Use limited numerical precision (e.g., float16, float32, or quantized precision).
  - Grow at a manageable rate (polynomially rather than exponentially) with input size or complexity

Notations

- TC°: constant-depth threshold-circuit tasks
- $O(\log n)$ embedding size means that the dimension of embeddings grows very slowly with input size—highly efficient but limited in representational power.
- Boolean circuits of size
  - Are computational models composed of AND, OR, NOT gates.
  - Have complexity measured by the total number of gates (size T)

Hypothesis

- CoT allows for performing more serial computations that a vanilla transformer cannot do without CoT

Methods

- the lens of expressiveness with and without CoT
- the language of circuit complexity to discuss the capability of transformers

Phenomenon

- transformers equipped with CoT—allowing the transformer to auto-regressively generate a sequence of intermediate tokens before answering the questions—can solve complex problems that inherently require serial computations
- without CoT, the number of serial computations conducted by the transformer is bounded by the depth (which is considered as a fixed constant for this work), whereas with T intermediate steps, the number of serial computations possible is boosted to T

Thank you very much!
Questions?