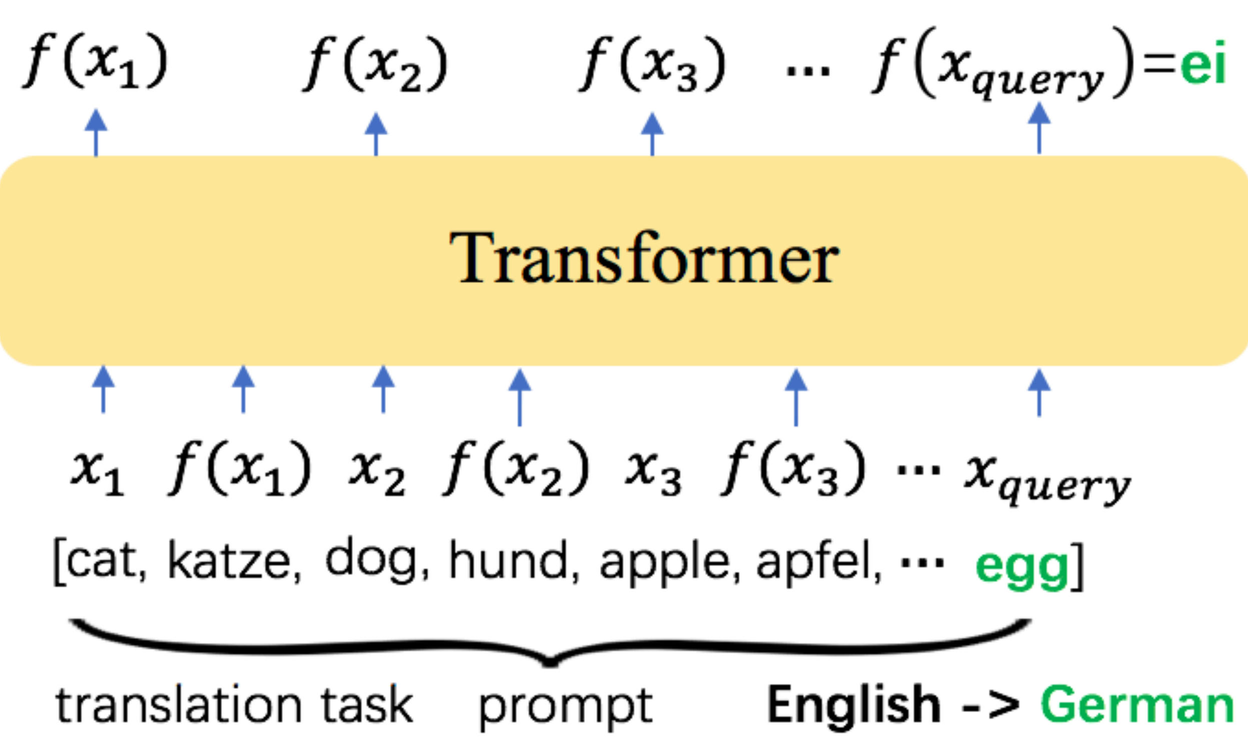
Theoretical and Algorithmic Foundations of In-Context Learning and reasoning Using Properly Trained Transformer Models

Hongkang Li[1], Meng Wang[1], Songtao Lu[1], Xiaodong Cui[1], Pin-Yu Chen[1]. 1: Rensselaer Polytechnic Institute. 2: IBM Research

Motivation

Transformer-based foundation models, e.g., GPT-4, Sora, have achieved great empirical success in many areas.

- Large foundation models are able to implement in-context learning (ICL) and reasoning.
- Theoretical understanding of **how a Transformer can be trained to perform ICL and generalize in and out of domain successfully and efficiently** is less investigated.

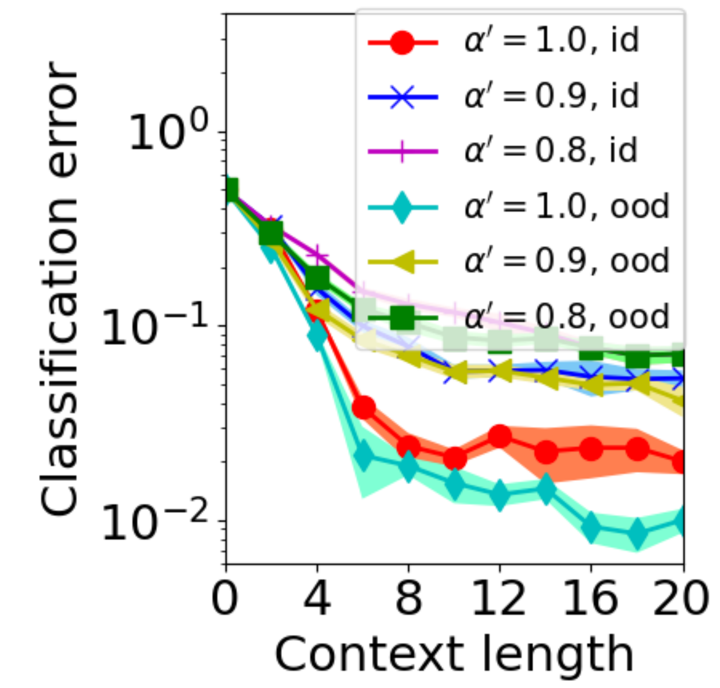


Current Progress

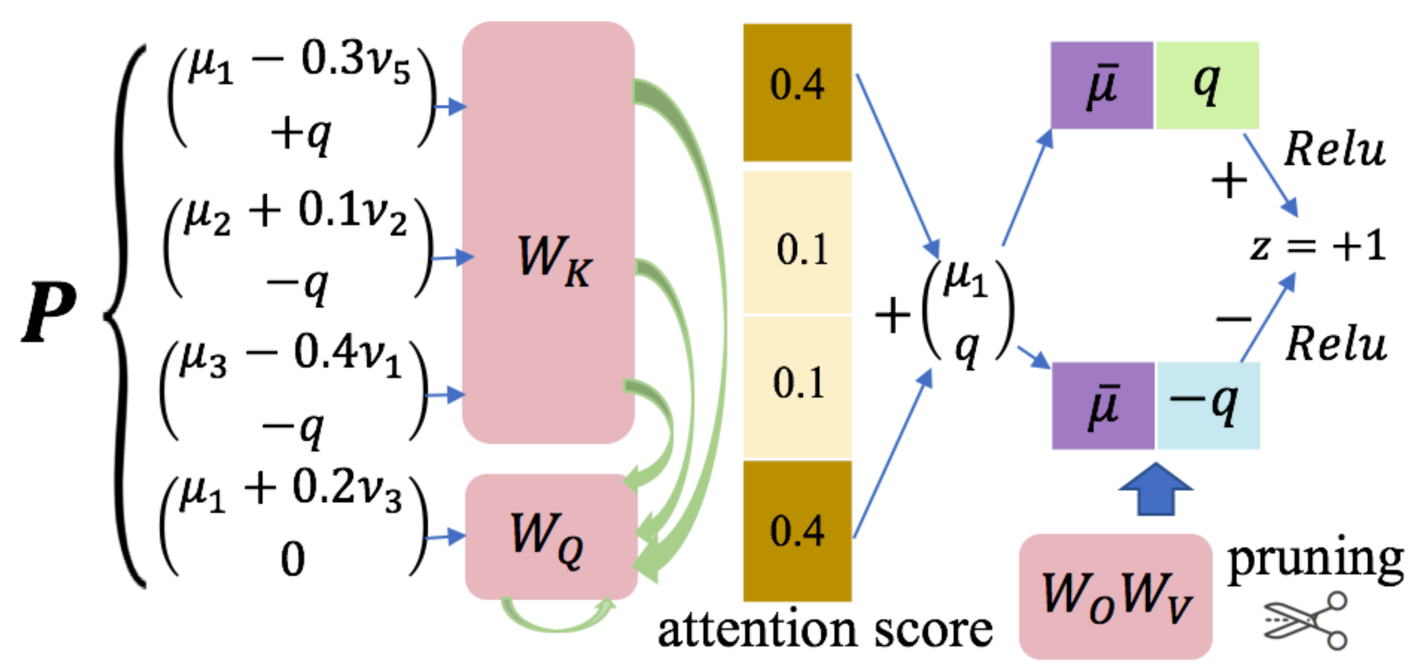
- We provide a theoretical characterization of how to train nonlinear Transformers to enhance their ICL capability on classification tasks. .

Theorem 1 (informal): Given enough neurons and a large batch, and prompt lengths inverse in the fraction of relevant tokens α, then after training with $\Theta(\alpha^{-1})$ steps,

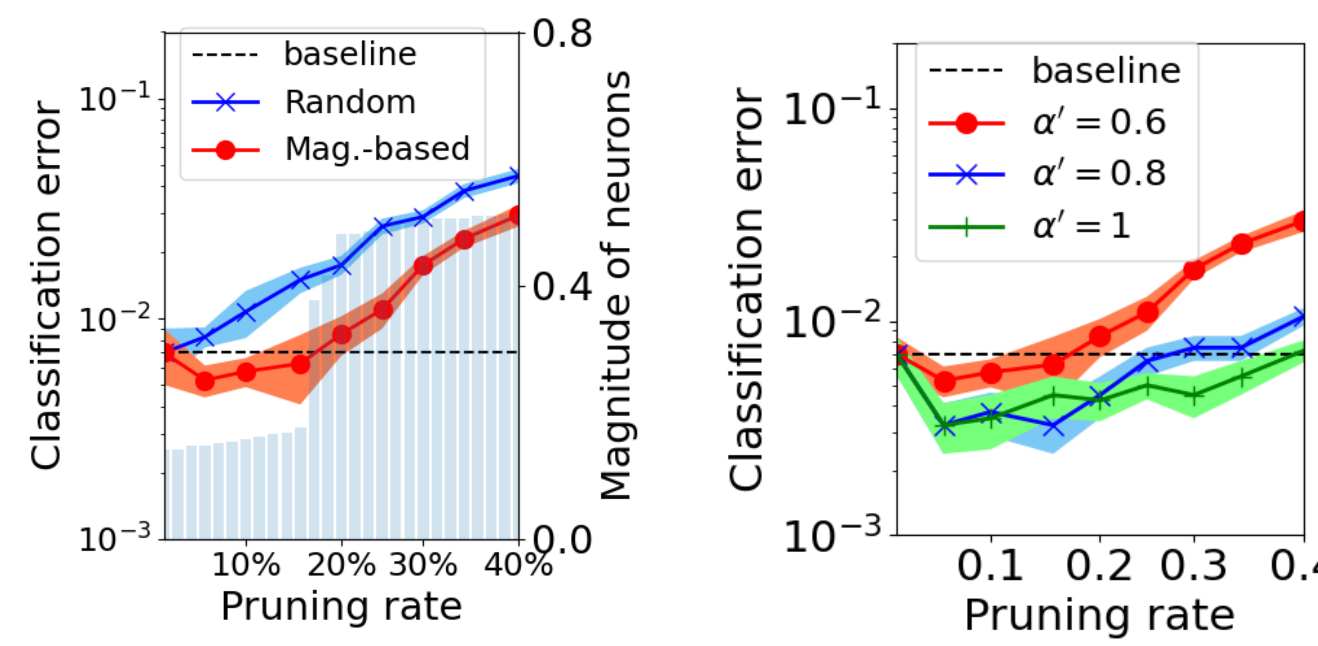
- *the returned one-layer Transformer model achieves an in-domain generalization error no larger than ϵ.*
- *If the testing relevant patterns are linear combinations of the trained ones with coefficient summation no larger than 1, the out-of-domain generalization error is no larger than ϵ.*



- We expand the theoretical understanding of the mechanism of the ICL capability of Transformers.

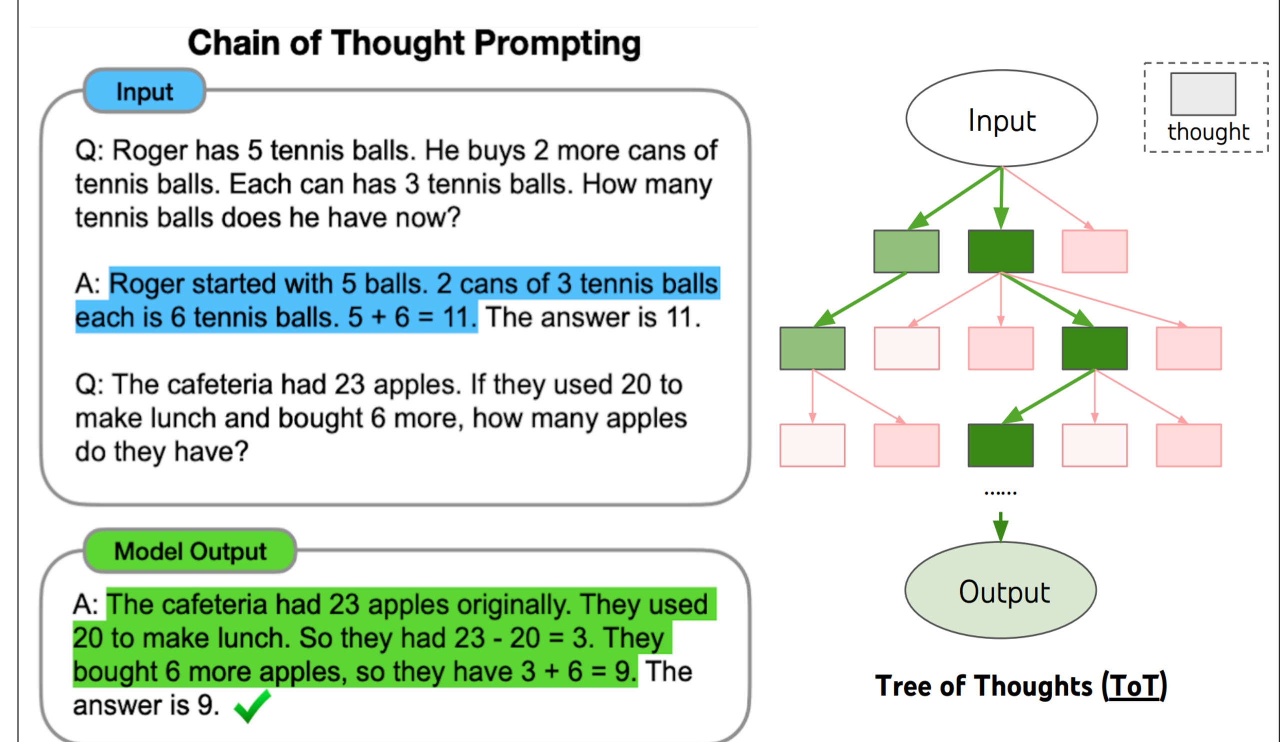


- We theoretically justify the Magnitude-based Pruning in preserving ICL.



Future Plan

LLM reasoning



Problems to solve

- How can a Transformer be trained to learn different hidden causal structure?
- Why does adding intermediate steps help the reasoning in theory?
- What is the mechanism of a Transformer implementing reasoning in context?

Theoretical contributions

- Hidden Markov chain modeling.
- Next token prediction beyond classification and regression.

Experiments

- Evaluate the results on the arithmetic reasoning dataset GSM8K and the commonsense reasoning dataset CSQA.

Future of Computing Research Collaboration at Rensselaer