Abstract

In-context Learning: Given a few task examples in context, a model learns on the fly to perform the task and complete a test query.

Hypothesis: LLMs simulate a few steps of GD on a small, simple model (e.g., logistic classifier) during inference [1, 2, 3]

Can a transformer simulate and train an internal transformer?

How many parameters to simulate-train a 125M transformer?

We construct a transformer with < 2B parameters that can simulate training a OPT-125M model during inference. We call this TinT.

In-Context Learning

- 4 major operations: linear, activation, self-attention, layer-norm.
- Create 3 simulator modules for each operation (12 total) to simulate forward, backward, and descent operations.
- Stack the simulator modules according to the auxiliary model.

Example: Simulating forward operation on linear layer

\[
x_1, \cdots, x_T \rightarrow y_1, \cdots, y_T
\]

Issues with a blind implementation:
- Sequence Length blowup
- Inefficient use of attention heads

Strategies:
- Stacking multiple params to prefix tokens
- Computation split across attention heads

Other modifications for parameter efficiency:
- Duality of self-attention for efficient descent through linear layer
- Zero-order gradient approximations
- Approximations via stop-gradient
- Parallelizing computation by Sub-MLP partition in MLPs
- Low-rank linear layers in TinT

Architecture Overview

Simulating GD with TinT: first few layers simulate the forward propagation, next few layers simulate backpropagation and descent, final layers simulate evaluation with the updated model.

Two necessities
A. Read and write access to T2
B. T2’s read access to in-context data

Strategy
A. Feed in T2’s parameters as input prefix tokens to each layer. Similar to Prefix tuning of LLMs [4]
B. Bidirectional attention on in-context and prefix tokens. Similar to Prefix LLMs [5, 6].

Empirical Validations

Drop in Perplexity w.r.t. OPT-125M

- 0.4-0.6 better perplexity over OPT-125M.
- TinT performs close to true one-step GD.

Language Modeling

- Achieves >4% better accuracy over OPT-125M.
- TinT performs close to true one-step GD.
- Competitive with comparatively sized OPT.

TinT’s size grows as internal model grows

- Backward and descent operations require ~2x more parameters than forward operations.

References

8. ITLS Input Language Toolset for Services. 2004
9. ITLS Input Language Toolset for Services. 2004
10. ITLS Input Language Toolset for Services. 2004