Recursive Reasoning with Tiny Networks

Group Project Deep Learning

October 26, 2025

Group 7

Tiny Recursive Models (TRMs)

TODO.

Motivation

TODO.

Background

TODO.

Milestones

Verifying the replacement of attention by MLP's Self-attention is good when $L \gg D$, i.e. the input length is way bigger than the embedding dimension. This is because we only need a $D \times 3D$ matrix for the keys, queries and values. But for $L \leq D$, we can use a linear map from $\mathbb{R}^L$ to $\mathbb{R}^L$ instead, which only requires an $L \times L$ matrix. This means that for small models, according to the authors, we can replace attention with MLP's.

Verifying that adding more layers leads to overfitting and 2 layers is optimal Decreasing the number of layers, while scaling the number of recursions n proportionally, the authors found that using 2 layers (instead of 4 layers) maximized generalization.

Verifying the removal of the continue loss By removing the continue loss with Adaptive computational time (ACT), there's supposedly no need for the expensive second forward pass, while still being able to determine when to halt with relatively good accuracy.

Increasing the number of recursions for hard problems The authors found that $T = 3$ (latent) recursion processes, each containing $n = 6$ evaluations of f was optimal for Sudoku-Extreme, but mentioned that they did not test more than $T \cdot n = 42$ total recursions due to limited resources.

Testing the TRM on different puzzles, comparing it to other models E.g. towers of Hanoi, ...

References

[1] Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks, 2025. `https://arxiv.org/abs/2510.04871`.