WHY ARE SMALL TRANSFORMERS NOT BETTER?

Anonymous authors
Paper under double-blind review

ABSTRACT

The attention mechanism is powering rapid progress in terms of large-scale generative AI. It is conversely exceedingly difficult to find small-scale applications for which attention-based models outperform traditional approaches, such as multilayer perceptrons or recurrent networks. We examine this problem in the context of 'task switching'. In this framework models work on ongoing token sequences with the current task being determined by stochastically interseeded control tokens. We show that standard transformers cannot solve a basic reference model, IARC, which is based on finite-domain arithmetics. The model contains a trivial unary operation, (I: increment the current input), a likewise trivial binary operation, (A: add last two inputs), and reverse copy, (R), a standard memory task. A fourth control token, (C), adds recursive context dependency by modifying current tasks. Tasks are maintained as long as no new control tokens appears in the prompt, which happens stochastially every 3-9 steps. We show that transformers, LSTM recurrent networks and plain MLPs of similar sizes ($\sim$1.5M parameters) achieve only modest prediction accuracies, of about 45%. As a countertest we trained transformers containing a modified attention mechanism, expressive attention, finding performance levels of around 95%. Our results indicate that the workings of attention can be understood better, and even improved, when comparing qualitatively different formulations is a task-switching setting.

1 INTRODUCTION

An undisputed advantage of the transformer architecture is that memory requirements scale only linearly with context length Vaswani et al. (2017). Compute scales however quadratically, a feature shared with fully-connected multi-layer perceptrons and recurrent networks. It remains an open question whether the success of transformers is due to particular properties of the underlying attention mechanism, or a consequence of the resulting improved size scaling. Alternative models with favorable scaling may be just as good in the later case Gu & Dao (2023). This question received further urgency by the recent observation that MLPs learn in-context a la par with transformers when given the same compute budget Tong & Pehlevan (2024). It is hence important to study to which extend transfomers excel or fail for small-sized applications, viz in a regime where scaling is not yet relevant.

Here we work below the scaling regime, typically for a context length $N_c = 24$. In this regime, in which transformers and classical models have about the same number of adjustable parameters, we evaluated to which extend several standard models are able to switch task upon the appearance of a suitable control token. Individual tasks work on sequences of encoded numbers, skipping the interseeded control tokens, as explained further in Sect. 3. The resulting evaluation protocol can be varied and/or extended by selecting appropriate tasks, making it a versatile tool for comparative performance tests. For real-world applications, task switching protocols are relevant, e.g., for steering robots by switching between motor primitives Saveriano et al. (2023).

Our results support the notions that small transformers are not generically better than MLPs or recurrent networks. This holds, however, only when attention weights are $\sim\exp(z_{ij}/\sqrt{N_c})$, the standard formulation, where z_{ij} is the dot-product between query and key. We also tested transformers based on expressive attention Gros (2024), for which attention weights are proportional to $z_{ij}^2/(1+z_{ij}^2)$, finding substantial performance leaps.

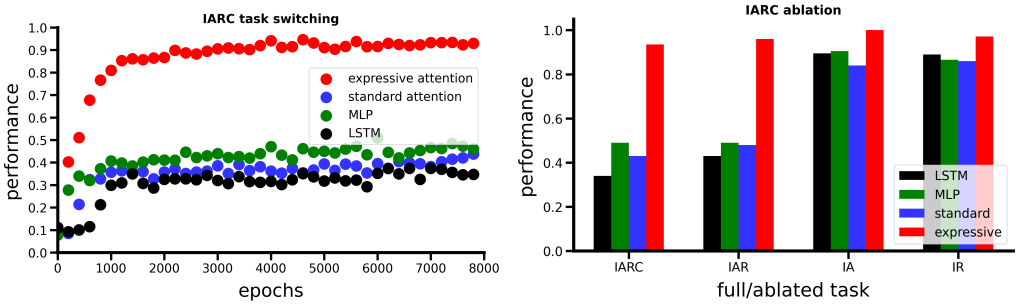

Figure 1: The IARC task switching evaluation framework, for details see Sect. 3. Shown are results for a LSTM recurrent network (black), a MLP (green), a standard transformer (blue), and a transformer with expressive attention (red). **Left:** As a function of training epochs, the prediction accuracy (performance). **Right:** Testing with various combinations of the fundamental tasks, including (I), incrementing the current input, (A), adding the last two inputs and (R), reverse copy. Recursive context dependency is encoded by (C).

2 Related Work

The task-switching protocol used here is a specification of multi task learning Zhang & Yang (2021); Chen et al. (2024), in the form of ever ongoing sequences of concatenated tasks. Task switching is in particular important for reset-free robotic applications Gupta et al. (2021), e.g., in the context of embodied robotics Kumar et al. (2024), or for large language models Knight & Duan (2024). Related to our approach are testing procedures involving synthetic reasoning tasks Zhang et al. (2022), which have been applied to small models in the form of in-context and global bigrams Bietti et al. (2024). Another example are toy models of superposition, which can be used to study polysemanticity Elhage et al. (2022). It is likewise important to find out which formal languages transformers can express Strobl et al. (2024); Deletang et al. (2022). Equivalently the question arises to which extend large-model scaling Kaplan et al. (2020); Hoffmann et al. (2022), and variantes thereof Naveed et al. (2023); Shen et al. (2024), is retained when models are small Ivgi et al. (2022).

3 IARC Task Switching Framework

A considerable number of benchmark tasks for the evaluation of transformer variantes have been developed Liu et al. (2024), however nearly exclusively for large model sizes Tay et al. (2020). The benschmark task introduced here is meant in contrast for the evaluation of small models, typically with a few million parameters or less.

The vocabulary consists of a finite set of N numbers, plus a limited number of control tokens, denoted here (I), (A), (R), and (C). Control tokens interseed the sequence of symbols, $S_t = \{x_{t'}|t' \leq t\}$. The task is to predict the next symbol, x_{t+1}, but not the occurance of future control tokens. The last control token, or the history of previous control tokens, determines the dependence of x_{t+1} on S_t. One has

$$x_{t+1}\big|_I = (x_t + 1)\%N, \qquad x_{t+1}\big|_A = (x_t + x_{t-1})\%N \tag{1}$$

for increment (I) and addition (A). For $N = 10$, an example would be

$$| \ 2 \ | \ 3 \ | \ 4 \overset{\text{A}}{|} \ 7 \ | \ 1 \ | \ 8 \ | \ 9 \ | \ 7 \overset{\text{I}}{|} 8 \ | \ 9 \ | \ 0 \ | \ 1 \tag{2}$$

Here with two interseeded control tokens, as indicated by the respective superscripts. Including reverse copy (R), one could have

$$| \ 2 \ | \ 3 \ | \ 4 \overset{\text{A}}{|} \ 7 \ | \ 1 \ | \ 8 \overset{\text{R}}{|} \ 8 \ | \ 1 \ | \ 7 \ | \ 4 \ | \ 3 \overset{\text{R}}{|} \ 3 \ | \ 4 \ | \ 7 \tag{3}$$

Tasks remain when (I/A) tokens are followed by (I/A) tokens, with the reverse copy process being restarted by subsequent (R) tokens. The action of the context token (C) depends on the current task, consecutively increasing the increment by one for (I) tasks:

$$| \ 1 \overset{\text{I}}{|} 2 \ | \ 3 \ | \ 4 \overset{\text{C}}{|} \ 6 \ | \ 8 \ | \ 0 \ | \ 2 \ | \ 4 \overset{\text{C}}{|} \ 7 \ | \ 0 \ | \ 3 \tag{4}$$

When the current task is (A/R), the context token (C) just acts as an additional (A/R) token, as illustrated in (3) for the case of two consecutive (R) token. As a basic protocol regulating the frequency of task switching, we use a 6 ± 3 setup, which means that the distance between subsequent control tokens is drawn from a flat distribution out of $[3, 9]$. The four control tokens, (A/I/R/C), appear with equal probabilities.

4 RESULTS

For all models the context length is $N_c = 24$, together with one-hot embedding and a fixed embedding dimension $d = 20$. We did set $d = N + S$, where S is the number of control symbols, compare (1). This implies $N = 16$ for IARC and $N = 17$ for an ablated version like IAR. Transformers have $L = 12$ layers, causal self-attention with four heads and ALiBi positional encoding Press et al. (2021), altogether 1.3M parameters. The multi-layer perceptron (MLP) had $L = 16$ layers with causal connections, resulting in 1.9M parameters, the LSTM recurrent net was somewhat larger, $L = 2$ with a size of 3.7M. For all models we did additional runs with increased numbers of layers, finding only marginal improvements. All runs used identical training protocolls, with a batch size of 500.

Our results are summarized in Fig. 1. The three basic models, LSTM, MLP and the standard transformer show roughly similar performances, both as a function of training compute and when ablating the IARC model. Given the same resources, small standard transformers are not better. A massive outperformance is seen however by transformers based on expressive attention Gros (2024). This is an interesting result, given that expressive attention is intrinsically quadratic in the scalar product between queries and keys, making it bi-quadratic in token activities. It also implies that there is room for gains when reformulating the core of the attention mechanism. Finally we want to point out that task-switching frameworks paralleling the one used here are relevant for real world applications taking place in dynamic environments, such as autonomous driving and robot control.

ACKNOWLEDGMENTS

– to be added –

REFERENCES

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36, 2024.

Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32, 2024.

Gregoire Deletang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, et al. Neural networks and the chomsky hierarchy. In *The Eleventh International Conference on Learning Representations*, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Claudius Gros. Reorganizing attention-space geometry with expressive attention. *arXiv preprint arXiv:2407.18601*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Abhishek Gupta, Justin Yu, Tony Z Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6664–6671. IEEE, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7354–7371, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Parker Knight and Rui Duan. Multi-task learning with summary statistics. *Advances in Neural Information Processing Systems*, 36, 2024.

Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Abhishek Gupta, and Aravind Rajeswaran. Robohive: A unified framework for robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Matteo Saveriano, Fares J Abu-Dakka, Aljaž Kramberger, and Luka Peternel. Dynamic movement primitives in robotics: A tutorial survey. *The International Journal of Robotics Research*, 42(13): 1133–1184, 2023.

Xuyang Shen, Dong Li, Ruitao Leng, Zhen Qin, Weigao Sun, and Yiran Zhong. Scaling laws for linear complexity language models. *arXiv preprint arXiv:2406.16690*, 2024.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

William L Tong and Cengiz Pehlevan. Mlps learn in-context. *arXiv preprint arXiv:2405.15618*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.