Transformers and Reinforcement Learning

August 21, 2026

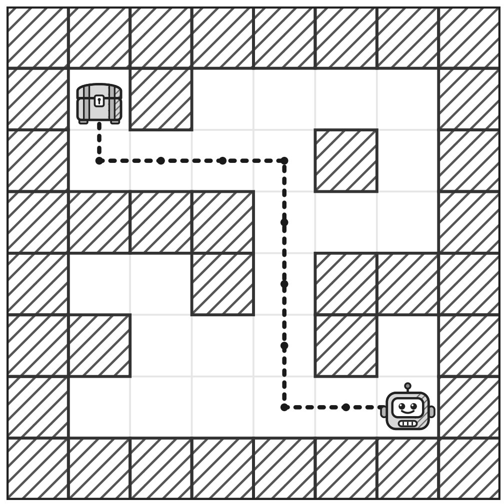

Figure 1: A test case in the qualifying problems.

Introduction

Based on the attention operation [Bahdanau et al., 2015], the transformer architecture [Vaswani et al., 2017] revolutionized natural language processing: its scalability allowed taking models pretrained on generic text corpora and quickly finetune them on various supervised [Devlin et al., 2019] and generative [Radford et al., 2018] natural language processing tasks. Scaling the models further, it was observed that new capabilities emerge such as in-context learning [Radford et al., 2019, Brown et al., 2020].

Reinforcement learning is the machine learning training paradigm where the model only receives a numerical score as feedback signal [Sutton and Barto, 2018]. Reinforcement learning on human feedback was used to finetune generative pretrained transformers into virtual assistants [Ouyang et al., 2022].

It was observed that one can elicit useful output to more complex queries by prompting the model to output its chain of thought [Wei et al., 2022]. This led to improved long-horizon capabilities such as coding and mathematical problem solving via reinforcement learning on verifiable rewards [Lambert et al., 2024, Guo et al., 2025].

Project Description

We will study how small transformer models learn to solve algorithmic tasks via reinforcement learning. This means that we will be able to do relevant research while keeping the computational requirements modest. You will be able to partake without a GPU or prior knowledge of transformers:

1. We will start by reviewing the transformer architecture and the reinforcement learning paradigm. Thus, you'll get to do research while getting acquainted with the most important machine learning model of our time.

2. By default (that is, unless you want to experiment doing this yourselves), I will train the models on the setups we discuss. I expect your main contribution to be designing experiments and interpreting their results. Therefore, you can participate perfectly well with any modern laptop.

Prerequisites

Strong command of the `Python` numerical library `numpy`.

Qualifying problems

See the contents of the attached `zip` file. `qualifying_problems.py` contains functions to be written, which you can test with `test.py`. The script `test.py` has many command line interface options, in particular, it can generate pictures that can help you debug you code or bask in the glory of your working solution.

I very much welcome partial solutions! The functions to be written are ordered in increasing difficulty. Please try to avoid explicit Python loops. Instead, whenever possible, use `numpy` functions that operate on the entire arrays at once.

You can hand in multiple versions. I will evaluate the latest submission. A valid submission must arrive to my email address by the deadline written in the general RES course description on the BSM webpage. You can expect me to answer a question before this time if it arrived to my email address at least 24 hours before this deadline.

In your email, please also write me the following:

1. Your Mathematics and Computer Science background.
2. Your Mathematics and Computer Science interests.
3. What do you find especially interesting in this project?

Have fun with the problems and I hope to see you in the group!

Using Virtual Assistants

Please do not use any AI tool (ChatGPT, Claude, Gemini, etc., or derivatives such as Codex or Copilot) when you work on the qualifying problems. The objective of the qualifying problems is to gauge your own programming capabilities. This is important even in the age of AI as you need to be able to verify the work of the virtual assistant.

This being said, in the group, I will highly welcome using virtual assistants as they can accelerate work greatly.

Contact

Pál Zsámboki, `zsamboki@renyi.hu`, HUN-REN Alfréd Rényi Institute of Mathematics

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL `https://www.nature.com/articles/s41586-025-09422-z`.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. doi: 10.48550/arXiv.2411.15124. URL `https://arxiv.org/abs/2411.15124`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=TG8KACxEON`.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018. ISBN 9780262039246. URL `http://incompleteideas.net/book/the-book-2nd.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=_VjQlMeSB_J`.