Mechanistic Interpretability on Irreducible Integers

Noah Syrkis

February 7, 2024

1 | Mech. interp. (MI)

2 | Grokking

3 | ℤ-sequences

4 | MIII

1 | Mech. interp. (MI)

- Reverse-engineering neural network circuits.
- Nanda et al. [3] shows MI modular addition transformer.
- There are (allegedly) low hanging fruits in MI.

2 | Grokking

- Grokking is when a model suddenly generalises.
- Nanda et al. [3] shows grokking in a transformer.
- Grokking means the weights represents an algorithm…
- … rather than a weired data base.

2 | Grokking (cont.)

- Since MI is about reverse-engineering circuits…
- … grokking is a good sign for MI …
- … as it means circuits are *there*.

3 | $\mathbb{Z}$-sequences

- Belcák et al. [2] shows that transformers can sequences $\in \mathbb{Z}$.
- They work in thousands of squences from OEIS [4].
- They have four tasks: (1) sequence classification, (2) sequence comparission, (3) sequence continuation, and (4) sequence unmasking.
- Each task is strictly harder than the previous one.

3 | $\mathbb{Z}$-sequences (cont.)

- Though $\mathbb{Z}$-sequences are simple to see, some can be hard to impossible to understand.
- $1, 2, 3, ..., 100$ is easy, while the busy beaver sequence [1] is hard/impossible.
- Complexity ranges from trivial to fuck-off-forever.

4 | MIII

- I want to explore the use of MI on $\mathbb{Z}$-sequences.
- Initially, I want to explore the classification task…
- … with possibility of moving up in task complexity.

References

[1] Scott Aaronson. "The Busy Beaver Frontier". In: *ACM SIGACT News* 51.3 (Sept. 2020), pp. 32–54. DOI: 10.1145/3427361.3427369.

[2] Peter Belcák et al. *FACT: Learning Governing Abstractions Behind Integer Sequences*. Sept. 2022. arXiv: 2209.09543 [cs].

[3] Neel Nanda et al. *Progress Measures for Grokking via Mechanistic Interpretability*. Oct. 2023. arXiv: 2301.05217 [cs].

[4] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*. Dec. 2003. DOI: 10.48550/arXiv.math/0312448. arXiv: math/0312448.