Mechanistic Interpretability on Irreducible Integer Identifiers

Noah Syrkis
University of Copenhagen
Feb. 23, 2024

"This disgusting pile of matrices with some non-linearities in between is an impressively poorly written beautiful and concise algorithm"[1]

[1] Neel Nanda (not verbatim)

1 | Mech. interp. (MI)

- To look into black (or opaque) machine learning (ML) boxes.
- Reverse-engineering deep learned circuits.
- Nanda et al. [7] shows MI on modular addition transformer.
- We show MI on prime number classification transformer.

1 | On complexity

- ML complexity (probably) grows faster than MI [9].

1 | On complexity

- ML complexity (probably) grows faster than MI [9].
- Lessons from MI might still inform ML development and risk.

1 | On complexity

- ML complexity (probably) grows faster than MI [9].
- Lessons from MI might still inform ML development and risk.
- Work on automatic MI [3]: ML might have a rôle in MI.

1 | On complexity

- ML complexity (probably) grows faster than MI [9].
- Lessons from MI might still inform ML development and risk.
- Work on automatic MI [3]: ML might have a rôle in MI.
- Current ML is sub-symbolic (lacks the rigor of formal systems).

1 | On complexity

- ML complexity (probably) grows faster than MI [9].
- Lessons from MI might still inform ML development and risk.
- Work on automatic MI [3]: ML might have a rôle in MI.
- Current ML is sub-symbolic (lacks the rigor of formal systems).
- Metric frenzy has made ML more engineering than science.

1 | Grokking

- When a model suddenly generalizes [8].
- Grokking means the weights represent an algorithm …
- … rather than a dataset.
- Good for MI, as it means circuits are there to be discovered.

2 | Transformers

- Famously introduced by Vaswani et al. [11].
- Batch normalization, residual streams, projections, etc.
- We use He and Hofmann [4]'s simplified transformer block.

"God made the natural numbers; all else is the work of man."[2]

[2] Leopold Kronecker (also not verbatim)

$3 \mid \mathbb{Z}$

- Complexity from trivial in seq. 1 to impossible in seq. 2 (busy beaver [1]).
- OEIS [10] is a big database of $\mathbb{Z}$-seqs.
- Four $\mathbb{Z}$ tasks: classify, compare, continue, and unmask [2].
- We focus on primes (seq. 3), which Lee and Kim [6] shows is doable.

$$0, 1, 2, 3, \ldots \tag{1}$$

$$6, 21, 107, \ldots \tag{2}$$

$$2, 3, 5, 7, 9, \ldots \tag{3}$$

3 | Irreducible integers[3]

- Given a sequence from $\mathbb{Z}$, which numbers are prime?
- Tests to determine primality include:
  - Wilson's Theorem: $n > 1$ is prime if $(n-1) \equiv -1 \mod n$.
  - Fermat's Little Theorem: $a^{n-1} \mod n = 1$, for $a < n$.
  - Euler's Criterion, AKS Primality Test, Miller-Rabin Primality Test, and more.

[3] Is what you call prime numbers when you really want the acronym of your project title to be MIIII.

3 | Irreducible integers (cont.)

Table 1: Four digit dataset with numbers and labels ($[\mathbf{X}|\mathbf{Y}]$).

x_0	x_1	x_2	x_3	y_0	y_1	y_2	y_3
1001	1003	1007	1009	0	0	0	1
1011	1013	1017	1019	0	1	0	1
$\vdots$							$\vdots$
9981	9983	9987	9989	0	0	0	0
9991	9993	9997	9999	0	0	0	0

4 | The MU puzzle [5]

- Can you get `MI` from `MU` by:
  1. Adding a `U` to the end of any string ending in `I`.
  2. Doubling the string after the `M`.
  3. Replacing any `III` with a `U`.
  4. Removing any `UU`.

5 | MIIII

- We something soemthign
- Then this
- Then this
- Then this

5 | MIIII (cont.)

- But then this.

5 | MIIII (cont.)

The algorithm

```
def fib(n):
    | -> return n :: []
    | Match P <- <> DIV (1 \in N)
```

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] Arthur Conmy et al. *Towards Automated Circuit Discovery for Mechanistic Interpretability*. Oct. 2023. DOI: 10.48550/arXiv.2304.14997. arXiv: 2304.14997 [cs].

[4] Bobby He and Thomas Hofmann. *Simplifying Transformer Blocks*. Nov. 2023. DOI: 10.48550/arXiv.2311.01906. arXiv: 2311.01906 [cs].

[5] Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*. 20th-anniversary ed. London: Penguin, 2000. ISBN: 978-0-14-028920-6.

[6] Serin Lee and S. Kim. *Exploring Prime Number Classification: Achieving High Recall Rate and Rapid Convergence with Sparse Encoding*. Feb. 2024. arXiv: `2402.03363 [cs, math]`.

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

[8] Alethea Power et al. *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*. Jan. 2022. DOI: `10.48550/arXiv.2201.02177`. arXiv: `2201.02177 [cs]`.

[9] Advait Sarkar. “Is Explainable AI a Race Against Model Complexity?” In: *Joint Proceedings of the IUI 2022 Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE*. Ed. by Alison Smith-Renner and Ofra Amir. Vol. 3124. CEUR Workshop Proceedings. Virtual Event, Helsinki: CEUR, Mar. 2022, pp. 192–199.

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

[11] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 2017. DOI: `10.48550/arXiv.1706.03762`. arXiv: `1706.03762 [cs]`.