Mechanistic Interpretability on Irreducible Integers

Noah Syrkis

February 9, 2024

1 | Mech. interp. (MI)

- Reverse-engineering neural network circuits.
- Nanda et al. [5] 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. [5] shows grokking in a transformer.
- Grokking means the weights represents an algorithm...
- ... rather than a dataset.

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 [7].
- 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

- MI on primes.
- Base 10 centric.
- Last digits $d_l \in \{1, 3, 7, 9\}$

4 | MIII (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 | MIII (cont.)

- I will focus on He and Hofmann [3]'s simple transformer (see sec. 5).
- With possible expansion into spiking neural nets.

5 | Simple trans. [3]

- Simple attention eq. 1

$$\mathbf{A}(\mathbf{X}) \leftarrow (\alpha I_T + \beta \mathbf{A}(\mathbf{X})) \quad (1)$$

5 | Simple trans. [3] (cont.)

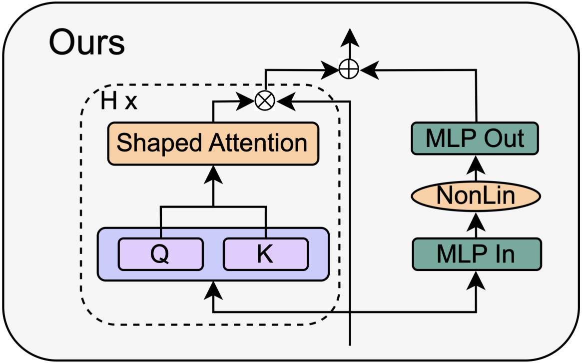

Figure 1: He and Hofmann [3]'s transformer block

6 | Spiking NN

- I think spiking neural nets are cool.
- More brain like.
- Could be cool to do mech interp on them.
- See Olin-Ammentorp and Bazhenov [6] and Hu et al. [4]

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

[4] Di Hu et al. "Digital Implementation of a Spiking Neural Network (SNN) Capable of Spike-Timing-Dependent Plasticity (STDP) Learning". In: *14th IEEE International Conference on Nanotechnology*. Aug. 2014, pp. 873–876. DOI: 10.1109/NANO.2014.6968000.

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

[6] Wilkie Olin-Ammentorp and Maxim Bazhenov. *Deep Phasor Networks: Connecting Conventional and Spiking Neural Networks*. Sept. 2021. DOI: 10.48550/arXiv.2106.11908. arXiv: 2106.11908 [cs].

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