Final LTFF Mamba Interp Progress Report

Danielle Ensign

September 2024

1 Difficulties with IOI circuit analysis

Our task was to identify a circuit on Mamba that is responsible for doing IOI (indirect object identification), and then later apply similar techniques to other tasks.

Unfortunately this hit a snag. We implemented traditional techniques on Mamba (resample ablation, path patching, ACDC, EAP) and they did not result in an informative circuit aside from the information we obtained on Layer 39.

The problem seemed to be that mamba was simply "enriching" tokens inplace, and all token information movement was occuring on layer 39. Because most traditional circuit analysis is about token movement, this limited how deep our analysis could go.

Regardless, we wrote up our results and published them in the ICMR workshop, presented as a Poster. See https://openreview.net/forum?id=lq7ZaYuwub, Investigating the Indirect Object Identification circuit in Mamba for further details.

We also published all code for our experiments at
https://github.com/Phylliida/investigating-mamba-ioi

2 SAE Analysis

In order to analyze behavior in the embedding space, we decided to pivot to SAEs. We trained top-k SAEs on the inputs to layers 0-28 of Mamba, and then ran attribution patching with integrated gradients on their outputs.

After obtaining attributions, we sort vertices in the SAEs by attribution score. For clarity, choosing vertex i (for some layer) means we keep the value of feature i unchanged. All non-chosen vertices have their feature value replaced with a corrupted run's value.

We can then binary search to find the smallest set of vertices that obtain a target normalized logit diff. This technique is standard, we are just applying it to Mamba.

We ran extensive ablation studies and found that

- Running EAP on SAEs and non-SAE edges at the same time resulted in much larger sets, so we recommend just studying SAEs alone. This is true even if we only count the number of SAE vertices in the resulting set. This is most likely because the attribution approximation becomes worse/noisy when patchings are done on SAEs and edge connections at the same time (vs just patching SAEs).

- Layer 0 had the majority of high attribution features, which is consistent with it being an extra embedding layer (as our previous experiements have suggested). We recommend excluding layer 0 features from the search.

- Smaller sets (20-30% smaller) could be obtained by training the SAEs on the the value added to the residual steam instead of inputs to layers. Thus we retrained SAEs for this instead. The justification for this is that layer inputs need to contain information from all previous layers, whereas layer outputs only need to contain information from that layer.

- Smaller sets (about 10% smaller) could be obtained by multiplying attribution score by count before sorting, and filtering by features that occurred more than 1/10th of the time. This was done because we observed many "outlier features" that had high attribution scores, but were only non-zero for a single data point. These were not useful to do well on the task in general, so it was better to ignore them.

Given these sets of features, we sorted them by attribution and found top-k activations on a large subset of the pile.

Using these, we could identify what the features are responsible for.

You can find our code at

https://github.com/Phylliida/mamba$_i$*nterp/blob/main/main_notebook.ipynb*

It is in the progress of being cleaned up into a library to help others to do SAE EAP circuit analysis on Mamba.

3 Feature analysis

We initially started with Layer 15, as resample ablation suggested 15 was worth investigating.

There were many standard features (spooky/Halloween features, weed/420 features, features for each letter, features for Hispanic names, Scottish names, etc.) that are also observed in transformers.

However, there were a few "copy" features that seemed worth further investigation. These were features that appeared to do the following:

- Only fire on a subset of tokens, and fire quite weakly.

- For a subset of those, fire highly the second time that token occurs

- For other tokens, only fire the first time that token occurs.

We could not identify a pattern in which subset they chose (besides simpsons-adjacent things firing more).

We plotted the pattern over repeating these tokens many times and found rich behavior that was difficult to explain cleanly.

Their primary purpose appeared to "fire on duplicate" but in a way that partitions tokens.

Using these features we constructed a hypothesis for how Mamba could be doing IOI: "duplicate name" tokens and "fire once" tokens together are sufficient to accomplish it on all of the examples we tested on. In the future, we will investigate this further.

4 Copy task

These sorts of features don't occur much in transformers, so they seemed worth investigating as a potential primitive that is unique to Mamba. Thus, we decided to investigate the copy task, to detect more of these features and investigate how they function.

In transformers, induction heads suffice to do the copy task well. We've only had a few weeks to study mamba's top features on the copy task, but so far every feature (sorted by attribution scores) appears to be of this "copy" form discussed above. Further work can investigate them in more detail and see how these primitives operate.

We were *not* able to find any induction features (AB...A→B). This does not mean they do not exist, but further study is required.

5 Conclusion

Overall, the results of our research are mixed. Techniques that work on Transformers also appear to work on Mamba, and give similarly sparse results. However, these are not easily interpretable, perhaps owing to:

1. The smallest mamba model that is worth studying is quite large, with 48 layers.
2. There are no "heads", which prevents more fine-grained analysis without resorting to SAEs.

SAEs result in features that appear to be similarly interpretable through traditional methods, and did not appear to be much more difficult to train. However, some of the features are copy features, which seem to be novel and are worth further study.