Short Summary

OpenAI: Toward Mechanistic Interpretability (MI)

Kris Carlson, Publisher and Editor-in-Chief

OpenAI Blog post

OpenAI Paper: Gao et al., Weight-sparse transformers have interpretable circuits

OpenAI explores reducing net density to see if a sparser net is more understandable than a denser one, and if they can then identify the specific circuit in the net that is computing a given function of interest. In both cases, the answer is yes, given some trade-offs. Here's are a couple of schematics of their goal:

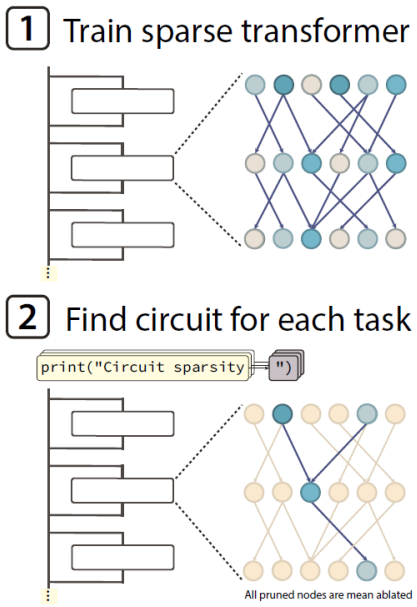
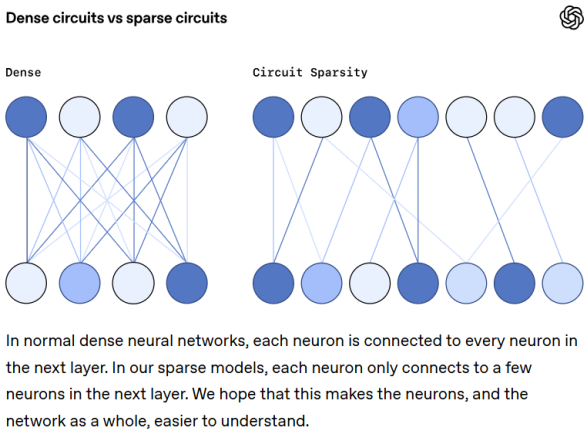


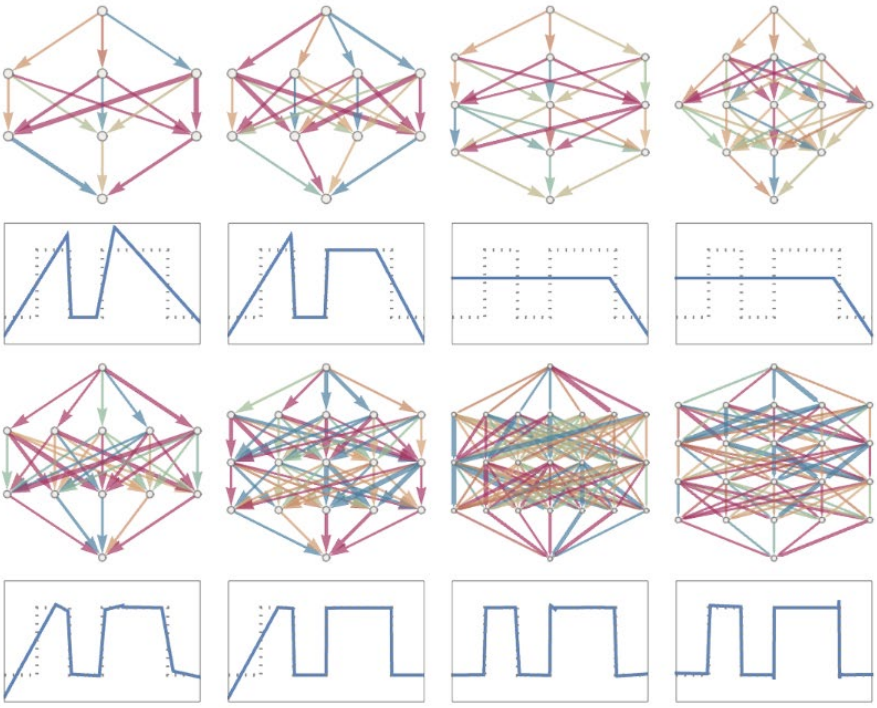
Figure 1. An illustration of our overall setup. We first train weight-sparse models. Then, for each of a curated suite of simple behaviors, we prune the model down to the subset of nodes required to perform the task. We ablate nodes by pruning to their mean activation value over the pretraining distribution.



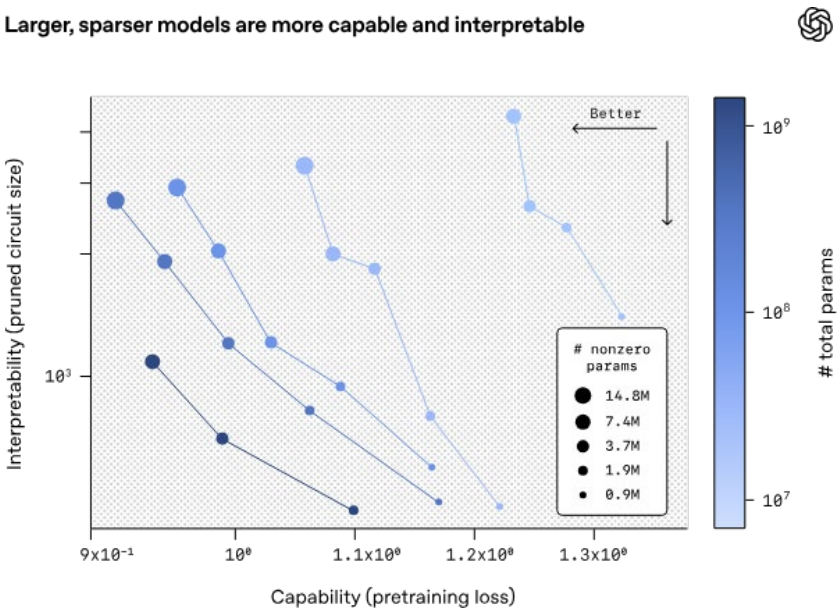
Wolfram did something similar in August 2024. But he didn't identify circuits performing specific tasks, he looked at the minimal size net required to reduce loss to acceptable levels, and then he went in different architectural directions and experimented with discrete transfer functions and evolutionary programming.

What's Really Going On in Machine Learning? Some Minimal Models:

But what happens if we use a different number of parameters, or set up the architecture of our neural net differently? Here are a few examples, indicating that for the function we're trying to generate, the network we've been using so far is pretty much the smallest that will work:



Here are some results in OpenAI's analysis of using sparser nets to induce interpretability. Lower-left is better, higher-right is worse: Smaller circuit size (y-axis, log) is better provided Capability (x-axis, low error in predicting next token, i.e., toward 0 is better) is preserved. Darker circles mean more total parameters, so total LLM size, while circle size shows sparsity, where smaller is sparser, and better since a sparser net is generally more interpretable than a denser one. But clearly there is a trade-off between total size and error. It's a lot to pack into one graph. From Fig. 2 in their paper, it appears they can reduce density by ~94% and get equal loss to the denser net.



We plot interpretability versus capability across models (lower-left is better). For a fixed sparse model size, increasing sparsity—setting more weights to zero—reduces capability but increases interpretability. Scaling up model size shifts this frontier outward, suggesting we can build larger models that are both capable and interpretable.

I don't think MI will be scalable in the sense that, as Wolfram predicts in his Bigger Brains post, if AI has a vocabulary of 10^5 – 10^6 or more concepts, compared to 10^3 – 10^4 for humans, we will be able to understand its more complex sub-circuits, which may be the bulk of them, as it is with humans compared to lifeforms with much smaller brains.

[What If We Had Bigger Brains? Imagining Minds beyond Ours]