Musings on Understanding of the Large Language Models

Written by David Sauerwein on Linkedin, May 31, 2025

Nobody knows how Large Language Models manage to display their impressive capabilities - they are incredibly powerful black boxes. Anthropic has now released methods to peek under the hood. It's an exciting step towards more interpretable LLMs.  
  
The transformer architecture contains very few building blocks, making it even harder to understand how they lead to these advanced, emergent capabilities. Here are two example challenges:  
  
1) Polysemantic neurons: The LLM needs to represent more concepts than it has neurons available. As a result, one neuron represents parts of many unrelated concepts, making it hard to understand what it means when it fires.  
  
2) Connectivity: Information passes sequentially through the transformer layers. It's unclear how concepts from earlier layers influence later ones.  
  
Anthropic's approach to untangle this is to create mechanistic interpretations of the transformer components in human-understandable language.  
  
They identify interpretable building blocks (features) the model uses, then describe processes (circuits) by which features interact to produce outputs.  
  
Concretely, Anthropic creates a second model (transcoder) that mimics the original LLM but with key differences:  
  
1) More neurons: More neurons in specific layers allow separate concept representation in individual "features."  
  
2) Direct connections: Transcoder layers receive direct input from all earlier transoder layers, passing features directly between layers.  
  
3) Sparsity penalty: The loss penalizes activating too many features per layer. This encourages assigning information across independent features instead of creating concept superpositions in single neurons.  
  
Anthropic provides interesting insights based on this method. For example:  
LLMs produce coherent output over thousands of tokens while only predicting the next token. But how do they think ahead?  
  
The creation of poems illustrates this particularly well:  
If the first line is "He saw a carrot and had to grab it," the next sentence must rhyme with "it." Indeed, the model continues with "His hunger was like a starving rabbit.", and Anthropic's transcoder model shows how the "rabbit" concept builds up way before the word itself appears.  
  
What are problems with/questions around this new approach?  
  
1) The transcoder isn't the original model. It's explanations might not apply to the LLM. This is a well-known problem with these "surrogate model" problem throughout ML.  
  
2) Why not train the transcoder directly? Sparsity and connectivity make it much harder to train with lower accuracy.  
  
3) Feature graphs are heavily pruned by humans, risking biased review and anthropomorphizing by the people analyzing the results.  
  
Despite this, it's exciting research with well-written, interactive papers and open-sourced analysis tools (see comments).  
  
The mechanistic approach to LLM interpretability is hard. But Anthropic has made great progress and I'm excited to see where the community goes next!

Reference

[Linkedin link of the article](https://www.linkedin.com/posts/davidsauerwein_ai-genai-llm-activity-7334325994964594688-8L4d?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAFZfUoBgPoGUucdnvtwuzPv79P8VHj6uvk)