ANTHROP\C

Mapping the Mind of a Large Language Model

This memo is a summary of research conducted at Anthropic. Templeton et al. (2024)
https://www.anthropic.com/research/mapping-mind-language-model

POLICY HIGHLIGHTS

- Dictionary learning is a technique utilized in the field of interpretability to uncover “features” which represent distinct concepts learned by large language models.
- Certain features are relevant to the safety of AI systems and map to concepts such as bias and toxicity, dangerous materials, insecure code, or manipulative behavior.
- When those features are forced to fire or turned off, we see altered model responses. For example, a response may become more or less biased in response to the same prompt when the “gender bias” feature is altered.
- Anthropic has now successfully applied this technique to full-scale models used in production, representing a major advance from earlier proof-of-concept experiments.
- The use of dictionary learning to understand model behavior is still nascent, but this work points to potentially promising avenues for new AI safety interventions.

Large language models (a type of artificial neural network) are often referred to as “black boxes” because we do not fully understand how models work. The field of **mechanistic interpretability seeks to understand these models at a detailed, mechanistic level by carefully examining their internal components, and determining how those components contribute to model behavior.**

This is challenging, as artificial neural networks are highly complex systems often comprising billions of “neurons”. Similar to how neurons in the human brain fire in response to specific inputs, neurons in language models also fire in response to specific concepts. However, a typical neuron in a large language model fires on multiple, unrelated inputs, a phenomenon referred to as “polysemanticity”. This means that the same neuron might fire on concepts as disparate as the presence of semicolons in computer programming languages, references to burritos, or discussion of the Golden Gate Bridge, giving us little indication as to which specific concept was responsible for activating a given neuron.

Descriptions of, or references to, the **Golden Gate Bridge**

```
WONDER OF THE WORLD," "THE GOLDEN GATE BRIDGE
come hurtling in through the Golden Gate Bridge
Golden Gate Bridge wind resistance barriers
I know the⏎⏎golden gate bridge often does,
```

Instead of dissecting model behavior at the individual neuron level, **Anthropic applied a technique called “dictionary learning” which uncovers patterns of activated neurons**, collectively referred to as “features”. Unlike individual neurons that are largely uninterpretable, we can interpret features because they map to specific, discrete concepts. Using the example above, we would expect to find one feature in the model that recognizes semicolons used in computer programming languages, another for burritos, and a third for the Golden Gate Bridge. Looking at features, **we can now begin to understand model behavior by seeing which features respond to a particular input, thus giving us insight into the model’s “reasoning” for how it arrived at a given response.**

[1] Read more about Anthropic’s previous research into dictionary learning at: https://www.anthropic.com/news/decomposing-language-models-into-understandable-components

MAY 2024

Building on past proof-of-concept work using dictionary learning on small "toy" models, **researchers at Anthropic have begun to apply this technique to full-size models used "in the wild"**: specifically, Claude 3 Sonnet. This represents a dramatic scale up in the use of dictionary learning, from tiny models purpose-built for testing, to ones several orders of magnitude larger and used by customers everyday. The fact that dictionary learning works not just in a "lab" setting on artificially small models, but also on large-scale production models, suggests it may be an effective tool for identifying—and potentially modifying—the components that affect model behavior in systems used throughout the economy.

Studying and manipulating safety-relevant features

With dictionary learning, we were able to identify several features that relate to safety-relevant concepts. These features map to a wide variety of risks including: bias and toxicity, insecure code, fraudulent or dangerous activity, and manipulative behavior, among several others.

Once we identify a particular feature, we can artificially stimulate it to enhance or suppress that concept and investigate its effect on model responses. For example, Anthropic researchers identified a feature corresponding to "unsafe code," which fires for pieces of computer code that disable security-related system features. When we prompt the model to continue a partially-completed line of code without artificially stimulating the "unsafe code" feature, the model helpfully provides a safe completion to the programming function. However, when we force the "unsafe code" feature to fire strongly, the model finishes the function with a bug that is a common cause of security vulnerabilities.

Generally, we see specific concepts within model responses manifest more strongly or become more subdued in response to manipulating a given feature, almost as if there were a "dial" that can be turned to tune the effect of that particular feature. In addition to forcing the "unsafe code" feature to fire, the researchers found that they could manipulate the effect of other features to make model responses more or less biased, more or less willing to draft a scam email, more or less willing to provide instructions on making dangerous items, among several other behaviors.

The ability to manipulate features may provide a promising avenue for directly impacting the safety of AI models, though the research here is in its early days. For example, it may be possible to prevent model jailbreaks by suppressing or "turning off" features related to unsafe behavior.

We are in the early days of dictionary learning

While dictionary learning has given us a tool to begin to understand which model components respond to particular inputs and how those components affect model responses, the research is still quite nascent. We haven't yet verified that manipulating specific features results in predictably safer behavior.

Additionally, we believe that we have only found a small subset of all of the features learned by the model during training and unfortunately, finding all available features would be cost-prohibitive (the computation presently required would vastly exceed the compute used to train the model in the first place). This is a significant limitation, since it's possible the "missing features" could implement safety-relevant behaviors, limiting our confidence in safety interventions based on the ones we've found. Finally, even if we were able to identify all possible features, we still would not have all the information needed to fully understand the inner workings of the model.

Our use of dictionary learning has dramatically improved our ability to break down the complexity of large language models into more understandable features, providing us with a better understanding of their internal workings and a potential tool for steering model responses. As we continue to apply this technique to large-scale production models, we hope to unlock new safety interventions and apply them to AI models that are used widely.

ABOUT US

Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems.