Peek inside a neural network

Starter · 1–3 hours · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/peek-inside-a-neural-network/

Summary

Use Neuronpedia, a free browser tool, to see which internal “features” of a language model light up for different ideas, then try to break your own explanation of one.

Why it matters

Interpretability aims to understand models from the inside, so we are not relying only on what they say. Getting a feel for features is the first step towards that whole research field.

What you will have at the end

Notes on three features: what you think each represents, and where your guess failed.

Steps (tick each box as you go)

- [ ] **1. Open Neuronpedia (10 min)**
  Go to Neuronpedia and pick a model with sparse-autoencoder features (Gemma Scope is a good choice). Search for a concept you like: a city, an emotion, a programming language.

- [ ] **2. Read one feature’s examples (15 min)**
  Open a feature from your search. Look at the text snippets that activate it most strongly. Write down, in one sentence, what you think it represents.

- [ ] **3. Test your guess with your own text (20 min)**
  Type new sentences into the feature’s test box. Try ones that should activate it and ones that should not. Where does your one-sentence guess break?

- [ ] **4. Repeat for two more features (20 min)**
  Choose features for quite different concepts. Note which were clean and which were messy or surprising.

- [ ] **5. Write it up (10 min)**
  Summarise: three features, your best description of each, one surprising failure. Add one question you would want a researcher to answer.

What you need

- A web browser
- About an hour

What you will learn

- What interpretability researchers mean by a “feature”
- How to test an explanation instead of trusting it
- Why understanding models from the inside is hard

Key terms

Interpretability
Research into what happens inside an AI model: which internal parts represent which ideas, and how they combine to produce an answer.

Feature
A direction inside a model’s internal activity that seems to stand for a recognisable concept, like “the Golden Gate Bridge” or “code that contains a bug”.

Sparse autoencoder (SAE)
A tool trained to split a model’s tangled internal activity into many separate, more understandable features.

Source and links

Neuronpedia (open interpretability platform)
https://www.neuronpedia.org/

Gemma Scope paper (the features you will explore)
https://arxiv.org/abs/2408.05147

Anthropic — Mapping the mind of a large language model
https://www.anthropic.com/research/mapping-mind-language-model

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Try next: Find a pattern-copying circuit
https://ai-safety-sandbox.vercel.app/projects/find-a-pattern-copying-circuit/

Try next: Does interpretability answer the octopus test?
https://ai-safety-sandbox.vercel.app/projects/interpretability-and-the-octopus-test/

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes