Find a pattern-copying circuit

Intermediate · A weekend · Free Colab / laptop · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/find-a-pattern-copying-circuit/

Summary

Follow a guided notebook to find the induction heads in a small transformer, the tiny circuit that lets models continue patterns they have seen earlier in the text.

Why it matters

Induction heads are the classic “hello world” of mechanistic interpretability. Finding them yourself teaches the tools (attention patterns, ablations) that interpretability researchers use.

What you will have at the end

A notebook showing the attention pattern, an ablation, and your explanation.

Steps (tick each box as you go)

- [ ] **1. Set up TransformerLens (30 min)**
  Open the ARENA interpretability materials (or the TransformerLens docs) in Google Colab. Load a small model such as GPT-2 small.

- [ ] **2. Make a repeated-sequence test (1 hr)**
  Feed the model random tokens repeated twice. Check that loss drops on the second half. That is the induction behaviour you are hunting for.

- [ ] **3. Find the heads (1–2 hr)**
  Plot attention patterns and look for heads that attend from a token to the token after its previous occurrence. Score every head this way.

- [ ] **4. Prove they matter (1–2 hr)**
  Ablate (switch off) your candidate heads and show that the in-context pattern-copying gets worse. Correlation is not enough. This is the causal test.

- [ ] **5. Explain it in plain English (1 hr)**
  Write a short post: what induction heads do, how you found them, and one thing that confused you.

What you need

- Python and basic PyTorch
- Free Google Colab
- A weekend

What you will learn

- Hands-on mechanistic interpretability
- Attention patterns and ablations
- TransformerLens

Key terms

Mechanistic interpretability
Interpretability that tries to reverse-engineer a model's internal computations into understandable mechanisms, a bit like reverse-engineering a compiled program.

Induction head
A simple, well-studied circuit in language models that spots a repeated pattern ("A B … A") and predicts what came next last time ("B").

Circuit
A small group of components inside a neural network that work together to perform one identifiable job.

Source and links

Anthropic — In-context Learning and Induction Heads
https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

ARENA materials
https://github.com/ARENA-education/ARENA_materials

TransformerLens
https://github.com/TransformerLensOrg/TransformerLens

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: Tackle a beginner-graded open problem in interpretability
https://ai-safety-sandbox.vercel.app/projects/open-problem-in-mech-interp/

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