Mechanistic Interpretability

11/10/2025

60 Points Possible

Attempt 1

In Progress

NEXT UP: Submit Assignment

Add Comment

Unlimited Attempts Allowed

Details

Mechanistic Interpretability Assignment

Explaining a Tiny Brain

This assignment invites you to step into the role of a "neuron detective" — your mission is to build a tiny neural network, explore how it works from the inside, and explain your findings as a story supported by code and visualizations.

Instructions

➡️ We have provided a starter notebook here: https://github.com/AIPI-590-XAI/Duke-AI-XAI/blob/main/assignments/mechanistic_interp_starter.ipynb (https://github.com/AIPI-590-XAI/Duke-AI-XAI/blob/main/assignments/mechanistic_interp_starter.ipynb)

Note: you must use the starter as just that - a starter. You will need to create your own model, task, and methods for this assignment.

Part 1 – Setup (Train Your Own Tiny Model on a Tiny Task)

Build and train your own tiny model on a toy task. This should be something simple that trains in under 10 minutes, such as:

- XOR or parity classification
- Reversing short sequences
- Any other small, interpretable task you invent

Part 2 – Explore

Dive into the internals of your model. Consider the following:

- Inspect weights, activations, or attention maps
- Try changing the input and observing how the internal state (e.g., hidden layer activations) responds.
- Identify at least one neuron, attention head, or component that seems to be doing something meaningful.

> *Focus on forming a mechanistic hypothesis: What does this part of the model "care about"? What feature is it detecting or encoding?*

Part 3 – Explain

Turn your observations into a story.

- Write a clear, concise explanation of what you believe one part of the model is doing.
- Use visuals (plots, heatmaps, diagrams) to support your explanation.
- Present this like a mini "computational case study" — tell us what the neuron or head is doing and why you think so.

Part 4 – Reflect

Include a short written reflection:

- What did you learn about how your model works?
- What was confusing, surprising, or challenging to interpret?
- What's one thing you wish you could understand better or explore further?

Submission

Submit a GitHub Repository link with a Google Colab Notebook that includes:

- All code, visualizations, and outputs you used to explore your model.
- A clear, well-organized explanation of one interpretable feature/component in the model.
- A brief reflection on what you learned and how you approached interpretation.
- A structure that reads like a mini blog post or report: narrative, code, and insight woven together.

Rubric

When grading, we will focus on:

- Clarity of thinking
- Creativity of explanation
- Evidence that you looked inside the model and formed a hypothesis
- Effective use of visuals and storytelling

Edit View Insert Format Tools Table

12pt Paragraph

p

0 words </> + −

(https://canvas.duke.edu/courses/62464/modules/items/572898)

S
Assignment

(https://canvas.duke.edu/courses/62464/modules/items/572860)