Mechanistic Interpretability

Looking Back
WT 2024/25

Frederick Riemenschneider

06.02.2025

Key Ideas

Fundamentals

Othello

- board game internally represented
- non-linear and linear probes
- black/white vs. mine/yours

Transformer Circuits

- IOI task
- responsibilities
- induction heads in action
- backup name mover heads
- path patching
- automated discovery with ACDC

Activation Patching

- knowledge stored in specific MLPs
- knowledge as dictionary lookup
- rank-one model editing
- different ways to think about paths in a transformer
- practical applicability of model patching

Dictionary Learning

- language models explain neurons
- sparse autoencoders to disentangle polysemanticity
- gated sparse autoencoders
- difficulty of evaluating autoencoders

Self-conditioning

- expert neurons predictive of concepts
- area under the precision-recall curve
- guided (self-conditioned) text generation
- languages as concepts

Grokking

- optimal semantic space arrangement leads to true generalization
- artifact of suboptimal training parameters
- transformer truly understandable?
- similarities to Othello

Multilingual Language Models

- language-specific vs. language-agnostic neurons
- LAPE and PROBELESS
- language models have a default language
- perturbation successful

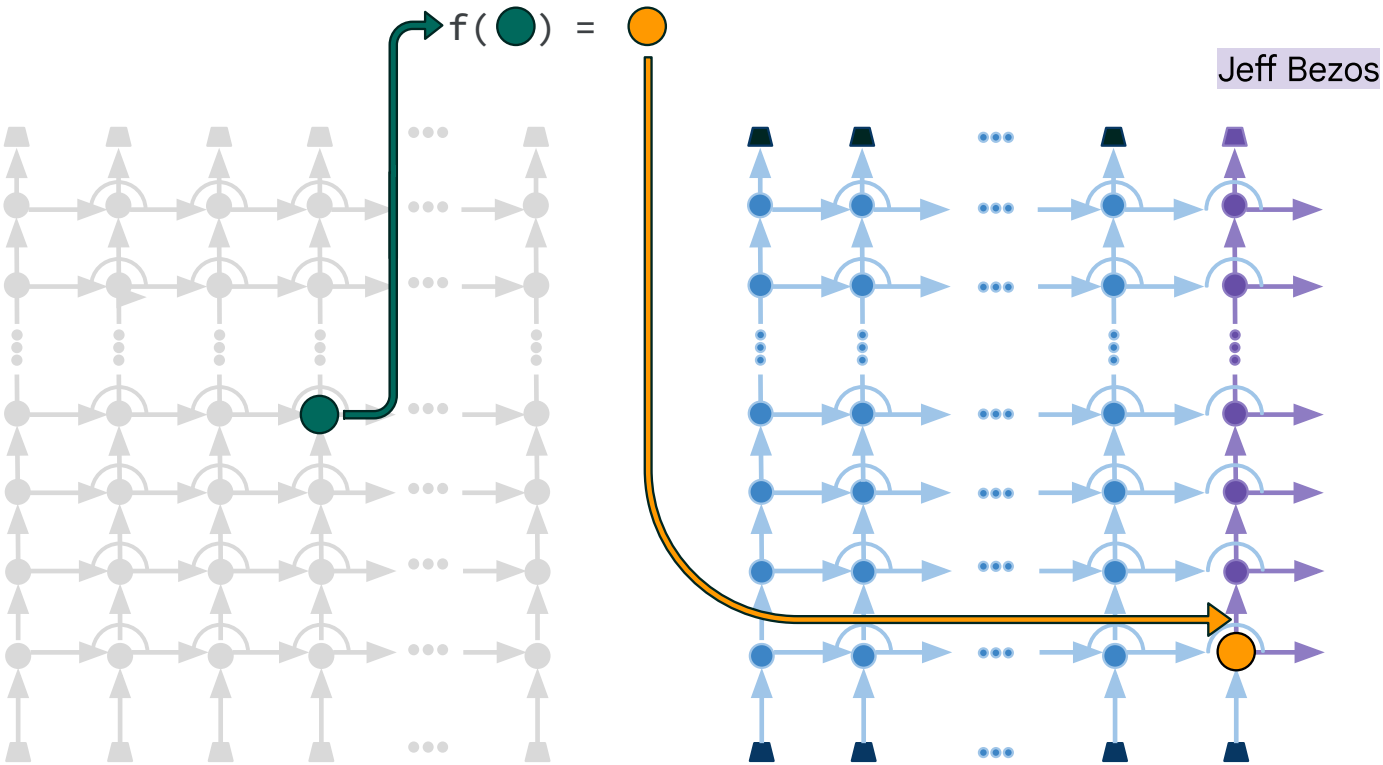
Frameworks

Step 1: Feeding **Source Prompt** to **Source Model**

Step 2: **Transforming Hidden State**

Step 3: Feeding **Target Prompt** to **Target Model**

Step 4: Running Execution on **Patched Target**



Detecting Lies in LLMs

- potential use-case
- two truth directions
- negated vs. affirmative statements

Your Comments

Applications

"How can we actually apply insights from Mechanistic Interpretability to advance machine learning systems?"

Subject of Investigation

"Are decoder-only transformers the only thing we should study?"

“Understanding”

“On one hand, if LLMs are not interpretable, they could be risky in areas like healthcare and autonomous driving. On the other hand, end-to-end assisted parking systems work so well that people do not worry about errors. Maybe in the future, we will accept LLMs even if we do not fully understand them.”

More Topics

"Are there any other topics in the field of Mechanistic Interpretability that we didn't cover in the seminar?"

Outlook

"What are current trends?"

Implementation Project

Implementation Project

- independent (!) reimplementation of one of the approaches
- exploration of one of my/your own ideas
- report: max. 8 pages
- deadline: 31st of March

Implementation Project

- Please present your work in a way that is both **clear** and **engaging**.
- Provide a **description** of your experiments, including:
  - the **motivation** behind them
  - the **methodology**
  - a **discussion** of your results
- You can choose **any format** that suits your work, such as a Jupyter Notebook with code and documentation. If **in doubt, ask me!**
- Please ensure your results are **easily reproducible**. I personally recommend dependency management with pdm. Ideally, I should be able to run a single script to reproduce your results (or one experiment).

Presentation

- By default, your project will be linked on the course page, so make sure it is appealing and accessible to your fellow students.
- If you prefer not to share your project, you can opt out. However, sharing is encouraged. It is more fun to see what others have created!

Plagiarism

- Please make sure to include a signed Declaration of Independent Work: https://www.cl.uni-heidelberg.de/-studies/Eigenstaendigkeitserklaerung_DE.pdf.
- You can find a non-binding English translation here: https://www.cl.uni-heidelberg.de/-studies/Eigenstaendigkeitserklaerung_EN.pdf
- general information: https://www.cl.uni-heidelberg.de/studies/faq/faqMa-plagiarism.mhtml#plagiarism001

Language Models

- my take:
  - proofreading, phrasing improvements, plotting and visualization assistance, basic coding support is entirely okay
    - basic coding support: using Copilot to complete trivial lines of code, auto-generating code documentation
  - **All other uses require prior discussion!**

Project Ideas

- probably fruitful: apply a given technique to a new question
- Can we evaluate autoencoders when testing them on artifical data?
- Can we find syntax circuits? Are they shared across languages?
- Do language models "think" in one specific (maybe abstract or predominant) language?
- Do multilingual language models have an accent that is influenced by other languages?