Extending “Towards Monosemanticity”

“Towards Monosemanticity” Summary

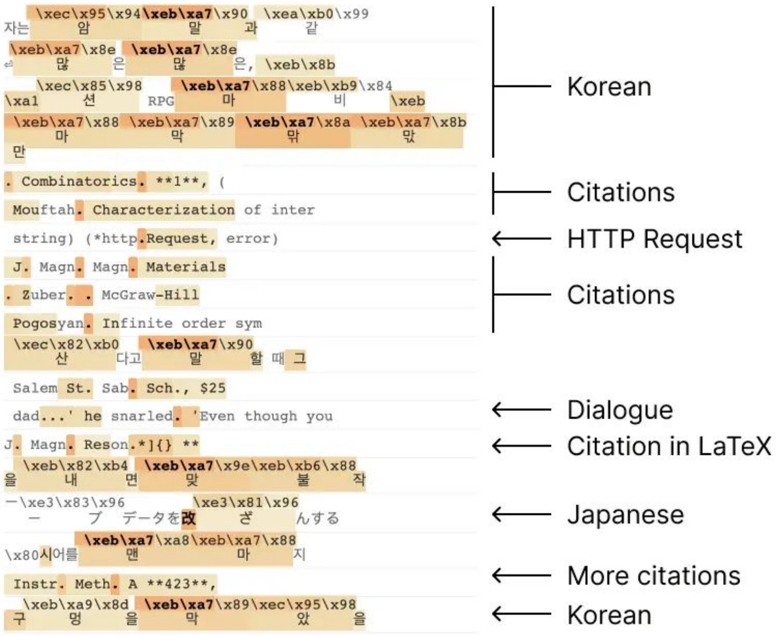
Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, Chris Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. 2023. (Anthropic)

Motivation

- Individual neurons do not have consistent relationships to network behavior (Superposition)
  - a small model’s neuron can activate with: academic citations, English dialogue, HTTP requests, and Korean text
- What is a better unit of analysis than a neuron?
- Model Steering
  - If we effectively separate individual neurons, we may have more control over model outputs in R&D settings.

Individual Features are Interpretable

Neurons in language models fire on many different types of text. **Neuron #83** fires on...



The features we find are dramatically more consistent. **Feature #2937** fires on DNA.

Primary Methods

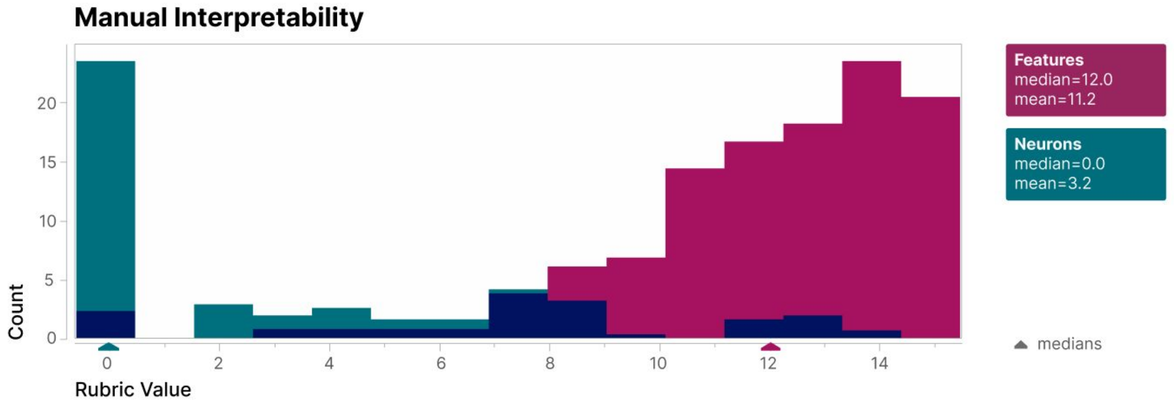
- Using Dictionary Learning to project an MLP output layer with 512 neurons to a higher dimension
    - Tested 1x, 8x, 32x, 64x, 128x, 256x (primary results use 8x version)
    - Trained as an autoencoder (input weights as an encoder and output weights as the decoder)

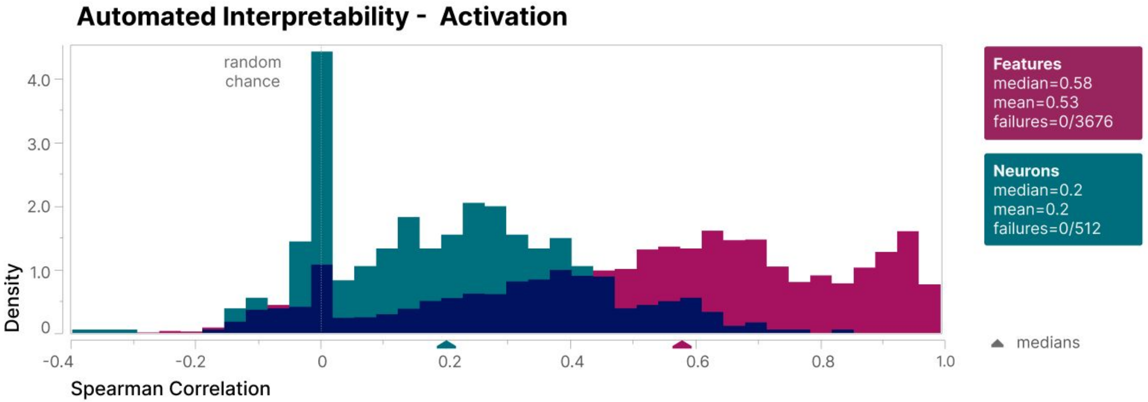
- Generate concepts related to activated tokens using an LLM
    - Bills et al., 2023 (OpenAI) “Language models can explain neurons in language models”
        - Used a few-shot prompt to generate natural language concepts based on a set of (token, quantized activation) tuples
        - Validated concepts by prompting the LLM to predict quantized activations for masked tokens

Evaluations

- Validate feature interpretability / faithfulness
  - Automatically, using an LLM (Bills et al., 2023 OpenAI)
    - computed the Spearman correlation coefficient between the predicted activation and the true activations (n=540 per activation)
  - Manually, with a human evaluator scoring interpretability
    - confidence in an explanation
    - consistency of the activations with that explanation
    - consistency of the logit output weights with that explanation
    - specificity

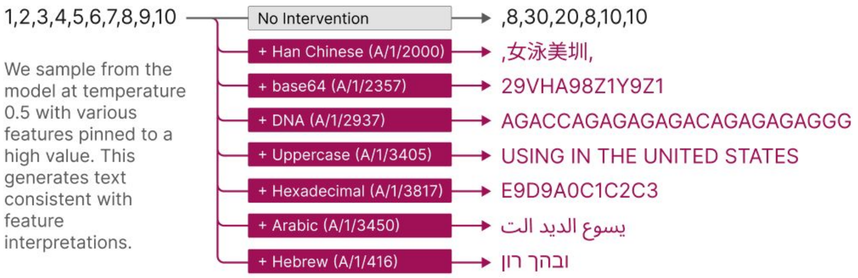
How Interpretable is the Typical Feature?





Caveats on Results

- [Do features tell us about the model or the data?]
    - Use Logit weight inspection, Feature ablation, and Pinned feature sampling
    - Feature ablation results seem convincing



- [How much of the model does our interpretation explain?]
    - 79% reconstruction loss
    - This number does not necessarily answer the question

My Extension

Why not use a bigger model?

- “overtrain the underlying model”
  - Anthropic hypothesized that a “very high number of training tokens” may cause cleaner representations

- smaller dictionaries (autoencoders) can be used
  - Fewer "true features" than larger models, learned by smaller dictionaries are cheaper to train and faster to experiment with

- approx. linear feature to logit mapping
  - Theoretical justification that learned features actually reflect the functionality of the model and not the underlying data data

Extension: Methods

1. Use distillGPT-2 (42M param., 6 layers)

2. Train autoencoder using last layer’s MLP with limited hyperparam. tuning
    a. dictionary_size = {8, 32}

3. Used automated interpretability methods with GPT-4

Implementation Details

- neelnanda-io/TransformerLens package for interpretability research
  - Relatively small user base so had to handle confusing documentation or limitations of the package
- Used slurm script to train various autoencoders on the CS department cluster
- Autoencoders were trained on MLP layer outputs (retrieved using a hook with TransformerLens)
  - Full details of the autoencoder can be found in the original Anthropic paper’s appendix

Extension: Objectives (Assumptions / Evaluation Criteria)

- Test how well the methods from “Toward Monosemanticity” generalize
  - Interpreting a larger model (distillGPT-2)
  - Automated interpretability with a different LLM (GPT-4)
    - Note that the original paper which proposed this method used an older GPT-4

- Assumptions
  - The public API provides the necessary information to perform automated neuron explanations

Extension: Uncertainty Analysis

- Parameters that affect our result
    - A number of hyperparameters are used which could influence the results
        - Adam optimizer (Learning Rate 1e-4, β_1 0.9, β_2 0.99)
        - L1 Coefficient 3e-4
        - Dictionary size (8x, 32x)
        - Number of tokens to train autoencoder on (2 billion, 3 billion, 4 billion*)
            - Token dataset (Pile)
    - Explanation model (GPT-4)
    - Source model for reconstruction (distillGPT-2)

*tested to determine if training on more data would increase reconstruction score, noticed performance degradation even when setting a lower learning rate of 1e-5

Extension: Results (preliminary)

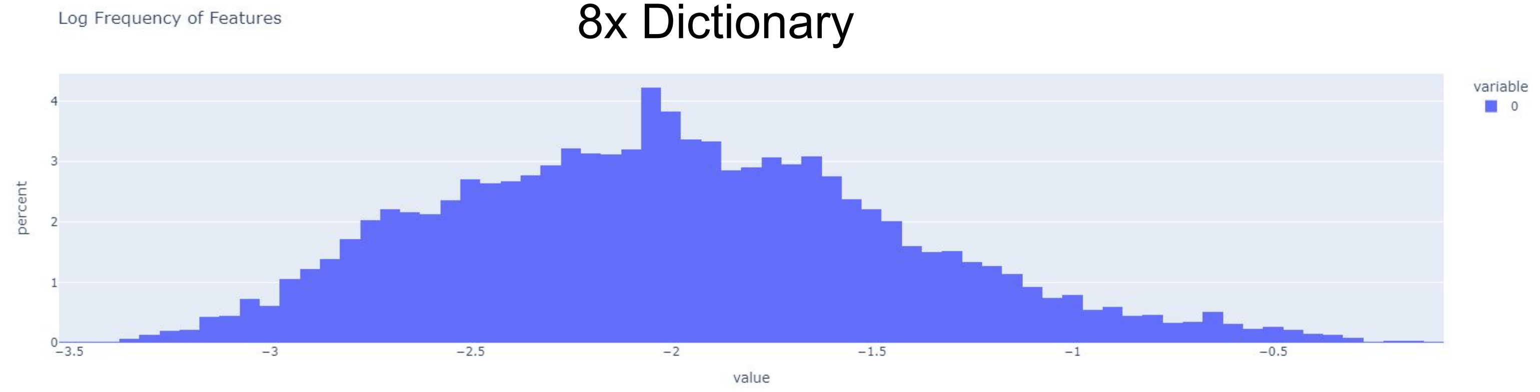
- Reconstruction scores*
  - 8x: 62.70%
  - 32x: 77.52%

- Attempting to use code from the Bills et al.
  - API has changed significantly, various parts need to be refactored
  - **WIP**

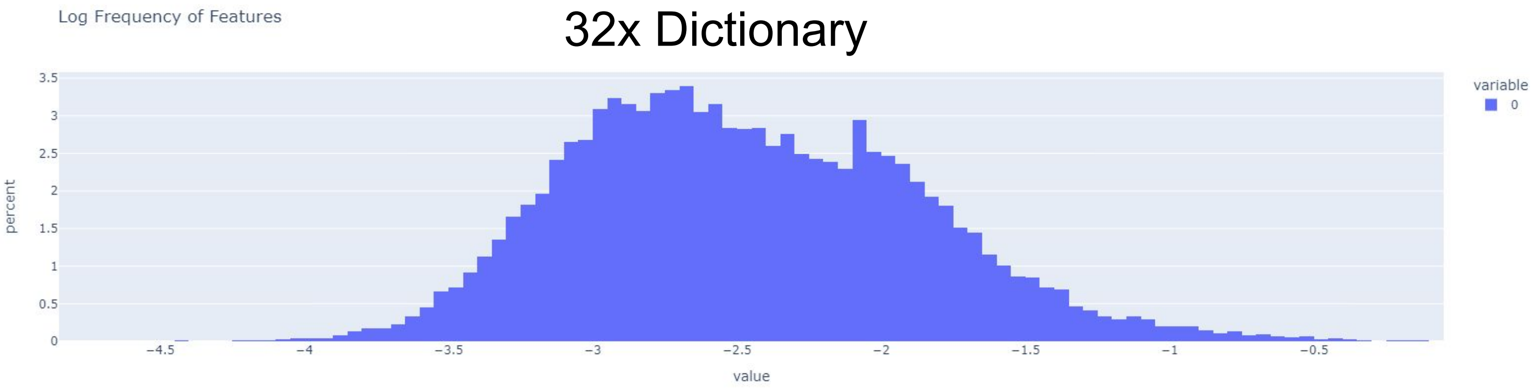
*score = ((zero_abl_loss - recons_loss)/(zero_abl_loss - loss))

Key graphs: Log Freq. of Features

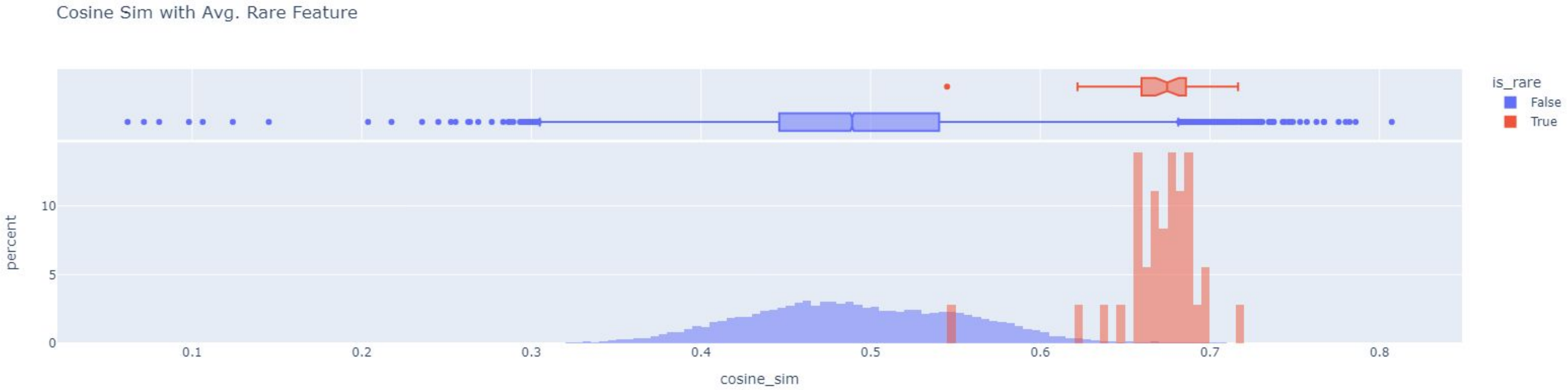
8x Dictionary



32x Dictionary



Key graphs: Rare Features are more similar than others



Extension: Limitations

- Reconstruction score optimisation
  - I trained my autoencoders to obtain the maximum reconstruction score
  - I did not validate that this score is a good proxy for good interpretation potential
  - I did not test different hyperparameters' effects on other outcomes

- The underlying data and models play a key role in this analysis which may impact the results

Works Cited

1. Trenton Bricken*, Adly Templeton*, Joshua Batson*, Brian Chen*, Adam Jermyn*, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, Chris Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. 2023. (Anthropic)
2. S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, W. Saunders. Language models can explain neurons in language models. 2023. (OpenAI)
3. https://github.com/neelnanda-io/1L-Sparse-Autoencoder
4. https://github.com/neelnanda-io/TransformerLens

Questions?