Finding Neurons in a Haystack: Case Studies with Sparse Probing

Mechanistic Interpretability Main Seminar Presentation

Raziye Sari

31.10.2024

University of Heidelberg
Institute for Computational Linguistics

Table of Contents

Motivation, Task and Approach

Sparse Probing

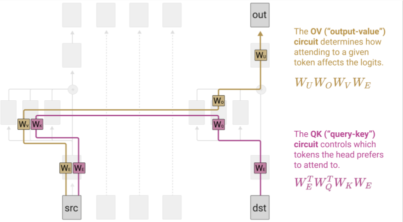
Experiments

Results

Discussion & Conclusion

Motivation, Task and Approach

Motivation



Feed forward layers function as key-value memories. [1]

- Multi head attention layers compute attention scores between tokens
- Multi layer perceptron (MLP) **responds** to input features (QK-circuit) by **updating** output vocabulary distribution (OV-circuit).

[1] Elhage et. al, 2021

Motivation

- Residual stream after MLP layers:
  $h_t^l = h_t^{l-1} + W_{proj}^l \sigma(W_{fc}^l \gamma(h_t^{l-1}) + b_{fc}^{(l)}) + b_{proj}^l$, where $\sigma = GeLU$
- Model parametrized by dense matrix multiplications and non-linearities
- n Features as linear directions in **activation space**, where $d < n$
  - Features in *superposition*
- $\rightarrow$ Train linear classifier (*probe*) on **internal activations** to predict feature

1

Probing

- Localization technique for testing feature representation
- Constrain model to use at most k neurons in predicting feature
  - Vary k to obtain information on sparsity of feature representation.
- $\rightarrow$ Limits model to **explicit** feature representation

Sparse Probing

Sparse Probing

- Transformer-based generative-pre-trained (GPT) language model $M : X \to Y, x = [x_1, \dots x_t]$
- Tokenized text dataset $X \in V^{nxT}$
- Labeled dataset $D_{probe} = \{x_{jt}, z_{jt}\}$, e.g. tense of every verb
- Binary classifier $g_l(a^l_{jt}) = \hat{z}_{jt}$, such that $L(z_{jt}, \hat{z}_{jt})$

Sparse Feature Selection Methods

Train Logistic regression probe for Optimal sparse probing (small k), else Adaptive thresholding:

1. Choose top neurons with max mean difference
2. Train series of probes with decreasing k:
3. Iteratively choose top k_t neurons with highest coefficient magnitude from k_{t-1}

Experiments

Probing in Praxis

- Challenge in conceptual separation of `isPolitician` vs. `isPolitical`, `isPerson`
- PR=$TP/(TP+FP)$, RE=$TP/(TP+FN)$, F1=$2PRxRE/(PR+RE)$
  - High PR: Either feature highly polysemantic OR model represents a more general feature
  - High RE: vice versa
- $\rightarrow$ Which features are most likely associated with the positive class ?

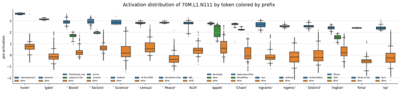
Empirical Overview

Models 7 GPT's from EleutherAI's Pyhia suite trained on 800gb dataset of diverse text

Data Ten different feature collections, including natural language, programming language and dependency & other morphological features (POS, tenses, compound words) & factual features

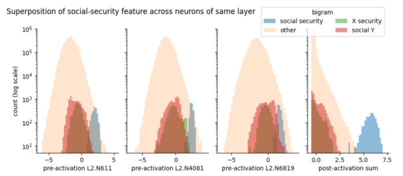
Results

Polysemanticity



(a) Activations of a single polysemantic neuron on different tokens when preceded by specific stimuli.

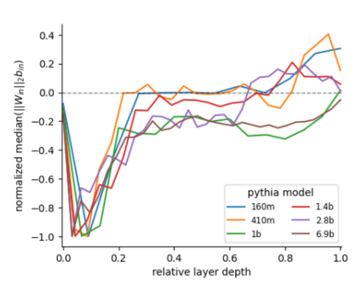
Polysemantic neuron activates on different tokens



Total activation magnitude

1. `social security` vs. `security`
2. Activations for 21 compound words were **perfectly discriminating**
3. Activation interference?

De-tokenization



Superposition in early layers

- Early layers "de-tokenize" tokens into n-grams $|V|^n$ by assigning **large input weights** and **negative biases**
    - High sensitivity towards input
    - Neuron activates very selectively

Monosemanticity

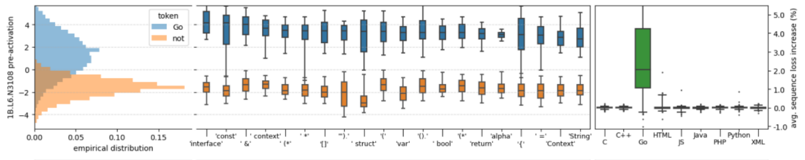


Figure 1: Single neuron activations

- Mean aggregate of activations across long sequences
- Ablation causes 6% average loss increase (70M parameter model)

Quantization Model of Scaling

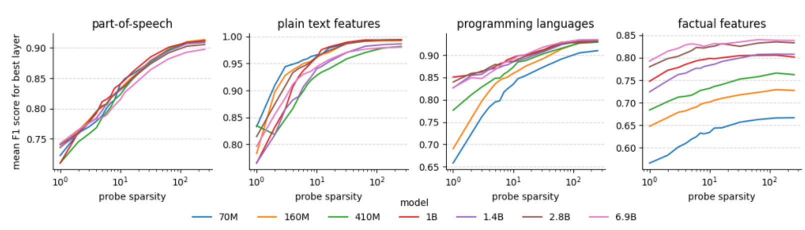


Figure 2: Caption

- Natural ordering of (rare) features learned
- Factual features learned sufficiently at lower sparsity

Feature Splitting

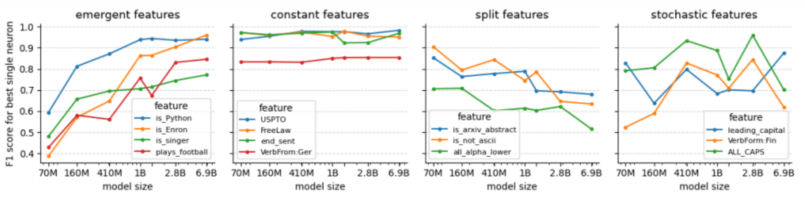


Figure 3: Caption

- Increasing model size enables more monosemanticity `allCaps` becomes `allCapsShouting, allCapsAbbreviation, ...`

Feature unions

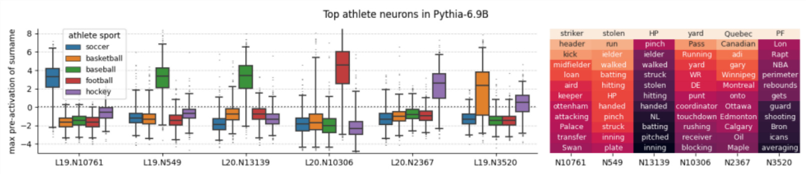


Figure 4: Coarse features represented as fine-grained features

- Feature with Low 1-sparse, but high 3-sparse may point to feature unions

Interpretability illusions

- Interpreting features for maximum activating dataset examples
    - May miss scope of representation

Judging outputs

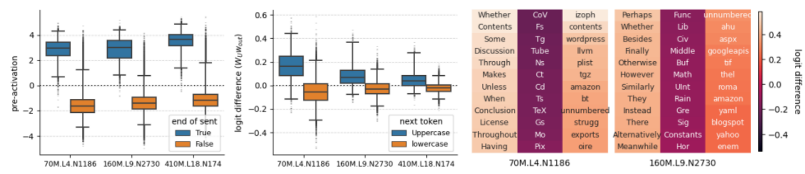


Figure 5: EOS-neuron activations

- Attaining logits by product of M^U and neuron output weight

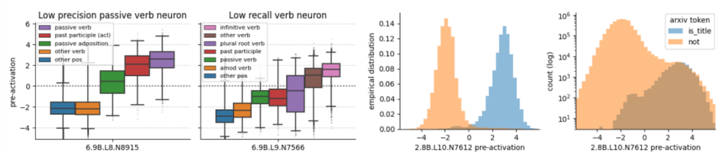


Figure 6: Caption

- Feature definition scope different for model
    - Low-recall-high-precision `isVerb`
    - Low-precision-high-recall `isPassiveVerb`
- Undefined, rare features drowned out by pre-defined features.

Discussion & Conclusion

Limitations

- Limited insights into causation
- Sensitive to implementation details
- Features in superposition vs. union of multiple independent features
- Increasing model scale harmful to transferability of feature dataset