Sparse autoencoders for mechanistic interpretability

Flora Chen
MSCS2201 - Mini Research Project

TLDR

- Motivation
  - It’s cool to shine a light into the black-box that is transformers
  - Mechanistic interpretability is important for deterministic measures of improving AI safety.
- Main idea:
  - Mid-layer neurons are polysemantic. This means single or small clusters of neurons can encode for many unrelated features.
  - This poses a problem for interpretability and to solve this Anthropic has proposed the use of Sparse AutoEncoders (SAE)
  - SAE have massive latent spaces and the sparsity encourages neurons to decompose activations into monosemantic features.

Results

Sentence topics clearly converged for single neuron after SAE. Before SAE sentences were varied

```
=== BEFORE (polysemantic neuron) ===
- Once upon a time there was a mommy, a daddy and a baby. Every day, Mommy and Daddy would take Baby t
- Once upon a time, there was a little girl named Lily. She loved to play with her toys and eat snacks
- Tilly had the best day! She woke up in the morning and put on her new pants. They were so big that t
- Once upon a time, there were two friends, Jack and Jane. They were playing together in the park with
- Mum and Dad were packing for a trip.

Mum said to Dad, "Let's get the egg in the suitcase."

Dad agr

=== AFTER (SAE monosemantic feature) ===
- Once upon a time there was a little girl named Jane. She left the house early one morning to go visi
- Once there was a little girl who wanted to watch the television. She approached the large television
- Once upon a time, there was a foolish little mouse. He was so foolish that he thought he could bite
- One day, there was a little girl called Sam who loved to play games. She particularly liked playing
- Once upon a time, there was a big pink elephant who had a bright yellow spot on his back. He liked t
```

Literature review

Main paper
Towards Monosemanticity: Decomposing Language Models With Sparse Autoencoders — Anthropic (Bakker et al., 2023)

- Introduces SAEs as a method to extract human-interpretable, low-overlap features.

Obviously important paper
Attention Is All You Need — Vaswani et al., 2017

- *The* paper on modern transformers

Related Techniques

- Feature attribution in transformers: Integrated Gradients, Activation Patching
- Concept activation: TCAV
- Visualization of internal representations: Projection heads, PCA/UMAP of activations

Approach used

Model: GPT-2 small (124M params), chosen for:

- strong polysemanticity
- manageable activation sizes
- widely studied architecture

Dataset: TinyStories (synthetic, small vocabulary, interpretable semantics)

- Used only to collect activations (no finetuning)

Target Layer: GPT-2 Layer 6 MLP post-GELU

- empirically known to exhibit polysemantic neurons
- mid-layers encode mixed semantic and syntactic structure

Summary

- Transformer neurons are **not interpretable units** due to superposition.
- SAEs recover **monosemantic sparse features** aligned with human concepts.
- This pipeline can scale to more layers, models, and more rigorous feature evaluation.