Faculty of Mathematics and Information Sciences

WARSAW UNIVERSITY OF TECHNOLOGY

Sparse Autoencoders Do Not Find Canonical Units of Analysis

ICLR 2025 Poster

About the Paper

- On ArXiv: 07.02.2025
- ICLR 2025 Poster: 22.01.2025
- Reviewers scores: two 8 and two 6
- Very long conversations with both 6 reviewers
- 8 Authors

Streamlit

OpenReview

About the Authors - Neel Nanda

- Runs the Google DeepMind mechanistic interpretability team
- 80% of SAE papers are his (just kidding)

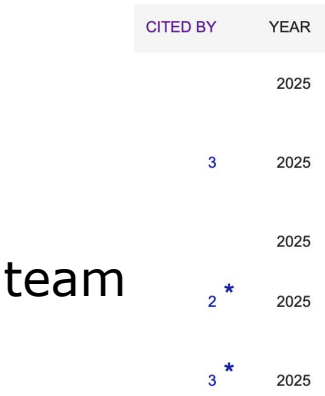
Neel Nanda

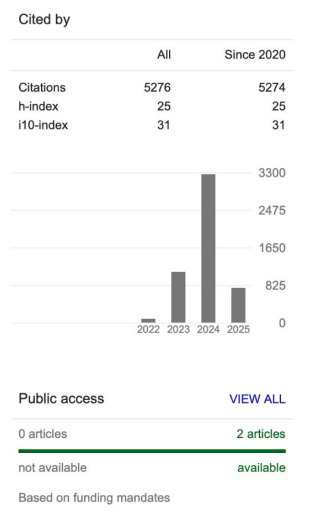
FOLLOWING

Mechanistic Interpretability Team Lead, Google DeepMind
Verified email at deepmind.com - Homepage

AI ML AI Alignment Interpretability Mechanistic Interpretability

TITLE	CITED BY	YEAR
Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback Y Bai, A Jones, K Ndousse, A Askell, A Chen, N DasSarma, D Drain, ... arXiv preprint arXiv:2204.05862	1866	2022
In-context Learning and Induction Heads C Olsson, N Elhage, N Nanda, N Joseph, N DasSarma, T Henighan, ... Transformer Circuits Thread	684 *	2022
A Mathematical Framework for Transformer Circuits N Elhage, N Nanda, C Olsson, T Henighan, N Joseph, B Mann, A Askell, ... Transformer Circuits Thread	679 *	2021
Progress Measures For Grokking Via Mechanistic Interpretability N Nanda, L Chan, T Liberum, J Smith, J Steinhardt ICLR 2023 Spotlight	410 *	2023
Predictability and surprise in large generative models D Ganguli, D Hernandez, L Lovitt, A Askell, Y Bai, A Chen, T Conerly, ... Proceedings of the 2022 ACM Conference on Fairness, Accountability, and ...	189	2022





The three of them just do SAE

Bart Bussmann

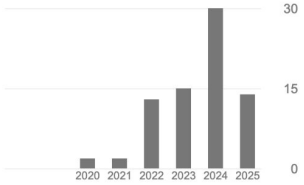
Independent
Verified email at student.uva.nl
Artificial Intelligence

FOLLOWING

TITLE	CITED BY	YEAR
Sparse Autoencoders Do Not Find Canonical Units of Analysis P Leask, B Bussmann, M Pearce, J Bloom, C Tigges, NA Moubayed, ... arXiv preprint arXiv:2502.04878		2025
Inferring the relationship between soil temperature and the normalized difference vegetation index with machine learning S Mortier, A Hamedpour, B Bussmann, RPT Wandji, S Latré, ... Ecological Informatics 82, 102730	7	2024
Showing sae latents are not atomic using meta-saes B Bussmann, M Pearce, P Leask, J Bloom, L Sharkey, N Nanda https://www.alignmentforum.org/posts/TMAmHh4DdMr4nCSr5/showing-sae-latents ...		2024

Cited by

	All	Since 2020
Citations	76	76
h-index	5	5
i10-index	2	2



Public access VIEW ALL

0 articles — 4 articles
not available — available

Based on funding mandates

Co-authors

Neel Nanda
Mechanistic Interpretability Team...

Jannes Nys
ETH Zürich

Patrick Leask

Durham University
Verified email at durham.ac.uk
Artificial Intelligence

FOLLOWING

TITLE	CITED BY	YEAR
BatchTopK Sparse Autoencoders B Bussmann, P Leask, N Nanda NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning	8 *	2024
Stitching sparse autoencoders of different sizes P Leask, B Bussmann, JI Bloom, C Tigges, N Al Moubayed, N Nanda NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning	1	2024
Sparse Autoencoders Do Not Find Canonical Units of Analysis P Leask, B Bussmann, M Pearce, J Bloom, C Tigges, NA Moubayed, ... ICLR 2025		2025

Learning

8 * 2024

3 * 2024

zotron images—generalisation dvanced deep-learning models — 7 — 2023

scovery in time series data — 37 * — 2020

...

TL;DR - SAEs May Not Provide Canonical Features in LLMs

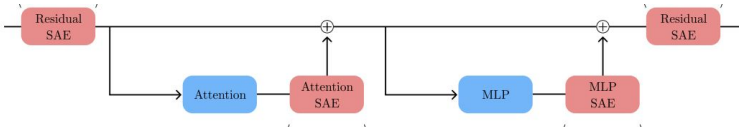
Mechanistic interpretability aims to decompose neural network activations into interpretable features. Authors **challenge** the belief that SAEs **find canonical units** in LLMs:

- SAE Stitching: Reveals incompleteness by showing larger SAEs contain novel latents that improve smaller SAEs' performance.

- Meta-SAEs: Demonstrates non-atomicity by decomposing SAE latents into combinations of smaller SAE latents (Use SAE on SAE).

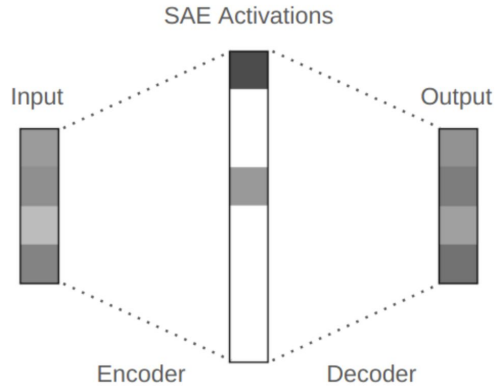
Recap on Sparse Autoencoders (SAEs)

- SAE is a technique for transforming a dense representation (with superposition) into a sparse monosemantic representation



- SAE decompose the activations of a model into more interpretable pieces.

- Can be applied to any activation representation
- It is simply an autonencoder with some proxy enforcing sparsity:
  - TopK
  - L1 regularization
  - JumpReLU



As a preprocessing step we apply a scalar normalization to the model activations so their average squared L2 norm is the residual stream dimension, D. We denote the normalized activations as $\mathbf{x} \in \mathbb{R}^D$, and attempt to decompose this vector using F features as follows:

$$\hat{\mathbf{x}} = \mathbf{b}^{dec} + \sum_{i=1}^{F} f_i(\mathbf{x}) \mathbf{W}^{dec}_{\cdot,i}$$

where $W^{dec} \in \mathbb{R}^{D \times F}$ are the learned SAE decoder weights, $\mathbf{b}^{dec} \in \mathbb{R}^D$ are learned biases, and f_i denotes the activity of feature *i*. Feature activations are given by the output of the encoder:

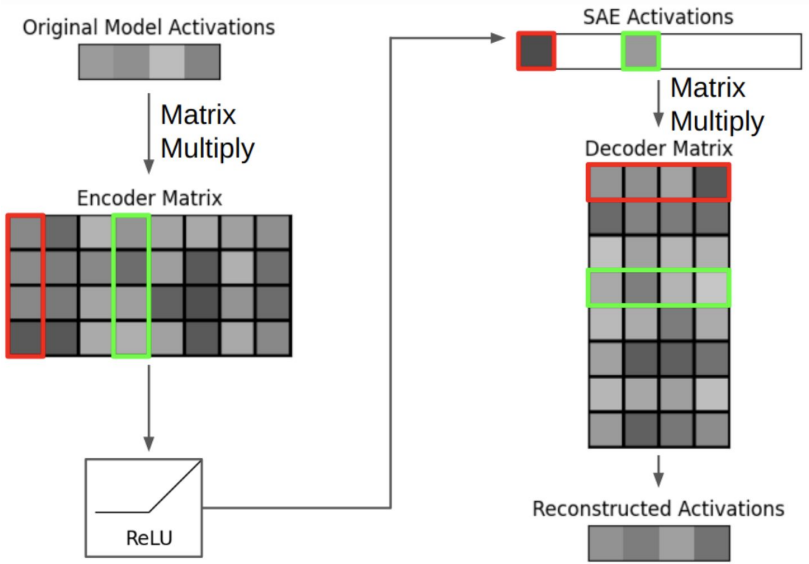
$$f_i(x) = \text{ReLU}\left(\mathbf{W}^{enc}_{i,\cdot} \cdot \mathbf{x} + b^{enc}_i\right)$$

where $W^{enc} \in \mathbb{R}^{F \times D}$ are the learned SAE encoder weights, and $\mathbf{b}^{enc} \in \mathbb{R}^F$ are learned biases.

The loss function $\mathcal{L}$ is the combination of an L2 penalty on the reconstruction loss and an L1 penalty on feature activations.

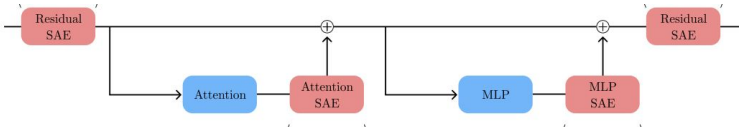
$$\mathcal{L} = \mathbb{E}_{\mathbf{x}}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda \sum_i f_i(\mathbf{x}) \cdot \|\mathbf{W}^{dec}_{\cdot,i}\|_2\right]$$

Including the factor of $\|\mathbf{W}^{dec}_{\cdot,i}\|_2$ in the L1 penalty term allows us to interpret the unit-normalized decoder vectors $\frac{\mathbf{W}^{dec}_{\cdot,i}}{\|\mathbf{W}^{dec}_{\cdot,i}\|_2}$ as "feature vectors" or "feature directions," and the product $f_i(\mathbf{x}) \cdot \|\mathbf{W}^{dec}_{\cdot,i}\|_2$ as the feature activations[2] . Henceforth we will use "feature activation" to refer to this quantity.



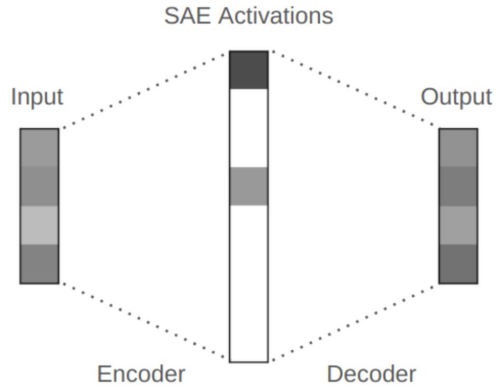
Recap on Sparse Autoencoders (SAEs)

- SAE is a technique for transforming a dense representation (with superposition) into a sparse monosemantic representation



- SAE decompose the activations of a model into more interpretable pieces.

- Can be applied to any activation representation
- It is simply an autonencoder with some proxy enforcing sparsity:
  - TopK
  - L1 regularization
  - JumpReLU



How to interpret SAE features

Just use LLM!

For each feature draw:

- 5 highest activating examples
- 20 random, non-activating examples

The model based on the k highest and n random samples suggest the description

Autointerpretable score is calculated by giving example and description and asking model to predict the magnitude of the feature.

Finally the correlation is calculated between ground truth and LLM prediction

The Shady Part

- Authors introduce BatchTopK in paper published on arXiv: **07.02.2025**

- The same authors published paper BatchTopK on arXiv: **09.12.2024**; and on ICLR: **22.01.2025**

In summary, our contributions are:

1. **SAE stitching**, as a method for comparing latents across different sizes of SAE. Latents in a larger SAE are either novel latents, missing in smaller SAEs, or reconstruction latents, similar to some latents in smaller SAEs.
2. **Meta-SAEs**, as an approach for decomposing the decoder directions of SAEs into interpretable, monosemantic meta-latents.
3. We also introduce **BatchTopK SAEs**, the state-of-the-art SAE architecture for sparse dictionary learning on language model activations at a fixed average sparsity.

arXiv > cs > arXiv:2412.06410

Computer Science > Machine Learning

[Submitted on 9 Dec 2024]

BatchTopK Sparse Autoencoders

Bart Bussmann, Patrick Leask, Neel Nanda

Stitching Sparse Autoencoders of Different Sizes

Patrick Leask, Bart Bussmann, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Neel Nanda

Published: 10 Oct 2024, Last Modified: 09 Nov 2024 SciForDL Poster Everyone Revisions BibTeX CC BY 4.0

SAE Stitching

$$\hat{\mathbf{x}} := \alpha \mathbf{b}_0^{\text{dec}} + (1-\alpha)\mathbf{b}_1^{\text{dec}} + \sum_{l_0 \in L_0} \mathbf{W}_{0,l_0}^{\text{dec}} f_{0,l_0}(\mathbf{x}) + \sum_{l_1 \in L_1} \mathbf{W}_{1,l_1}^{\text{dec}} f_{1,l_1}(\mathbf{x}) \quad (5)$$

where $\alpha = \frac{|L_0|}{|L_0|+|L_1|}$, $\mathbf{W}_{0,l_0}^{\text{dec}}$ and $\mathbf{W}_{1,l_1}^{\text{dec}}$ represent individual decoder directions and L_0 and L_1 are the set of latents we include from the respective SAEs. Unlike with model stitching (Bansal et al., 2021), SAE decoder directions are privileged and require no transformations to stitch.

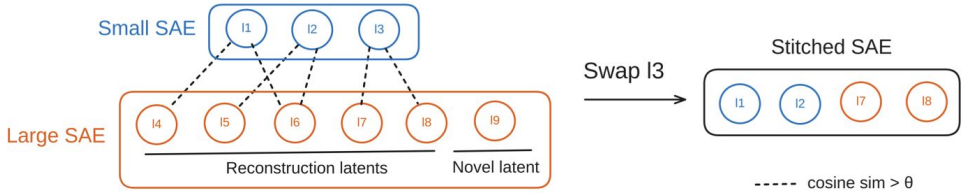
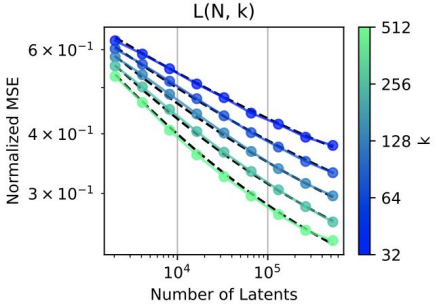


Figure 3: SAE stitching operation: connected subgraphs of latents can be swapped between SAEs based on cosine similarity

two categories, i.e. for an SAE decoder direction $\mathbf{W}_{1,i}^{\text{dec}}$, the feature is assigned to the novel group if

$$\max_j \left(\frac{\mathbf{W}_{1,i}^{\text{dec}} \cdot \mathbf{W}_{0,j}^{\text{dec}}}{\|\mathbf{W}_{1,i}^{\text{dec}}\| \|\mathbf{W}_{0,j}^{\text{dec}}\|} \right) < \theta \quad (6)$$

where $\mathbf{W}_{0,j}^{\text{dec}}$ represents any decoder direction in $\mathbf{W}_0^{\text{dec}}$. If the maximum cosine similarity is larger than the threshold, it is assigned to the reconstruction group. Figure 4 shows the relationship between



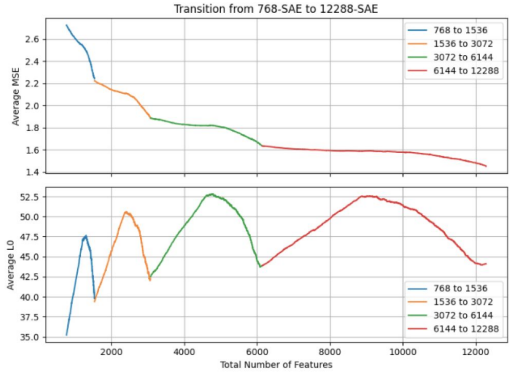


Figure 5: Interpolating between SAE pairs of increasing dictionary size (768→1536→3072→6144→12288) through two steps per phase: adding novel latents (increasing L0) then swapping groups of reconstruction latents (decreasing L0 on average). Both steps on average improve reconstruction (MSE). The L0 and MSE are averages over input samples.

SAE Stitching

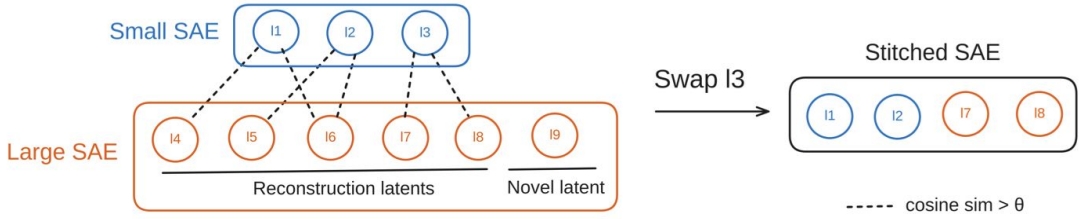


Figure 3: SAE stitching operation: connected subgraphs of latents can be swapped between SAEs based on cosine similarity

- "The existence of the novel group of latents demonstrates that smaller SAEs are incomplete, and that larger SAE learn features that are missed by the smaller SAE."

Figure 5: Interpolating between SAE pairs of increasing dictionary size (768→1536→3072→6144→12288) through two steps per phase: adding novel latents (increasing L0) then swapping groups of reconstruction latents (decreasing L0 on average). Both steps on average improve reconstruction (MSE). The L0 and MSE are averages over input samples.

Meta-SAE

To decompose the latents of larger SAEs, we introduce meta-SAEs. Meta-SAEs are SAEs trained to reconstruct the decoder directions $\mathbf{W}_i^{\text{dec}}$ of a standard SAE using a dictionary of meta-latents, rather

SAE Latent Description	Meta-Latent Descriptions
Albert Einstein	Science & Scientists, Famous People, Space & Astronomy, Germany, Electricity, Words starting with a capital E
Rugby	Sports activities, Words starting with 'R', References to Ireland, References to sports leagues, activities & actions
Android Operating System	Mobile phones, operating systems, Californian cities

"some reconstruction group latents have high decoder cosine similarity to multiple latents in the smaller SAE, suggesting the large SAE latent is an interpolation or composition of the smaller SAE latents"

"The latents of meta-SAEs have similar decoder directions to those found in SAEs of comparable size trained directly on the same network activations. This observation further supports the hypothesis that larger SAE latents are not entirely new features but may be compositions of features already learned, albeit less precisely, by smaller models."

A.8 METASAE ADDITIONAL FIGURES

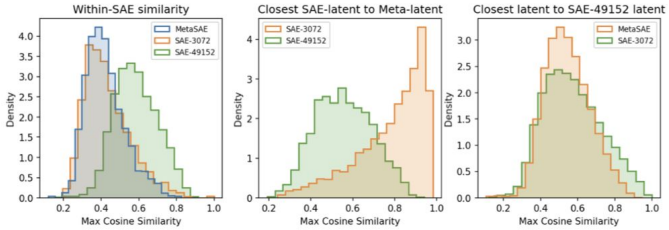


Figure 22: Cosine similarity between SAE latents and meta-SAE latents. Note the high maximum cosine similarity between latents from a meta-SAE with 2304 latents, and a standard SAE with 3072 latents.

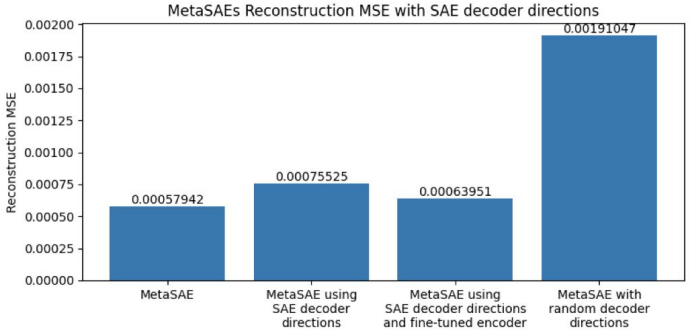


Figure 23: Change in reconstruction performance of a meta-SAE when its decoder directions are replaced with the most similar decoder direction from an SAE with a similar dictionary size.

Meta-SAE

To decompose the latents of larger SAEs, we introduce meta-SAEs. Meta-SAEs are SAEs trained to reconstruct the decoder directions $\mathbf{W}_i^{\text{dec}}$ of a standard SAE using a dictionary of meta-latents, rather

SAE Latent Description	Meta-Latent Descriptions
Albert Einstein	Science & Scientists, Famous People, Space & Astronomy, Germany, Electricity, Words starting with a capital E
Rugby	Sports activities, Words starting with 'R', References to Ireland, References to sports leagues, activities & actions
Android Operating System	Mobile phones, operating systems, Californian cities

"some reconstruction group latents have high decoder cosine similarity to multiple latents in the smaller SAE, suggesting the large SAE latent is an interpolation or composition of the smaller SAE latents"

"The latents of meta-SAEs have similar decoder directions to those found in SAEs of comparable size trained directly on the same network activations. This observation further supports the hypothesis that larger SAE latents are not entirely new features but may be compositions of features already learned, albeit less precisely, by smaller models."

BatchTopK

The training objective for BatchTopK SAEs is defined as:

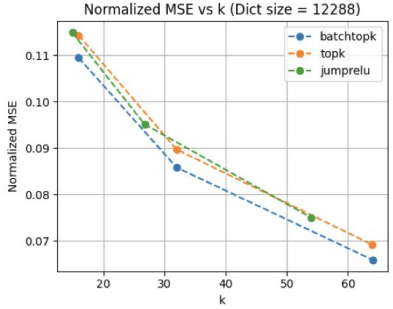
$$\mathcal{L}(\mathbf{X}) = \|\mathbf{X} - \text{BatchTopK}(\mathbf{W}^{\text{enc}}\mathbf{X} + \mathbf{b}^{\text{enc}})\mathbf{W}^{\text{dec}} + \mathbf{b}^{\text{dec}})\|_2^2 + \alpha\mathcal{L}_{\text{aux}} \tag{7}$$

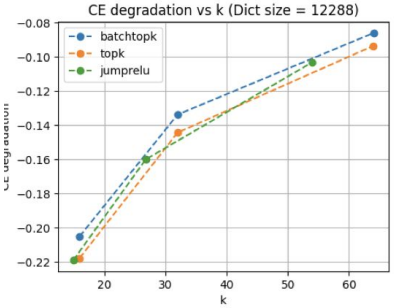
Here, $\mathbf{X}$ is the input data batch; $\mathbf{W}^{\text{enc}}$ and $\mathbf{b}^{\text{enc}}$ are the encoder weights and biases, respectively; $\mathbf{W}^{\text{dec}}$ and $\mathbf{b}^{\text{dec}}$ are the decoder weights and biases. The BatchTopK function sets all activation values to zero that are not among the top $n \times k$ activations by value in the batch, not changing the other values. The term $\mathcal{L}_{\text{aux}}$ is an auxiliary loss scaled by the coefficient α, used to prevent dead latents, and is the same as in TopK SAEs.

BatchTopK introduces a dependency between the activations for the samples in a batch. We alleviate this during inference by using a threshold θ that is estimated as the average of the minimum positive activation values across a number of batches:

$$\theta = \mathbb{E}_{\mathbf{X}}[\min\{z_{i,j}(\mathbf{X}) | z_{i,j}(\mathbf{X}) > 0\}] \tag{8}$$

where $z_{i,j}(\mathbf{X})$ is the jth latent activation of the ith sample in a batch $\mathbf{X}$. With this threshold, we use the JumpReLU activation function during inference instead of the BatchTopK activation function, zeroing out all activations under the threshold θ.





Why is this better? Because it lifts the constraint of **always using k** SAE latents

Discussion Time

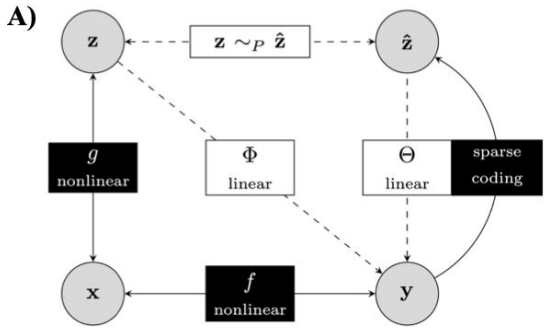
- is monosemanticity achievable at all?

2012). Assuming the setting described above, we know that $f \circ g : \mathbb{R}^N \to \mathbb{R}^M$ is a linear function. Thus, there exists a matrix $\Phi \in \mathbb{R}^{M \times N}$ (see Fig. 3A) such that

$$f \circ g(z) = \Phi z \quad \forall z \in \mathbb{R}^N. \tag{6}$$

Moreover, assuming the distribution over latent variables $z \sim P(Z)$ is sparse (at most K active components) (Donoho, 2006), we can recover the latent variables with high probability if

$$M > \mathcal{O}\left(K \log\left(\frac{N}{K}\right)\right). \tag{7}$$



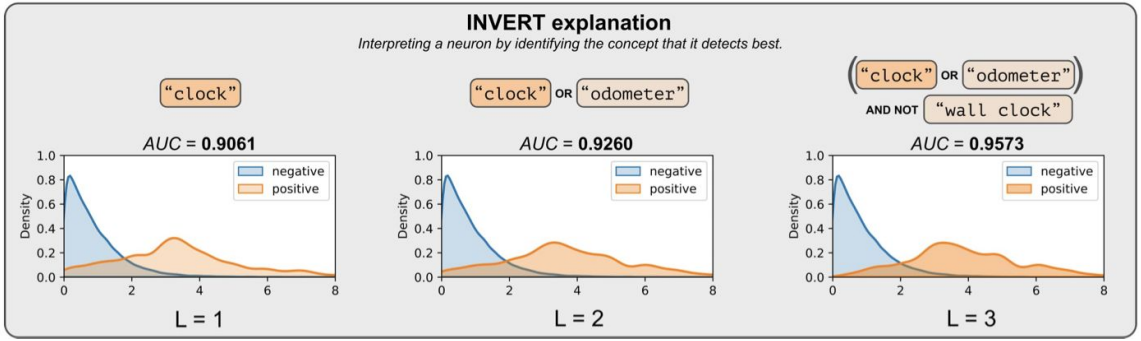
[Submitted on 3 Mar 2025]

From superposition to sparse codes: interpretable representations in neural networks

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, Nina Miolane

Discussion Time

- do we need monosemanticity to be able to explain NNs?



Discussion Time

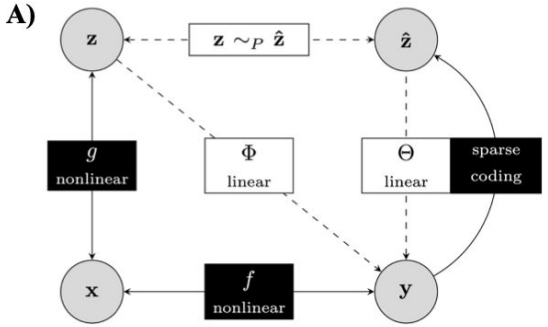
- is monosemanticity achievable at all?

2012). Assuming the setting described above, we know that $f \circ g : \mathbb{R}^N \to \mathbb{R}^M$ is a linear function. Thus, there exists a matrix $\Phi \in \mathbb{R}^{M \times N}$ (see Fig. 3A) such that

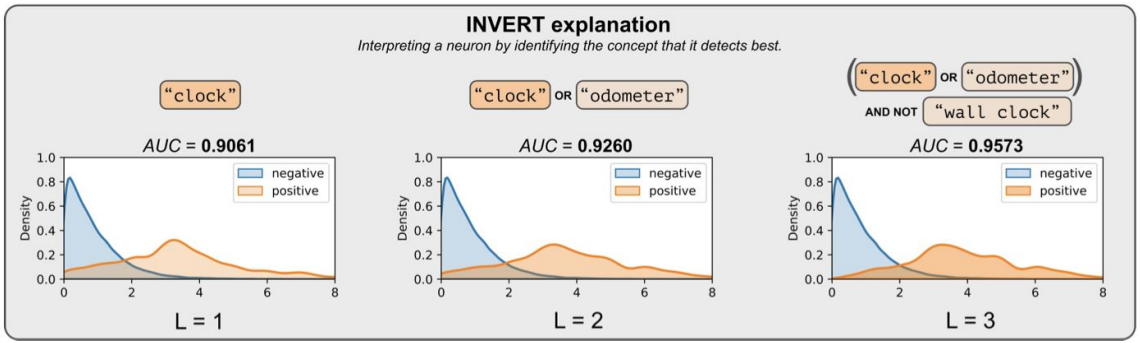
$$f \circ g(z) = \Phi z \quad \forall z \in \mathbb{R}^N. \tag{6}$$

Moreover, assuming the distribution over latent variables $z \sim P(Z)$ is sparse (at most K active components) (Donoho, 2006), we can recover the latent variables with high probability if

$$M > \mathcal{O}\left(K \log\left(\frac{N}{K}\right)\right). \tag{7}$$



- do we need monosemanticity to be able to explain NNs?



Faculty of Mathematics and Information Sciences

WARSAW UNIVERSITY OF TECHNOLOGY

Sparse Autoencoders Do Not Find Canonical Units of Analysis

ICLR 2025 Poster