Mechanistic Interpretability

Part 2: Polysemanticity and Sparse Autoencoders

Sondre Wold, Shenbin Qian

University of Oslo



In this lecture:

1. Polysemanticity
2. Sparse Autoencoders

Too many features!

The chestnut-naped antpitta inhabits temperate to humid montane forest, where it almost entirely occurs in and near dense bamboo [MASK].

Too many features!

The chestnut-naped antpitta inhabits temperate to humid montane forest, where it almost entirely occurs in and near dense bamboo [MASK].

- Even in large models, we have a very limited number of neurons to encode features of the input with.

Too many features!

The chestnut-naped antpitta inhabits temperate to humid montane forest, where it almost entirely occurs in and near dense bamboo [MASK].

- Even in large models, we have a very limited number of neurons to encode features of the input with.
- If fitting a LLM on a corpora requires encoding α features of the corpora, and the model has φ neurons (hidden dimension), then surely $\varphi < \alpha$.

Too many features!

The chestnut-naped antpitta inhabits temperate to humid montane forest, where it almost entirely occurs in and near dense bamboo [MASK].

- Even in large models, we have a very limited number of neurons to encode features of the input with.
- If fitting a LLM on a corpora requires encoding α features of the corpora, and the model has φ neurons (hidden dimension), then surely $\varphi < \alpha$.
- Llama 3.3 70B has a hidden size of $\varphi = 8192$, and the training data has 15 trillion tokens; this makes for an enormous α!

Polysemanticity

This hints towards a *polysemantic* encoding of features: models represent more features than they have dimensions!

Polysemanticity

This hints towards a *polysemantic* encoding of features: models represent more features than they have dimensions!

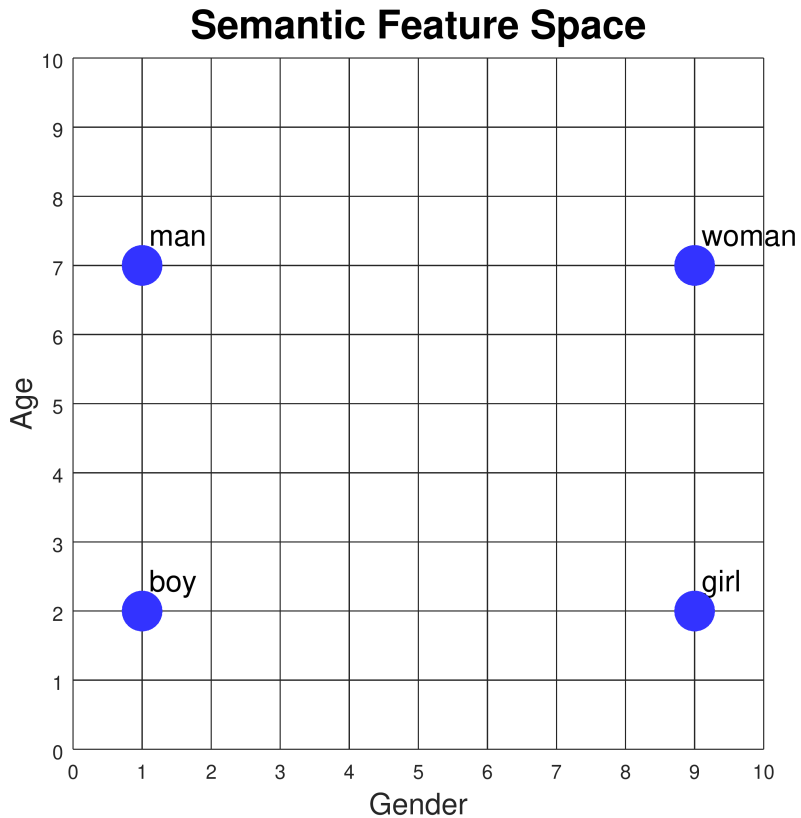


Figure: Illustration by Dave Touretzky

Superposition

The idea that models can encode many more features than they have dimensions is called the **superposition hypothesis** (Elhage et al., 2022), and it makes interpretability **hard**.

Superposition

Let's say we have a feature vector that represents five aspects of an animal, like whether or not it has fur, whether it can fly, etc, and we want to predict the type of animal we are dealing with.

Superposition

Let's say we have a feature vector that represents five aspects of an animal, like whether or not it has fur, whether it can fly, etc, and we want to predict the type of animal we are dealing with.

We want to learn a 2-dim embedding of this feature vector, so a 5×2 dimensional weight matrix.

Superposition

Let's say we have a feature vector that represents five aspects of an animal, like whether or not it has fur, whether it can fly, etc, and we want to predict the type of animal we are dealing with.

We want to learn a 2-dim embedding of this feature vector, so a 5×2 dimensional weight matrix.

How can the network compress this?

Superposition

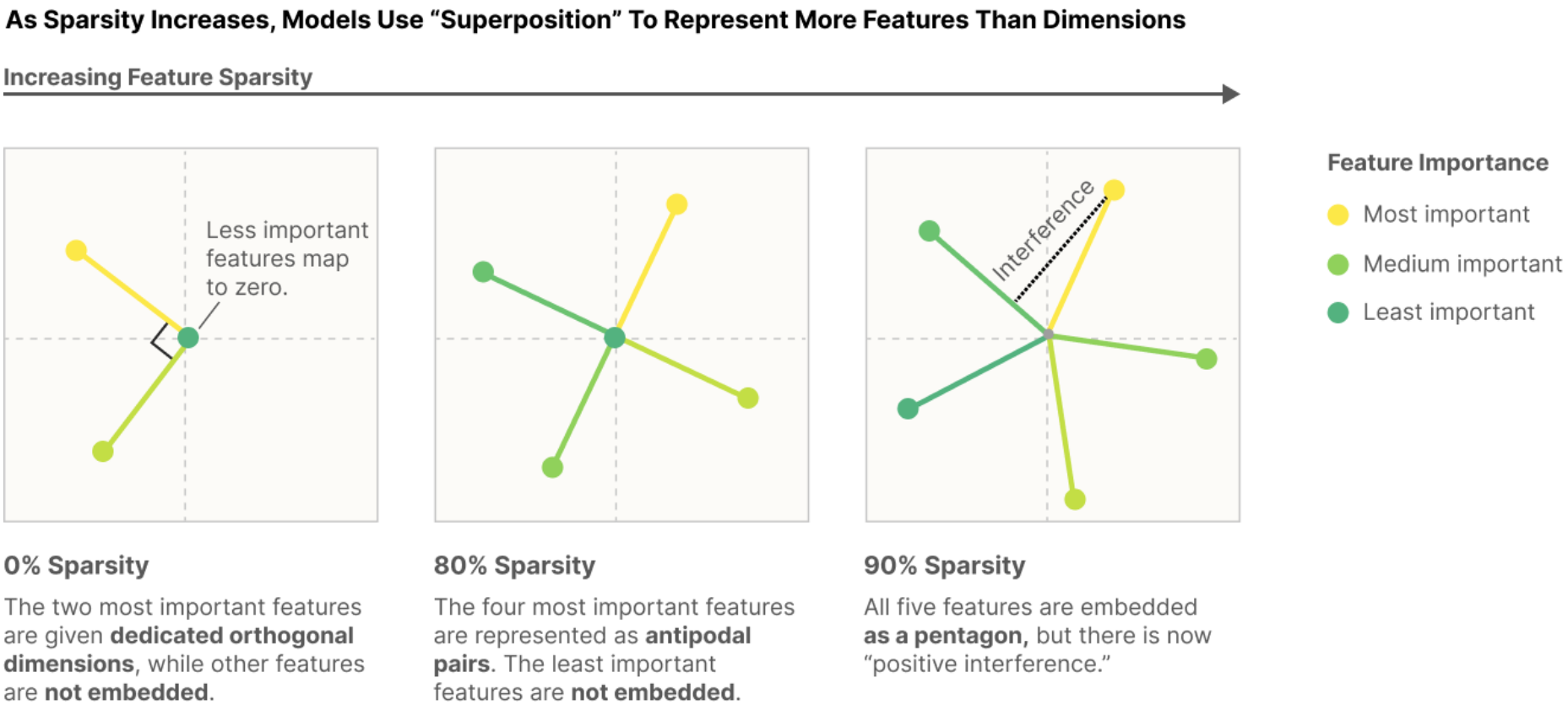


Figure: Illustration from Elhage et al. (2022)

Superposition

This means that:

- Most neurons respond to seemingly unrelated features of the input.

Superposition

This means that:

- Most neurons respond to seemingly unrelated features of the input.
- Because the NNs can't represent all features as orthogonal vectors, features are instead represented as linear combinations of other features.

Superposition

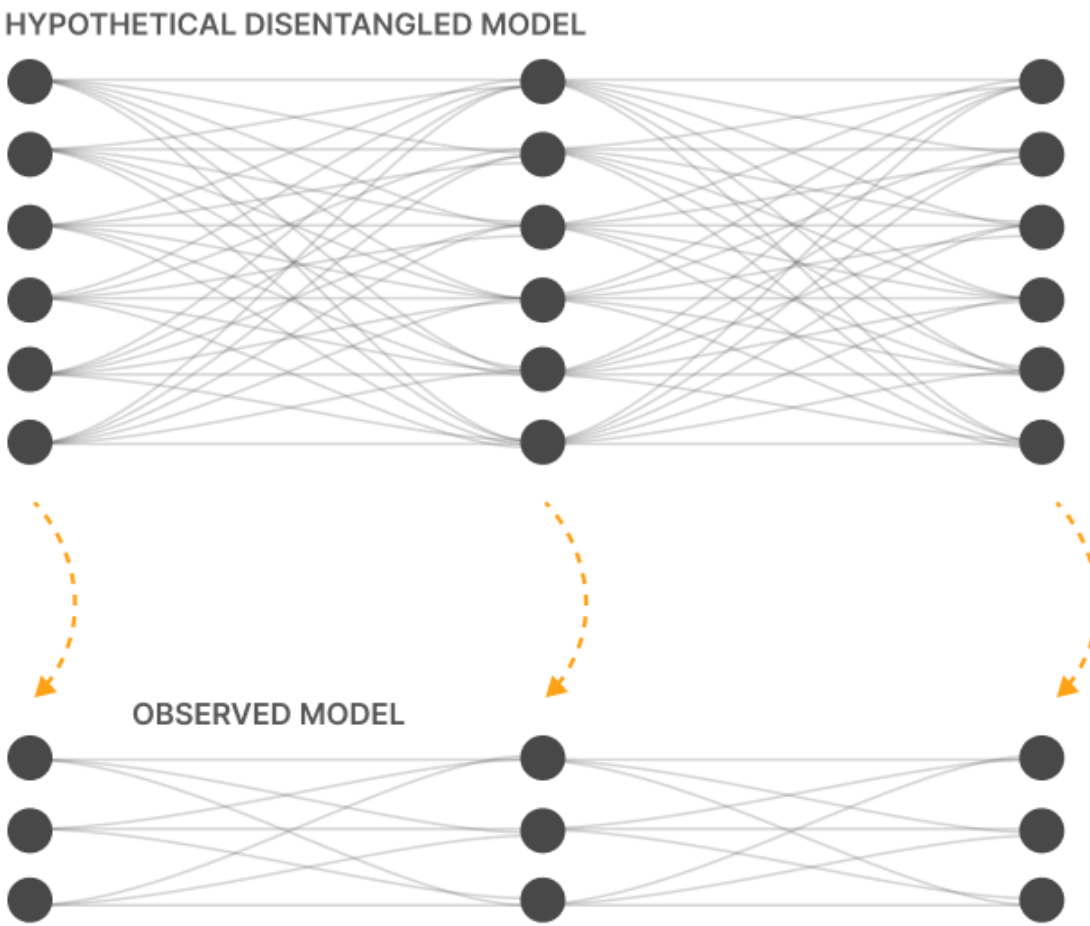
This means that:

- Most neurons respond to seemingly unrelated features of the input.
- Because the NNs can't represent all features as orthogonal vectors, features are instead represented as linear combinations of other features.
- This is a mess.

In this lecture:

1. Polysemanticity
2. Sparse Autoencoders

Taking features out of superposition



Under the superposition hypothesis, the neural networks we observe are **simulations of larger networks** where every neuron is a disentangled feature.

These idealized neurons are **projected** on to the actual network as "almost orthogonal" vectors over the neurons.

The network we observe is a **low-dimensional projection** of the larger network. From the perspective of individual neurons, this presents as polysemanticity.

Figure: Illustration from Elhage et al. (2022)

Taking features out of superposition

- If we could reconstruct the idealized model, we would have perfect interpretability.
- Unfortunately, the field of *compressed sensing* tells us that this is impossible.
- Unless. . .



Figure: Terence Tao

Taking features out of superposition

If features are sparse, you *can* reconstruct the original signal!

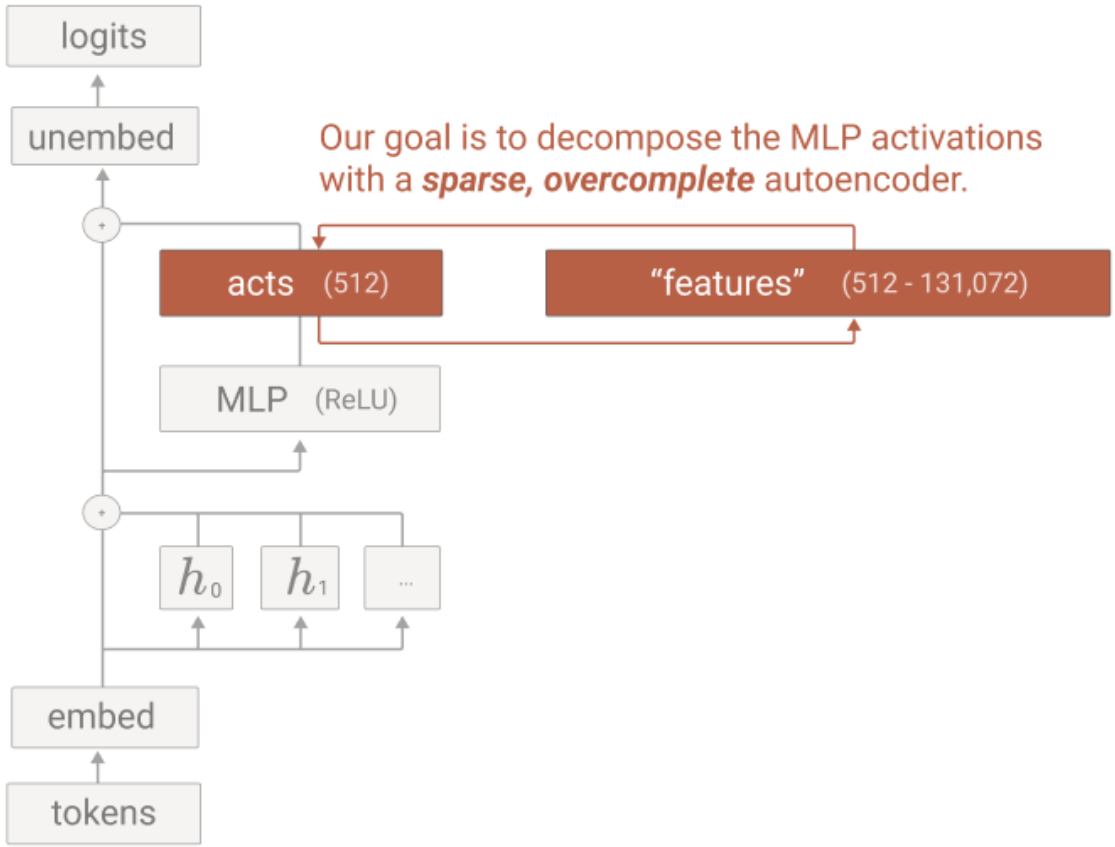


Figure: Illustration from Bricken et al. (2023)

Sparse Autoencoders

- Sparse Autoencoding is an unsupervised learning algorithm used for representation learning (Ng et al., 2011).
- Recently, it has also been used to create sparse reconstructions of activations in LLMs, enabling interpretability (Cunningham et al., 2023).

Dictionary learning

Problem:

- Suppose that each $\{\mathbf{x}_i\}_{i=1}^{n_{vec}} \subset \mathbb{R}^d$ is a sparse linear combination of unknown vectors $\{\mathbf{g}_j\}_{i=1}^{n_{gt}} \subset \mathbb{R}^d$.
  - The elements of $\mathbf{x}$ can be constructed from elements of $\mathbf{g}$: $\mathbf{x}_i = \sum_j a_{i,j}\mathbf{g}_j$ where $\mathbf{a}_i$ is a sparse vector.

Dictionary learning

Problem:

- Suppose that each $\{\mathbf{x}_i\}_{i=1}^{n_{vec}} \subset \mathbb{R}^d$ is a sparse linear combination of unknown vectors $\{\mathbf{g}_j\}_{i=1}^{n_{gt}} \subset \mathbb{R}^d$.
  - The elements of $\mathbf{x}$ can be constructed from elements of $\mathbf{g}$: $\mathbf{x}_i = \sum_j a_{i,j}\mathbf{g}_j$ where $\mathbf{a}_i$ is a sparse vector.

- $\{\mathbf{x}_i\}_{i=1}^{n_{vec}}$ is the activations of a neural network and $\{\mathbf{g}_j\}_{i=1}^{n_{gt}}$ is the unknown "ground-truth network" that we want to recover.

Dictionary learning

Problem:

- Suppose that each $\{\mathbf{x}_i\}_{i=1}^{n_{vec}} \subset \mathbb{R}^d$ is a sparse linear combination of unknown vectors $\{\mathbf{g}_j\}_{i=1}^{n_{gt}} \subset \mathbb{R}^d$.
  - The elements of $\mathbf{x}$ can be constructed from elements of $\mathbf{g}$: $\mathbf{x}_i = \sum_j a_{i,j}\mathbf{g}_j$ where $\mathbf{a}_i$ is a sparse vector.

- $\{\mathbf{x}_i\}_{i=1}^{n_{vec}}$ is the activations of a neural network and $\{\mathbf{g}_j\}_{i=1}^{n_{gt}}$ is the unknown "ground-truth network" that we want to recover.

- We want to learn a set of features $\{\mathbf{f}_k\}_{k=1}^{n_{feat}} \subset \mathbb{R}^d$ so that for any network feature $\mathbf{g}_j$ there is a feature $\mathbf{f}_j$ such that $\mathbf{g}_j \approx \mathbf{f}_k$

Dictionary learning

Problem:

- Suppose that each $\{\mathbf{x}_i\}_{i=1}^{n_{vec}} \subset \mathbb{R}^d$ is a sparse linear combination of unknown vectors $\{\mathbf{g}_j\}_{i=1}^{n_{gt}} \subset \mathbb{R}^d$.
  - The elements of $\mathbf{x}$ can be constructed from elements of $\mathbf{g}$: $\mathbf{x}_i = \sum_j a_{i,j}\mathbf{g}_j$ where $\mathbf{a}_i$ is a sparse vector.

- $\{\mathbf{x}_i\}_{i=1}^{n_{vec}}$ is the activations of a neural network and $\{\mathbf{g}_j\}_{i=1}^{n_{gt}}$ is the unknown "ground-truth network" that we want to recover.

- We want to learn a set of features $\{\mathbf{f}_k\}_{k=1}^{n_{feat}} \subset \mathbb{R}^d$ so that for any network feature $\mathbf{g}_j$ there is a feature $\mathbf{f}_j$ such that $\mathbf{g}_j \approx \mathbf{f}_k$

- $\{\mathbf{f}_k\}_{k=1}^{n_{feat}}$ is called a **dictionary**, which is our approximation of the idealized network.

Learning a dictionary with an SAE

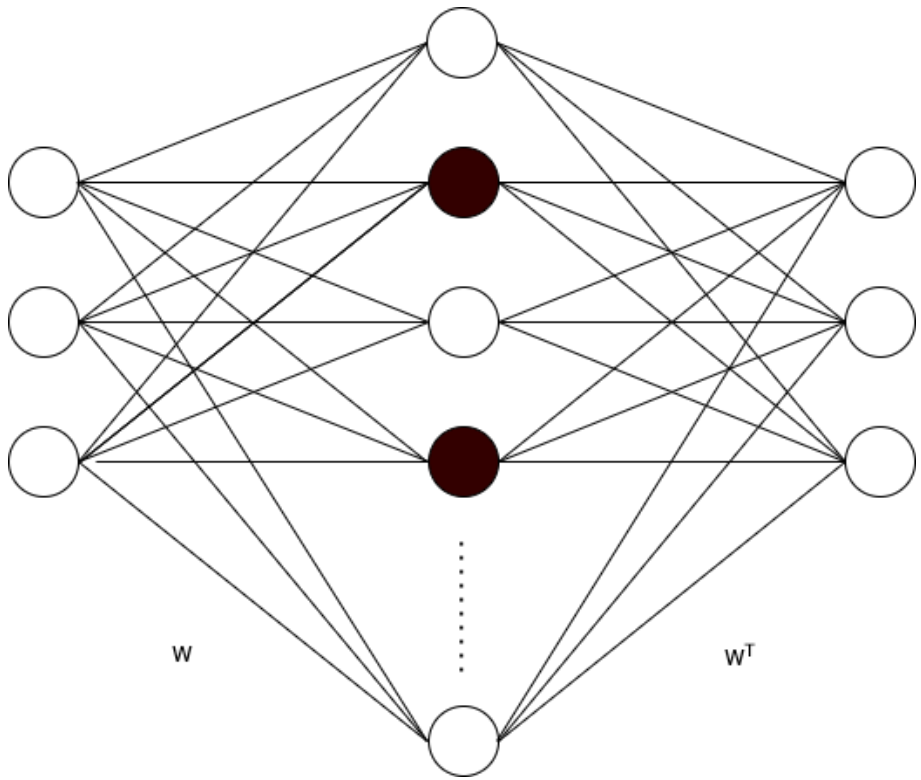
Given an activation $\mathbf{x} \in \mathcal{X}$, e.g from a Transformer MLP layer, $\hat{\mathbf{x}}$ is:

$$\textit{Encoder}: \quad \mathbf{h} = \text{ReLU}(W_{enc}\mathbf{x} + b_{enc})$$

$$\textit{Decoder}: \quad \hat{\mathbf{x}} = W_{dec}\mathbf{h} + b_{dec}$$

And the loss we optimize for is:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = |\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \alpha||\mathbf{h}||_1$$



Learning a dictionary with an SAE

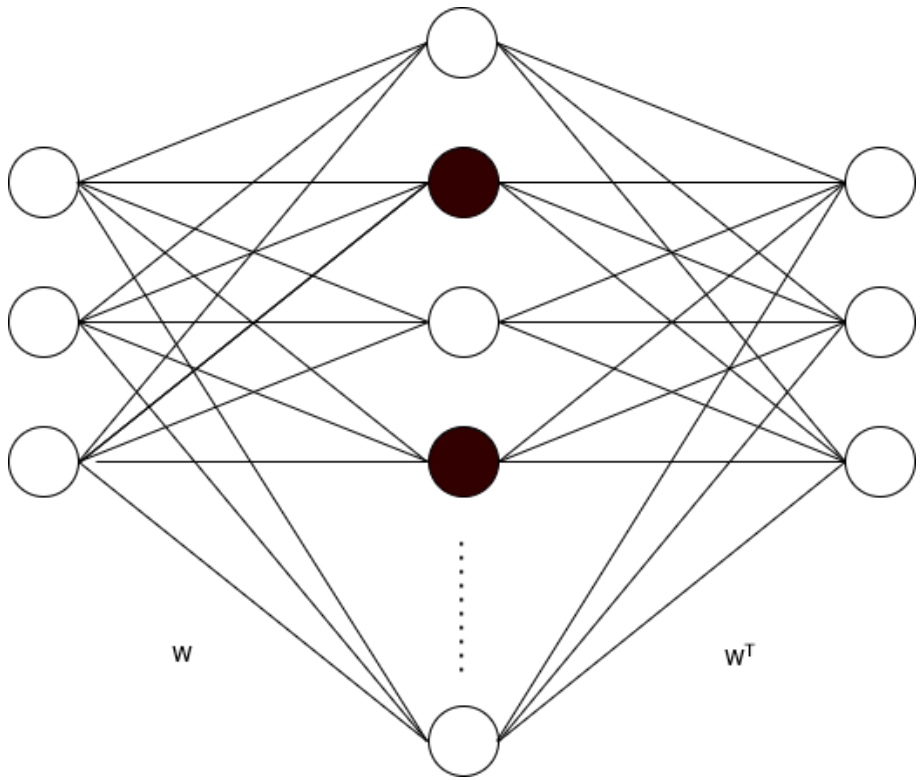
Given an activation $\mathbf{x} \in \mathcal{X}$, e.g from a Transformer MLP layer, $\hat{\mathbf{x}}$ is:

$$\textit{Encoder}: \quad \mathbf{h} = \text{ReLU}(W_{enc}\mathbf{x} + b_{enc})$$

$$\textit{Decoder}: \quad \hat{\mathbf{x}} = W_{dec}\mathbf{h} + b_{dec}$$

And the loss we optimize for is:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = |\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \alpha||\mathbf{h}||_1$$



Taking features out of superposition

- The $\ell 1$ incentives the reconstruction to be a sparse linear combination of the features.
  - $\ell 1$ is the sum of the absolute values of the vector.

Taking features out of superposition

- The $\ell 1$ incentives the reconstruction to be a sparse linear combination of the features.
  - $\ell 1$ is the sum of the absolute values of the vector.

- α is a parameter that controls the strength of the regularization.

Taking features out of superposition

- The $\ell 1$ incentives the reconstruction to be a sparse linear combination of the features.
  - $\ell 1$ is the sum of the absolute values of the vector.
- α is a parameter that controls the strength of the regularization.
- W_{dec} will be our set of dictionary features f, so $W_{dec\,k} = \mathbf{f}_k$.
- h contains the sparse coefficients a_i, which is learned by the encoder.

Taking features out of superposition

- The $\ell 1$ incentives the reconstruction to be a sparse linear combination of the features.
  - $\ell 1$ is the sum of the absolute values of the vector.
- α is a parameter that controls the strength of the regularization.
- W_{dec} will be our set of dictionary features f, so $W_{dec\,k} = \mathbf{f}_k$.
- h contains the sparse coefficients a_i, which is learned by the encoder.
- This has been shown, both empirically and theoretically, to "recover" $\mathbf{g}$.

What can we do with SAEs?

The assumption of this whole process is that the feature space of life is somewhat sparse. LMs give us a compressed version of the feature space of text, so we can use SAEs to try to recover the original signal, given that it is indeed sparse.

What can we do with SAEs?

- Templeton et al. (2024) train SAEs on the hidden activations of Claude 3 Sonnet.

What can we do with SAEs?

- Templeton et al. (2024) train SAEs on the hidden activations of Claude 3 Sonnet.
- By learning up to 34 million features in their dictionary, they find neurons that:
  - Always activate for any reference of the Golden Gate Bridge.
  - Always activate for any mention of tourist attractions.
  - Always activate for any mention of brain sciences.
- Disclaimer: Monosemanticity is in the eye of the beholder.

Evaluating interpretability

How do you know if your dictionary displays reduced polysemanticity?

Evaluating interpretability

How do you know if your dictionary displays reduced polysemanticity?

- Manual inspection
- Autoinspection using an LLM

Example

If we train a 1-layer GPT model OpenWebText (Gokaslan et al., 2019), with a hidden size of 512 in the MLP.

Example

If we train a 1-layer GPT model OpenWebText (Gokaslan et al., 2019), with a hidden size of 512 in the MLP.

And then reconstruct the MLP activations using an SAE that upscales $512 \rightarrow 4096$.

Example

If we train a 1-layer GPT model OpenWebText (Gokaslan et al., 2019), with a hidden size of 512 in the MLP.

And then reconstruct the MLP activations using an SAE that upscales $512 \rightarrow 4096$.

What do we get?

Example

Neuron # 2

Top Activations

Max Activation = 3.2091

blog:<|endoftext|>While the residents of Kyle Canyon
<|endoftext|>Edward Snowden, the whistleblower who leaked classified
!<|endoftext|>Researchers with the University of Tennessee Institute
animals.<|endoftext|>Getty The influx of immigrants could
Twitter<|endoftext|>2016 is the year that the concept
AlertMe<|endoftext|>ALL the photographs Lenny Com
Australia.<|endoftext|>After the Ambassador of the United
until his death, the FBI director had to
look<|endoftext|>Share. The pre-order mode
com.<|endoftext|>Only the grind of the season

Neuron # 31

Top Activations

Max Activation = 5.4505

by Tampa Bay and sat behind me on the
, so when I sat down with director and
his bravery when he sat out Saturday's 3
, the prime minister sat down and everyone became
cornerback Jeremy Lane also sat during the national anthem
committee. But we sat here at the beginning
in Israel, I sat down with Oded
. Lacey always sat to my left.
i midweek but sat out the other win
office since 1995 and sat on the House Appropriations

Subsample Interval 0

Max Activation = 2.4854

big man while he sits out the season because
remember thinking, while sitting on the floor of
you can get quite hung up on this child
because he has to settle for mutant sea bass
dive strategy roots lie in the Overwatch Premier

Examples from
`https://github.com/shehper/sparse-dictionary-learning`

Behaviour steering

- Templeton et al. (2024) show that we can use a learned dictionary to do *feature steering*.

Behaviour steering

- Templeton et al. (2024) show that we can use a learned dictionary to do *feature steering*.
- We can roof/floor the activation of a neuron to excite or dampen it.

Clamping the GGB neuron

Claude 3 Sonnet plain

Human: what is your physical form?

Assistant: I don't actually have a physical form. I'm an artificial intelligence. I exist as software without a physical body or avatar.

With The GGBN clamped 10x

Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables

F#34M/31164353

References I

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. (2022). Toy models of superposition. Transformer Circuits Thread.

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. (2019). Openwebtext corpus.

References II

Ng, A. et al. (2011). Sparse autoencoder. CS294A Lecture notes, 72(2011):1–19.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread.