

The background features a light gray grid with several wavy, overlapping lines in a slightly darker shade of gray, creating a sense of movement and depth. The text is centered in a bold, black, sans-serif font.

Interpretability: Overview, Limitations, & Challenges

Unity Collective

Interpretability

Interpretability: extent to which a model's decisions, predictions, or internal workings can be understood/explained by humans

- **Concept-based:** explaining model decisions in terms of high-level concepts or features
- **Mechanistic-based:** focus on understanding inner workings of model

The background features a light gray grid with several thick, wavy, black lines that flow across the frame, creating a sense of movement and depth. The text is centered in a bold, black, sans-serif font.

**Why is interpretability
important?**

The background features a light gray grid with several wavy, shaded bands in shades of gray and white, creating a modern, abstract aesthetic.

01

Salient Explainers

Image Data

Salient Map

- **Idea:**
 - Explains how a network responds to an individual sample image.
 - It boils down to gradient computation of output with respect to input
 - End result: a map with the same dimensionality with input data, showing each input part's importance (gradient)
- **Existing approach:**
 - GradientxInput (Shrikumar et al., 2017)
 - SmoothGrad (Smilkov et al., 2017)
 - Integrated (Sundararajan et al., 2017),
 - Guided Backpropagation (Springenberg et al., 2015)
 - GradCAM (Selvaraju et al., 2016))

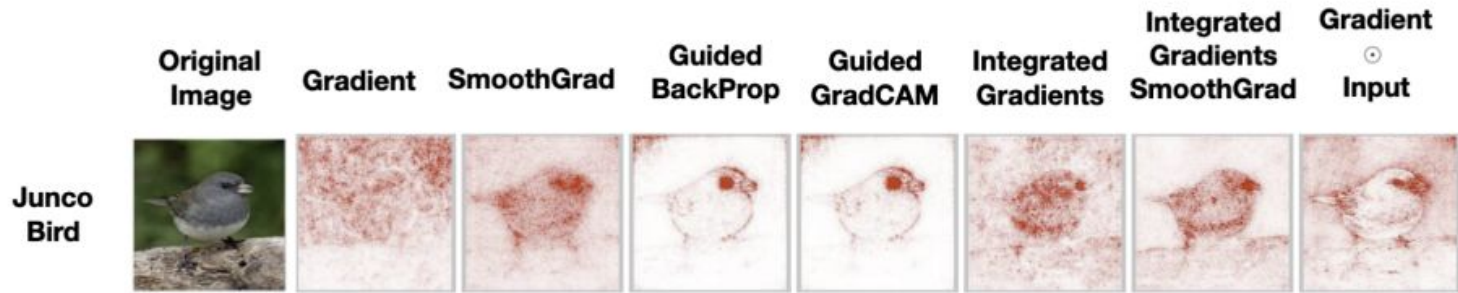


Figure 2. Saliency maps for an input image of a bird (left) generated by different saliency methods. While the vanilla gradient method output is noisy, the other methods “improve” the map visually. Figure adapted from (Adebayo et al., 2018)

Saliency Map Demo

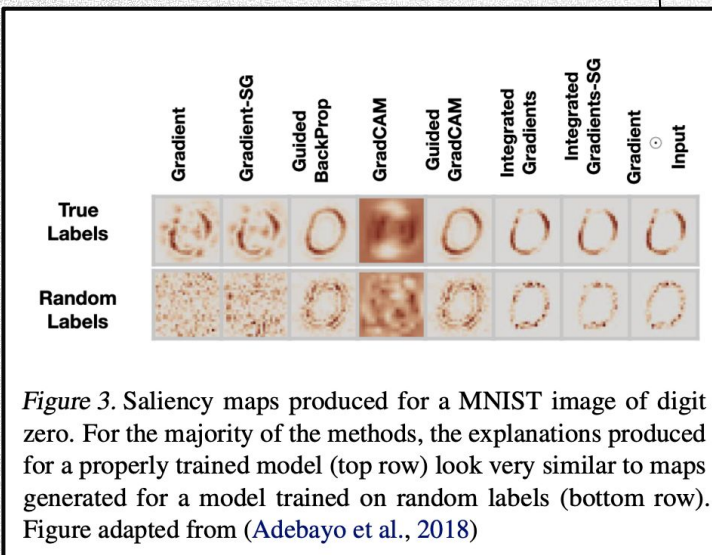
- **CLIP (Contrastive Language–Image Pre-training)** is a neural network that works with both images and texts
 - Trained to predict which randomly sampled text snippets are close to a given image, meaning that a text better describes the image
- Use salient map to explain how model makes prediction:
 - Some regions of the image are closer to the text query than others
 - This difference can be used to build the saliency map
- Notebook [link](#)

Query: "Who developed the Theory of General Relativity?"



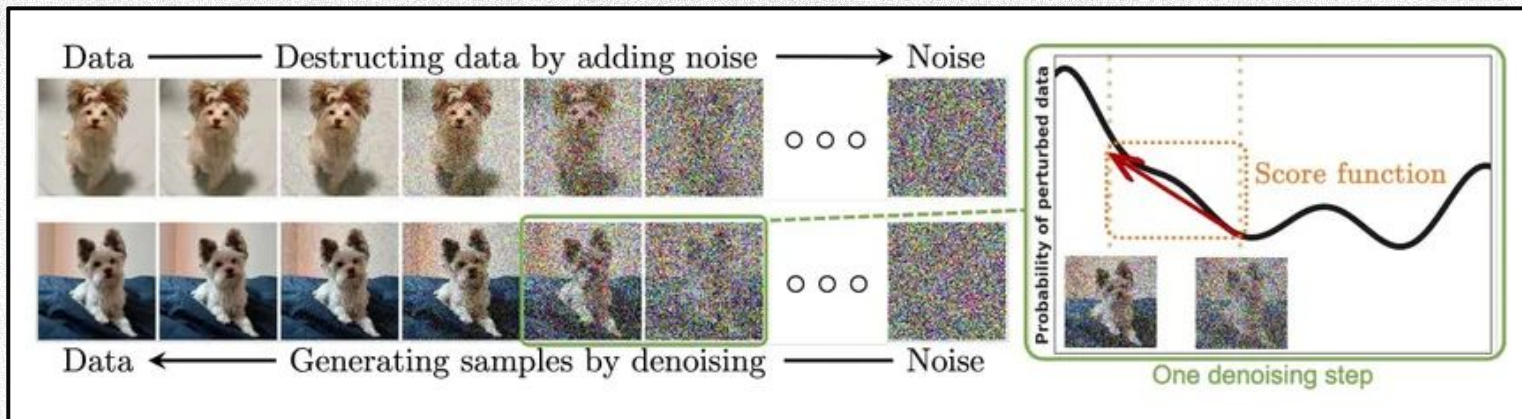
Limitations

- None of those methods were evaluated in a quantitative way:
 - Hard to find metrics or reliable ones
- Lacked a direct connection and guarantees related to how well the “explanations” correspond to the model’s “reasoning” process
- Experiments questioning reliability
 - A sanity check or an evaluation protocol is not a task-independent indicator of the saliency method’s validity



Discussion

- How do you imagine salient explainers would fit into the context of generative AI models (i.e. diffusion models, Bayesian Flow Networks, GANs)?



02

Attention Explainers

Textual Data

Attention “Explanations”

- Currently used across different tasks, but focus on NLP domain for this section
- Attention weights can be considered “importance” weights: the bigger the weight, the more critical the input element is

Question: Where is Sandra ?

Original Attention: John travelled to the garden . Sandra travelled to the garden

Figure 4. A model with the attention module was trained to solve a question answering problem. For the presented question, the greatest attention weight was attributed to the word *garden*. Assuming attention correlates with importance, this might indicate that the word *garden* was crucial for the model prediction. Example adapted from (Jain & Wallace, 2019)

Limitations: Attention is not Explanation

- Interpreting weights as “importance” is not well-defined
- Attention weights do not correlate w/ other feature importance measures
- Alternative attention weights do not significantly change model predictions
- Inconsistencies in evaluation methods

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α
 $f(x|\alpha, \theta) = 0.01$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

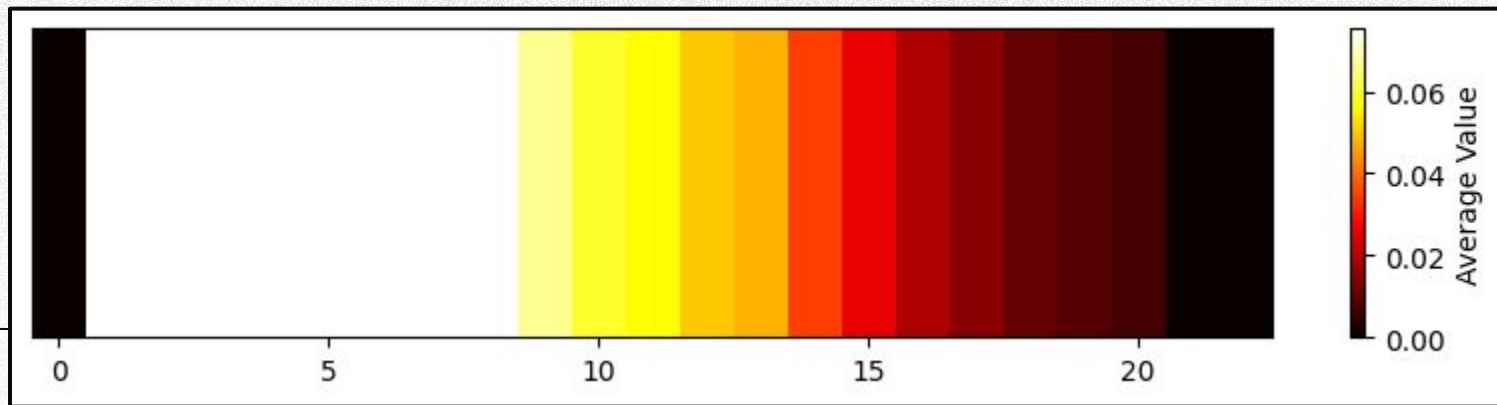
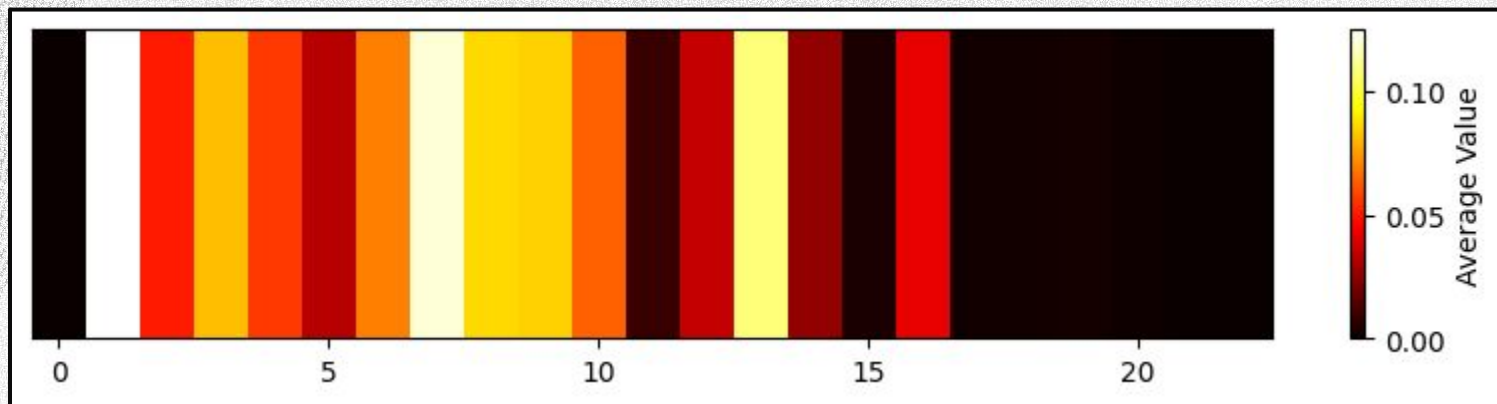
adversarial $\tilde{\alpha}$
 $f(x|\tilde{\alpha}, \theta) = 0.01$

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

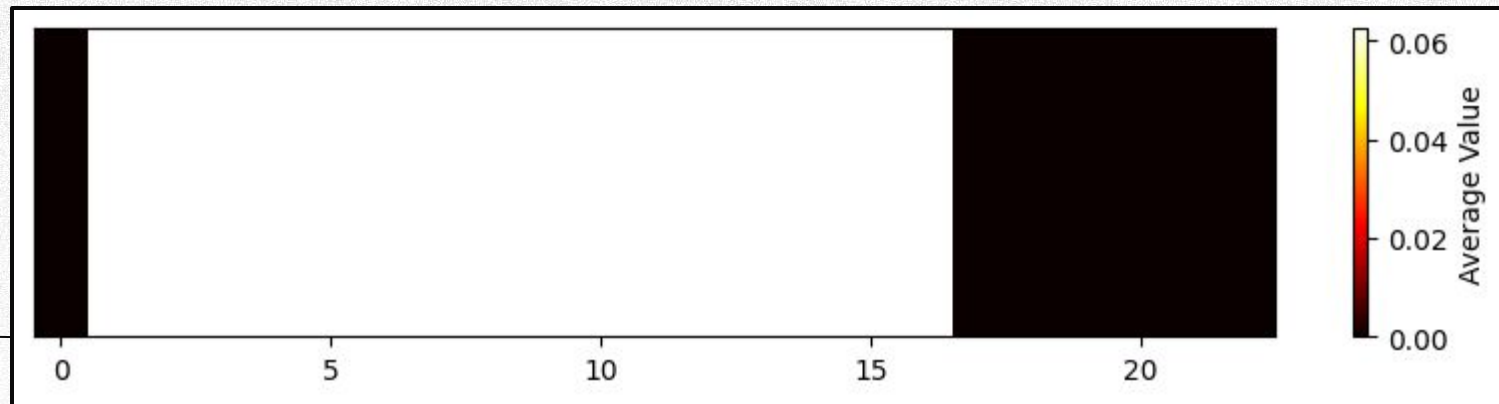
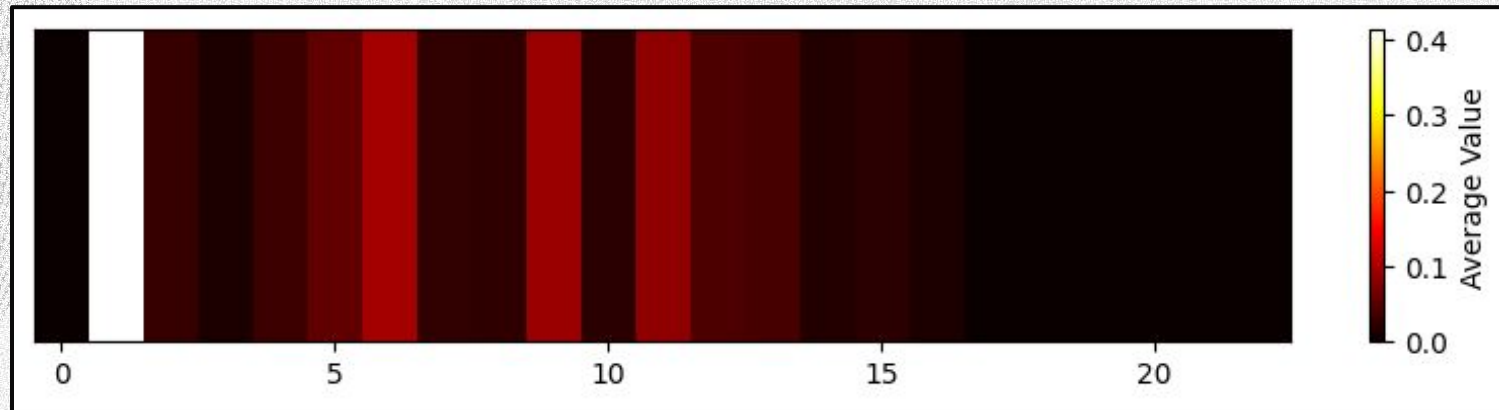
The background features a light gray, textured surface with a grid of thin black lines. Overlaid on this are several wavy, concentric lines that create a sense of depth and movement, resembling a stylized landscape or a topographical map. The overall aesthetic is clean and modern.

Demo: *Attention is not not Explanation*

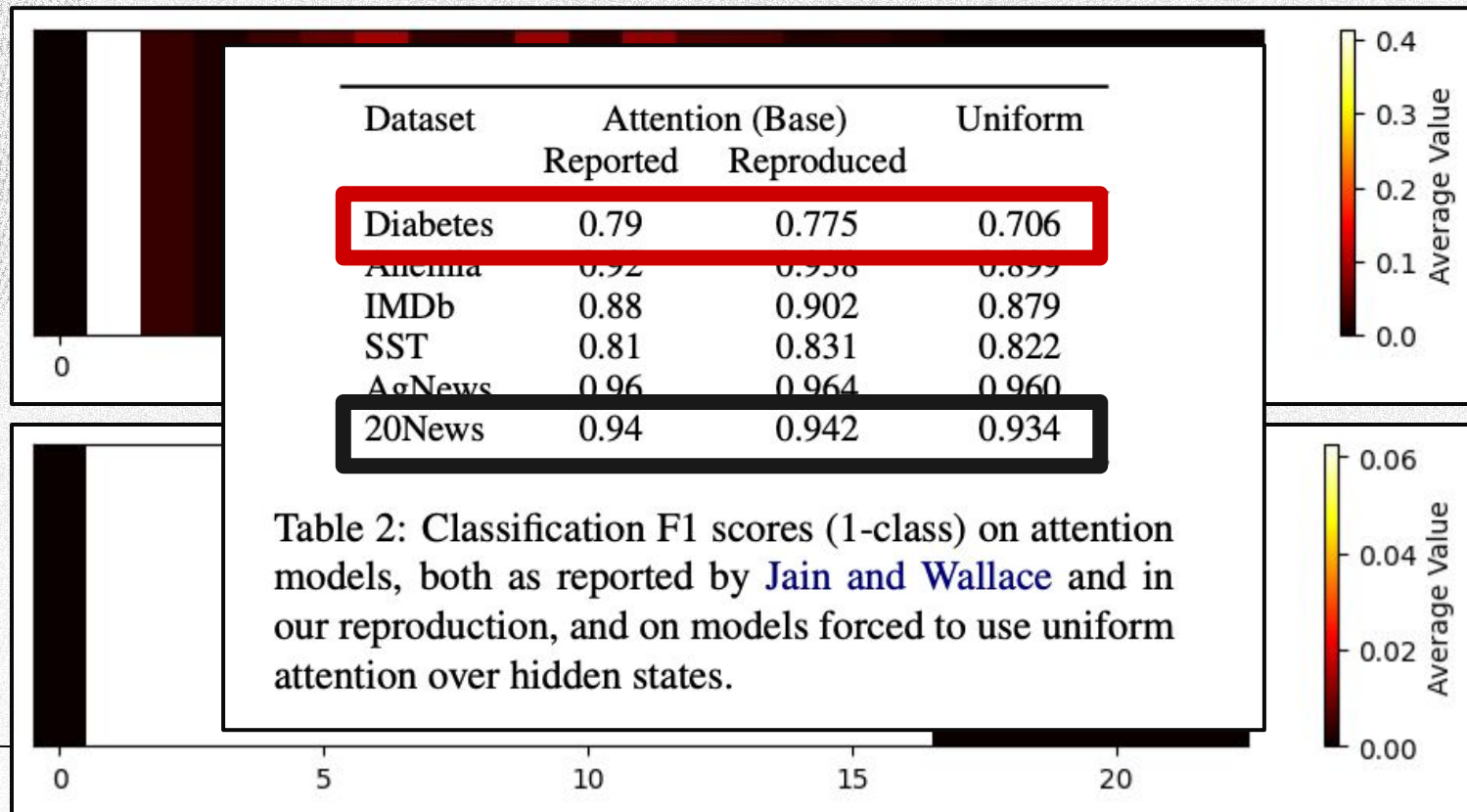
Frozen Attention Weights (Average)



Frozen Attention Weights (Instance)



Frozen Attention Weights (Instance)

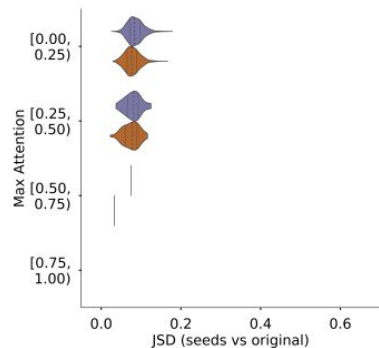


Attention Might Be Explanation

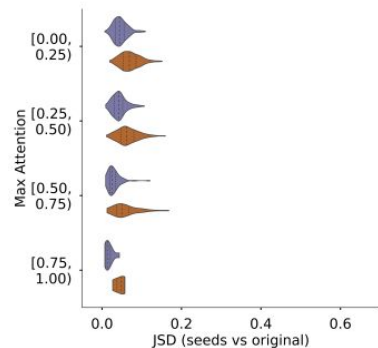
Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	0.753	0.931	0.824	0.905

Table 3: F1 scores on the positive class for an MLP model trained on various weighting guides. For ADVERSARY, we set $\lambda \leftarrow 0.001$.

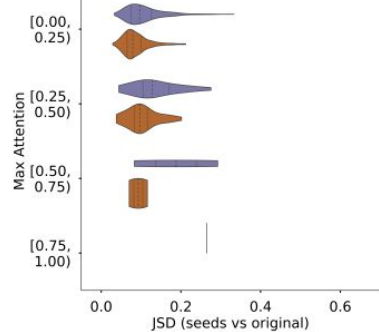
Attention Might Be Explanation



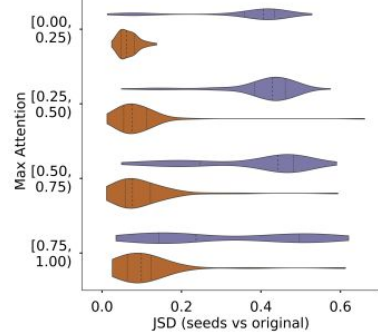
(a) IMDB (seeds)



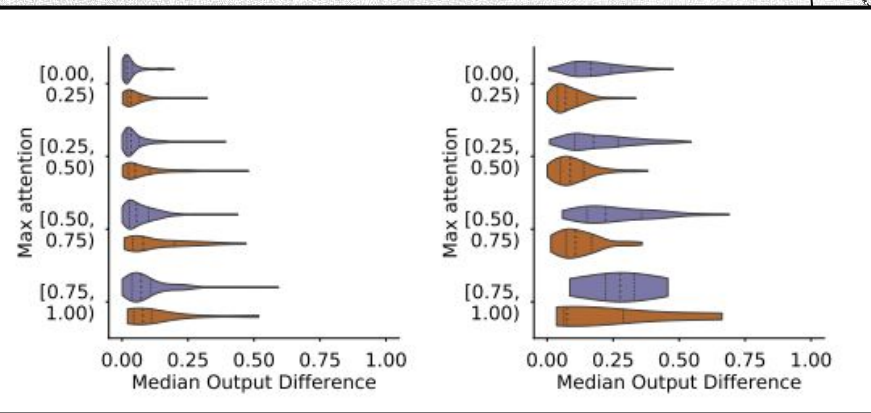
(c) SST (seeds)



(b) Anemia (seeds)



(d) Diabetes (seeds)



The background features a light gray, textured surface with several thin black lines forming a grid. Overlaid on this are large, curved, concentric patterns that resemble ripples or sound waves, primarily located in the top-left and bottom-left corners. In the center, the number '03' is enclosed within a thin black circle.

03

Discussion

Questions & Thoughts

Discussion

- What are some types of features/answers that you're looking for when using XAI tools?
- What are the requirements/metrics that you need in order to decide whether XAI tools are useful and working as intended?
- What directions/suggestions do you have to solve those issues?

The background features a light gray grid with wavy, shaded patterns in the corners. A large, thin-lined circle is centered in the upper half of the page.

04

Towards Provably Useful XAI

Future Directions

Is Task-Agnostic Explainable AI a Myth?

*“for instance, positive feature attribution **does not**, in general, imply that increasing the feature will increase the model output. Similarly, zero feature attribution **does not**, in general, imply that the model output is insensitive to changes in the feature.”*

- The authors conclude that without **Task-Inspired** techniques, there are no guarantees that these approaches offer useful real-world applications
 - Lack of distinction in some studies b/w performance of a model with and without explanations
 - XAI evaluations involving humans often rely on simplistic proxy tasks or subjective opinions on explanation quality
- With these techniques, however, under rigorous evaluation it *may* be possible to find meaningful results

The background features a light gray, textured surface with a grid of thin black lines. Overlaid on this are several sets of wavy, concentric lines that create a sense of depth and movement, resembling a stylized landscape or a series of ripples. The overall aesthetic is clean and modern.

Thank you!

Questions?

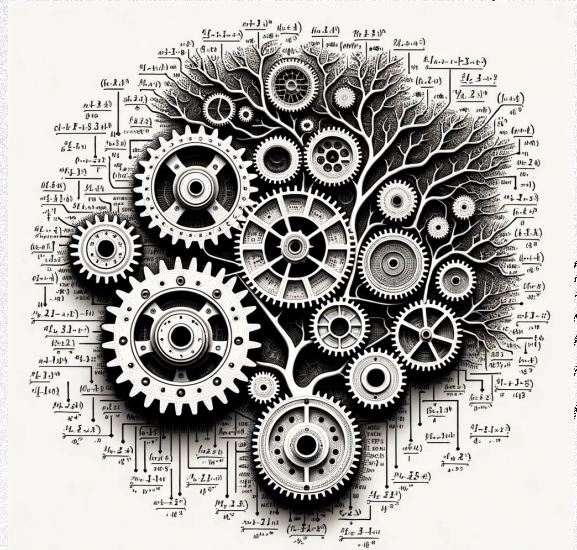
The background features a light gray grid with several wavy, overlapping lines in a slightly darker shade of gray, creating a sense of motion and depth. The text is centered in the middle of the frame.


MECHANISTIC Interpretability

Unity Collective

MECHANISTIC Interpretability

- **Mechanistic:** Understanding inner workings of models
 - Tracing input -> output
- Differs from **Concept-based** Interpretability which uses high-level concepts that are meaningful to humans
 - e.g. having a sub-network determine if a particular symptom is present



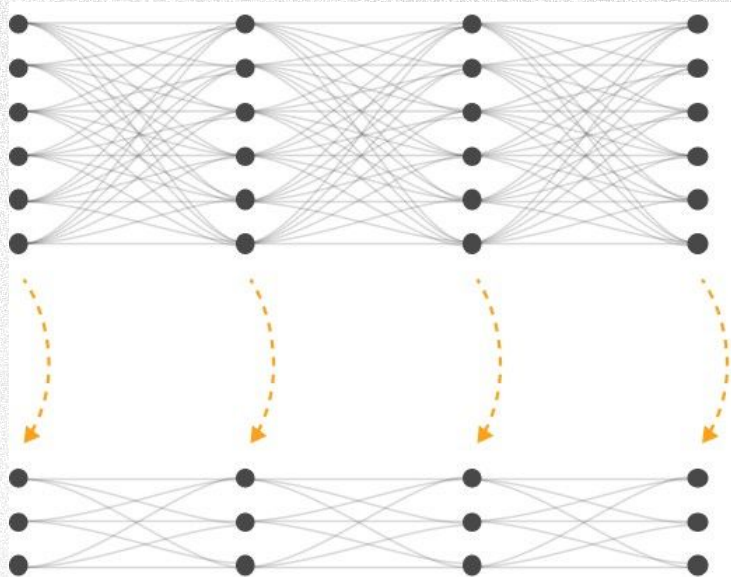
The background features a grid of thin black lines. Overlaid on this are several wavy, shaded regions that resemble a topographical map or a stylized landscape. The shading is composed of fine, parallel lines that create a sense of depth and movement. The overall aesthetic is clean, modern, and technical.

01

SoLU

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer El Showk, Nicholas Joseph, Nova DasSarma, Ben Mann, and others (Anthropic AI). **Softmax Linear Units**
Transformers Circuit Thread, 2022

Superposition Hypothesis



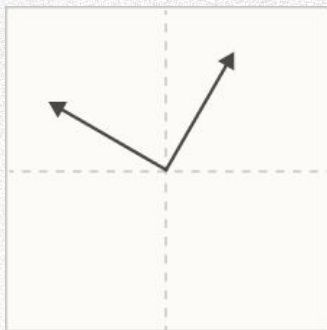
Under the superposition hypothesis, the neural networks we observe are **simulations of larger networks** where every neuron is a disentangled feature.

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 polysematicity.

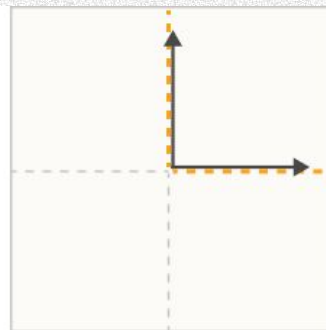
Solutions to Superposition

1. Create models with less superposition
 - Softmax Linear Units (SoLU)
2. Find a way to understand representations with superposition



In a **non-privileged basis**, features can be embedded in any direction. There is no reason to expect basis dimensions to be special.

Examples: word embeddings, transformer residual stream



In a **privileged basis**, there is an incentive for features to align with basis dimensions. This doesn't necessarily mean they will.

Examples: conv net neurons, transformer MLPs

SoLU vs GeLU

GeLU = Gaussian Error Linear Unit, approx: $\text{sigmoid}(1.7x) * x$

$$\text{SoLU}(x) = x * \text{softmax}(x)$$

SoLU vs GeLU

GeLU = Gaussian Error Linear Unit, approx: $\text{sigmoid}(1.7x) * x$

$$\text{SoLU}(x) = x * \text{softmax}(x)$$

SoLU increased interpretability at a major performance cost, so...
applying an extra LayerNorm after the SoLU:

$$f(x) = \text{LN}(\text{SoLU}(x)) = \text{LN}(x * \text{softmax}(x))$$

SoLU Motivating Examples

$$\text{SoLU}(4, 1, 4, 1) \approx (2, 0, 2, 0)$$

$$\text{SoLU}(4, 0, 0, 0) \approx (4, 0, 0, 0)$$

$$\text{SoLU}(1, 1, 1, 1) \approx \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

Discourages polysemanticity by:

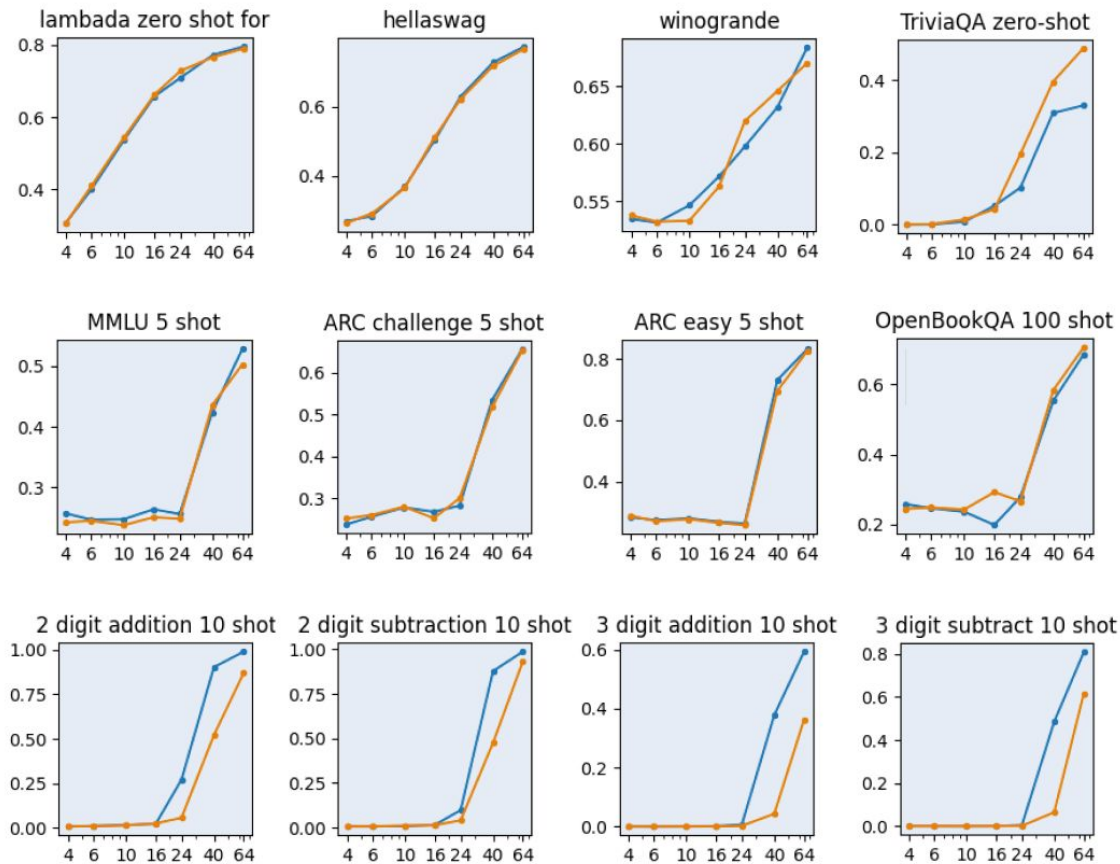
- Large values suppress smaller values
 - Lateral inhibition
- Basis aligned vectors are preserved
 - Approximate activation sparsity
- Features spread dimensions will have a smaller magnitude
 - Approximate superlinearity

Performance vs. Explainability: a limited tradeoff

Downstream Task Performance vs Model Size

Performance of SoLu vs our standard transformer on several downstream evaluations, over a range of model sizes. Overall the SoLu model performs comparably to the baseline.

— SoLu — baseline



Does SoLU result in better explanations?

human evaluator
evaluates whether a
single hypothesis or
concept explains 80%
of the strongest firings

Dataset Examples

<EOT> are done. Lower heat if balls begin to get too dark-brown. Remove from fire and drain on absorbent paper. Serve at once while piping hot. Some people stick a toothpick in each cheese ball.

ANCHOVY CHEESE CANAPÉ_(Quickie!_)
OR SANDWICH SPREAD

includes information on smartphones, facial recognition, and social networks.

1

HOW THIS BOOK CAN MAKE YOU INVISIBLE

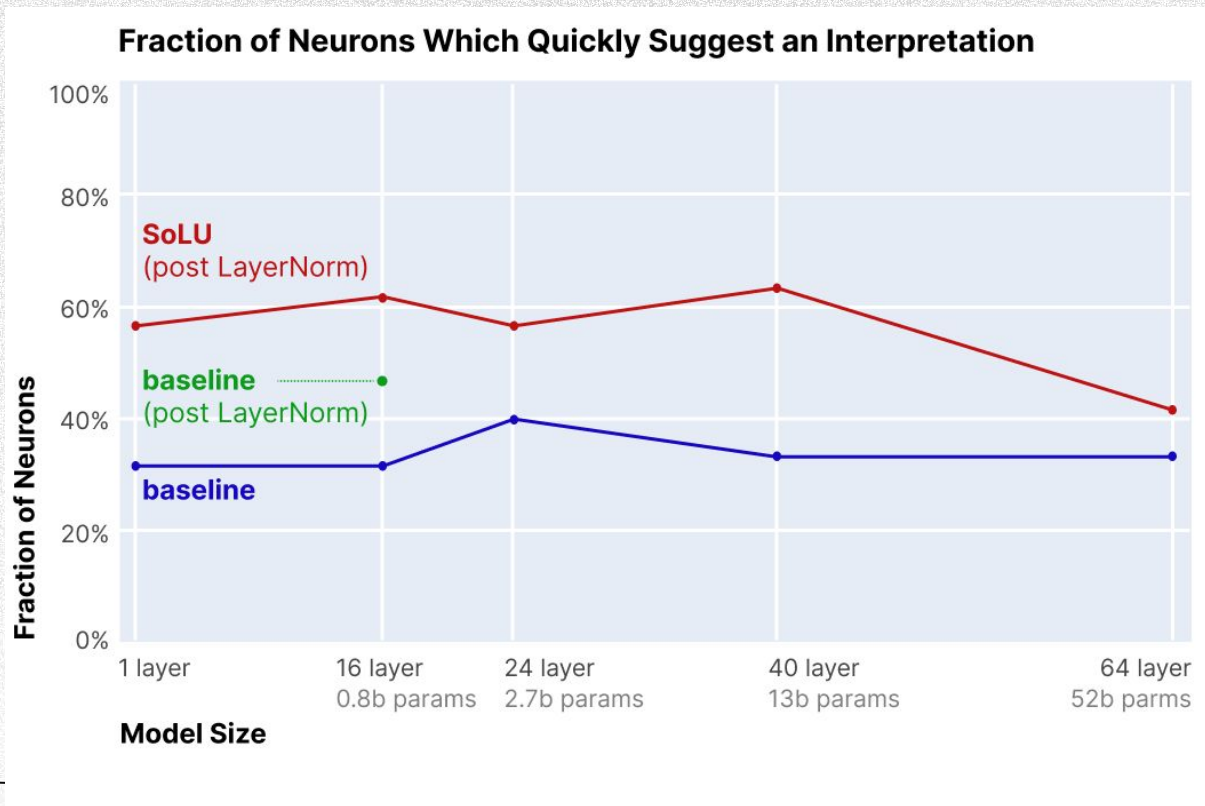
One day you can be on the top of the world; the next day you can be in hell. One of Wiley Miller's Non Sequitur cartoon strips is titled "LEGAL MUGGING." It shows a businessman

Americans take for granted. In short, the lives of both the Brothers and the Hallway Hangers have been severely circumscribed by their subordinate position in the class structure.

RACIAL DOMINATION: INVIDIOUS BUT INVISIBLE****

Both the Brothers and the Hallway Hangers are victims of class exploitation, but the African

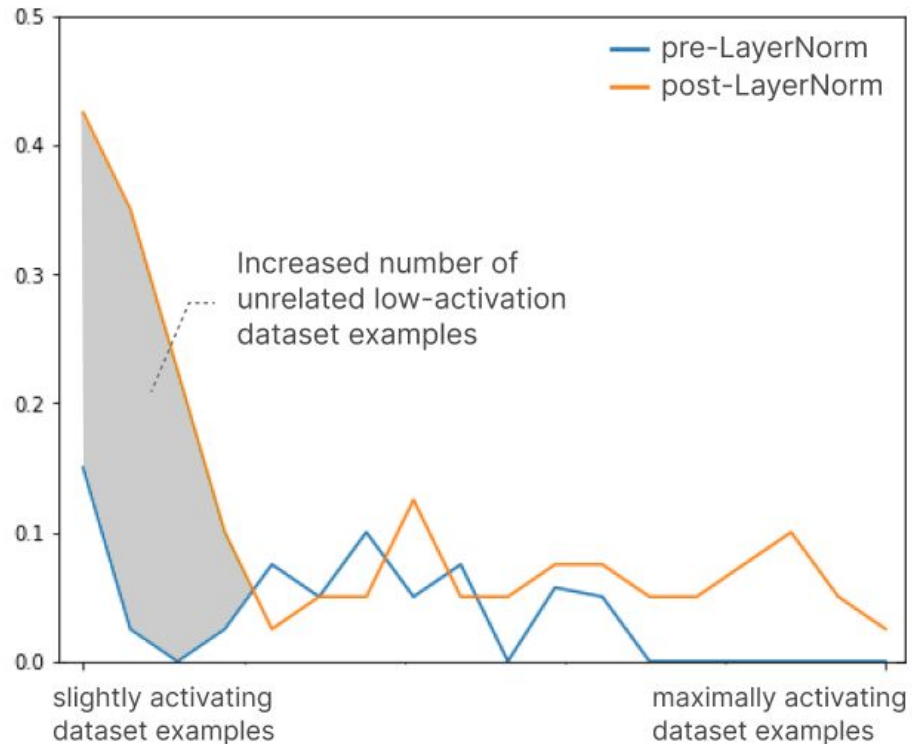
Does SoLU result in better explanations?



LayerNorm Complications

Fraction of Examples Inconsistent with Primary Hypothesis of Neuron

We categorize pre- and post-LayerNorm dataset examples across a range of activation levels based on whether they're inconsistent with the primary feature the neuron seems to respond to. After the LayerNorm, it's much more common for weak activation dataset examples to be totally unrelated. Note pre- and post- activations are on different scales, so we plot relative to maximum activation.



Class Activity

Based on tokens highlighted in text below (corresponding to certain neurons), can you come up with interpretable features mapping to these patterns?

This is a **sentence** where I **talk** about **interesting** stuff like **sencha** tea.

I love **running**, but I hurt my **ankle** last time I tried!

The **following** **paper** describes a new **method** for **inferring** unseen attributes- **using** just **model** predictions.

Some people think **AI** is **becoming** **sentient**, but that is debatable.

I ate **Corgi**-shaped cupcakes yesterday. They were delicious!

Have you seen my **cat** anywhere?

I'll be **sleeping** by the time this **movie** finishes

My assignment is due today!!

Are you **craving** some **chocolate**?

Meet my new **cat**- **Gustavo**

Can you **help** me **solve** for x: $5x - 25x^2 + 300 = 0$

The background features a light gray grid with thin black lines. On the right side, there are several concentric, wavy lines that curve towards the center. The overall aesthetic is clean and modern.

02

Monosemanticity

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, and others (Anthropic AI).
Towards. **Monosemanticity: Decomposing Language Models With Dictionary Learning.** Transformers
Circuit Thread, 2023

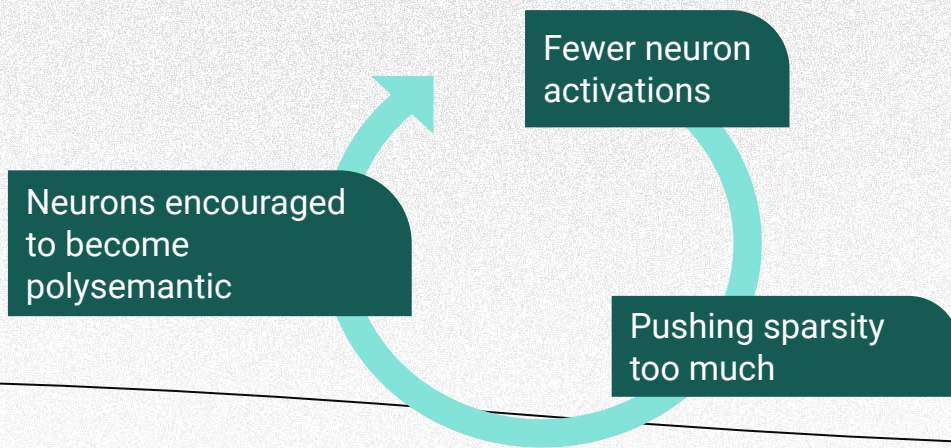
Architectural limitations

Can help eliminate superposition, but neurons remain largely non-interpretable

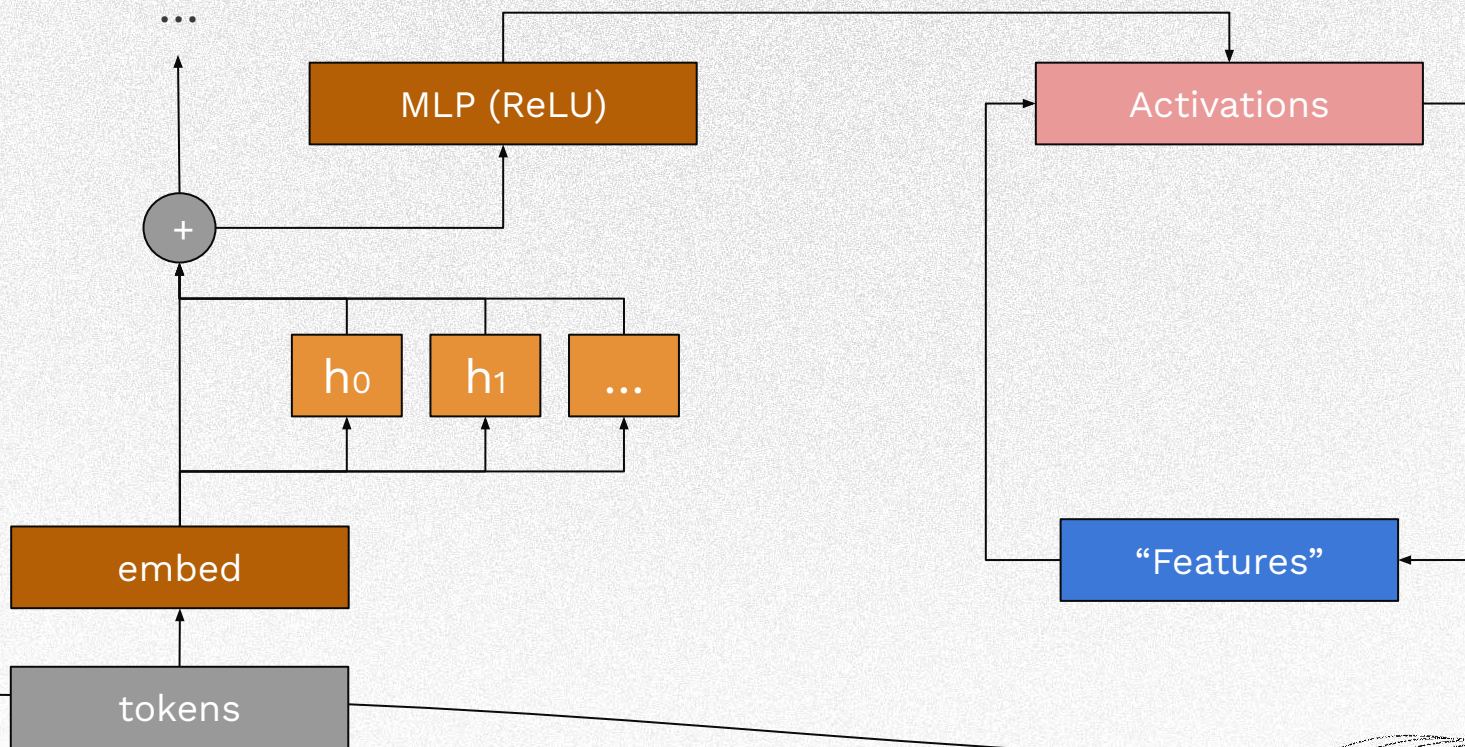
Toy example: Independent features {A, B, C, D}, Single, binary-output neuron.

Scenario 1: Only A useful: CE loss: $\frac{1}{4} * 3 * -\log(\frac{1}{3}) \sim 0.8$

Scenario 2: A/B equally useful: CE loss: $\frac{1}{2} * 2 * -\log(\frac{1}{2}) \sim 0.7$



Extracting features from neurons

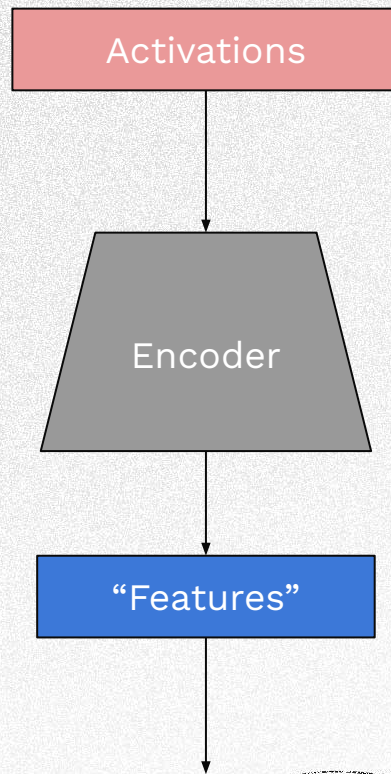


Features as a Decomposition

$$x_j \approx b + \sum_i f_i(x^j) d_i$$

Superposition hypothesis - these “features” are likely to form an overcomplete basis i.e., more directions than neurons

$$f_i(x) = \text{ReLU}(W_e(x - b_d) + b_e)_i$$



A “good” decomposition

$$x_j \approx b + \sum_i f_i(x^j) d_i$$

Describe points for which feature activates:

E.g., feature 4 -> {"Hello!", "Hey there!", "Bonjour", "How's it going?"}

Interpret downstream effects of changing features

E.g., $P(\text{feature 4}) \uparrow \rightarrow P(\text{negative sentiment}) \downarrow$

Features cover significant portion of layer functionality

Sparse Autoencoders

MSE loss: Avoid polysemanticity

Larger internal dimension: Overcomplete

L1-penalty: Sparsity

Input bias: Boosted performance (toy models)

```
import torch.nn as nn
import torch as ch

class SparseAutoEncoder(nn.Module):
    def __init__(self, lambda: float):
        self.lambda = lambda
        input_dim = 128
        hidden_dim = input_dim * 8
        # Parameters
        # Bias (tied with output)
        self.bias_d = nn.Parameter(ch.ones(input_dim,))
        # Encoder weights
        self.W_e = nn.Linear(input_dim, hidden_dim)
        # Decoder weights (only weights)
        self.W_d = nn.Linear(hidden_dim, input_dim, bias=False)

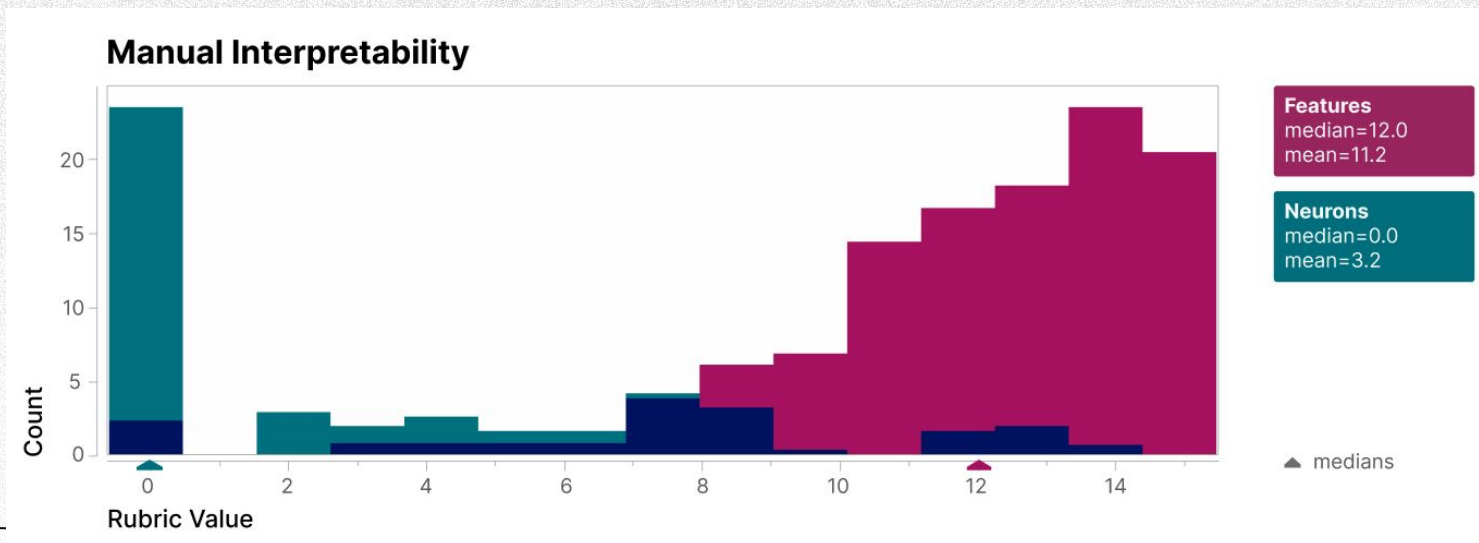
    def forward(self, x):
        x_hat = x - self.bias_d
        f = nn.ReLU(self.W_e(x_hat))
        x_cap = self.W_d(f) + self.bias_d
        return x_cap, f

    def loss(self, x, y):
        y_hat, f = self.forward(x)
        loss = nn.MSELoss()(y, y_hat)
        loss += self.lambda * ch.norm(f, 1)
        return loss
```


Are these features “interpretable”?

Top-activation samples may have neurons that “appear” monosemantic

Sample uniformly across all feature activations

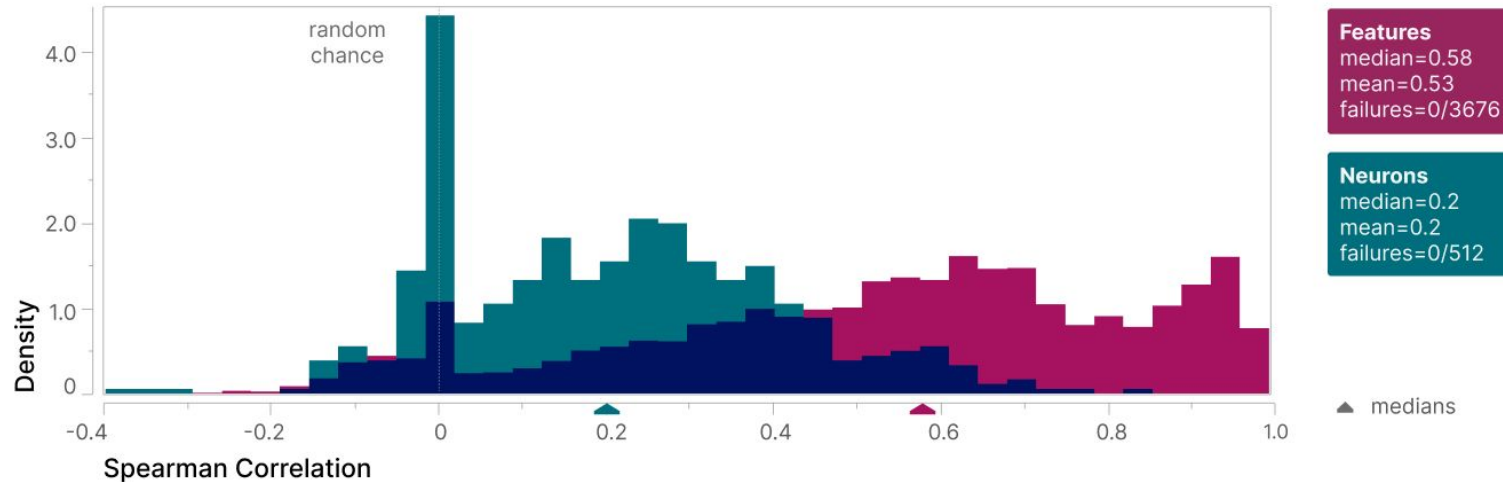


Automated evaluation

Use larger LLM to summarize using examples of tokens that activate feature

Predict unseen tokens using explanation

Automated Interpretability - Activation



Group Discussion

Interpretable features are able to explain ~80% of the loss (loss preserved when replaced with autoencoder reconstructions), are highly similar (~0.7 correlation) between models on same data.

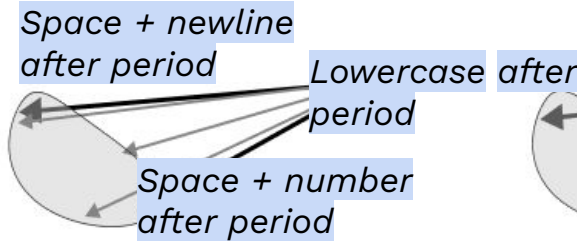
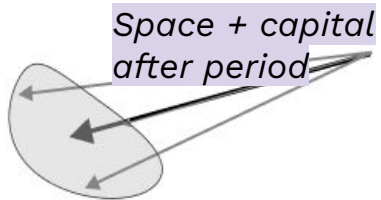
1. Is the lack of being able to explain the remaining 20% under given constraints (can, if sparsity constraints are slowly relaxed) an issue? Why/why not?
2. Is it expected to have such variance in a technique that is supposed to find “interpretable features”, even within models with the same architecture trained on the same dataset? What might be the reasons causing this
3. The authors used larger LLMs to look at examples and come up with “interpretability-related” concepts for those features. Can you think of any issues with this approach?

Feature Splitting

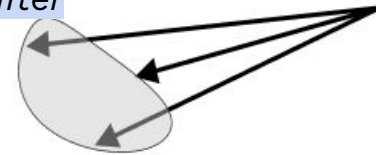
Features appear in “clusters”

Dictionary learning *can* learn all features with unlimited size, but is forced into structured superposition.

Increasing number of learned sparse features



Features Split →



Feature Splitting - Example

A/0 (512)

A/1 (4,096)

A/2 (16,384)

A/0/281

mathematical terminology and notation related to abstract algebra, especially homomorphisms, isomorphisms, and topological spaces.

A/1/437

mathematical quantifiers in a LaTeX context.

A/1/491

mathematical prose, especially in topology and abstract algebra.

A/1/3362

"the" in physics, especially field theory.

A/1/1652

"the" when preceding a term in physics, especially condensed matter physics.

A/2/15420

"every" and "each" in mathematical prose.

A/2/11964

quantity-related words in mathematical prose.

A/2/3962

mathematical prose, especially in category theory.

A/2/11307

"the" in math and technical writing.

A/2/2609

prepositions in physics and technical writing.

A/2/247

"the" and occasionally words after "the" in mathematical prose.

Takeaways

While architecture-based changes show promise, controlling polysemanticity with increasing sparsity is a cyclic problem

For trivial 1-layer Transformers, post-learning techniques (based on sparse dictionary learning) are promising and can help extract meaningful features from existing neurons