

CSCI 5541: Natural Language Processing

Lecture 14: Interpretability

Ryan Peters



UNIVERSITY OF MINNESOTA

Driven to Discover®

Outline

- ❑ General Overview of AI Interpretability
- ❑ Interpretable models
 - Linear regression, Decision trees
- ❑ Attribution and explainable methods
 - Integrated gradients and LIME
- ❑ Mechanistic Interpretability (of LLMs)
 - Transformer Overview
 - Superposition
 - SAEs
 - Introduction to Circuits: Induction Heads
 - Copying? (If time permits)



What is AI Interpretability

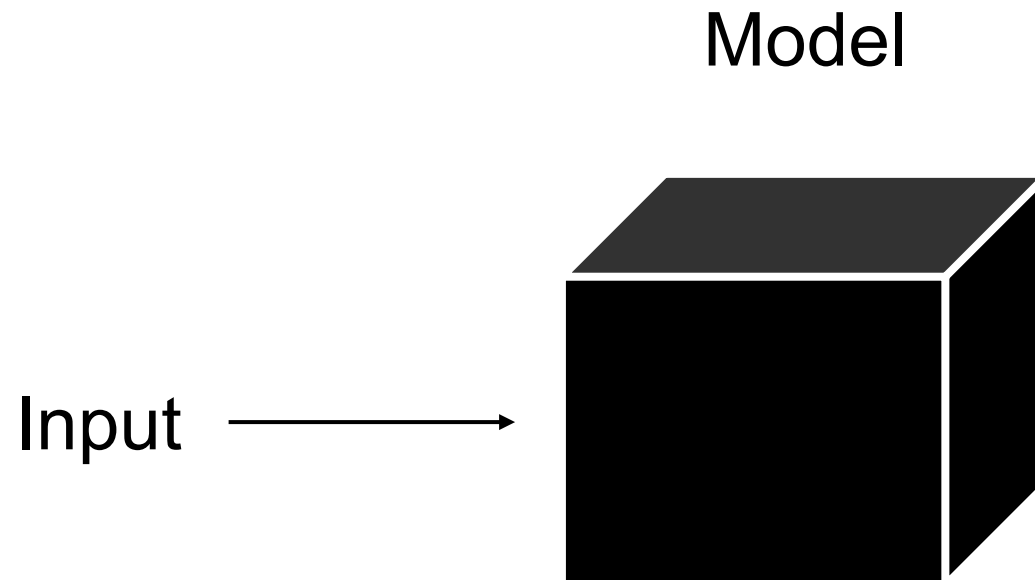


What is AI Interpretability

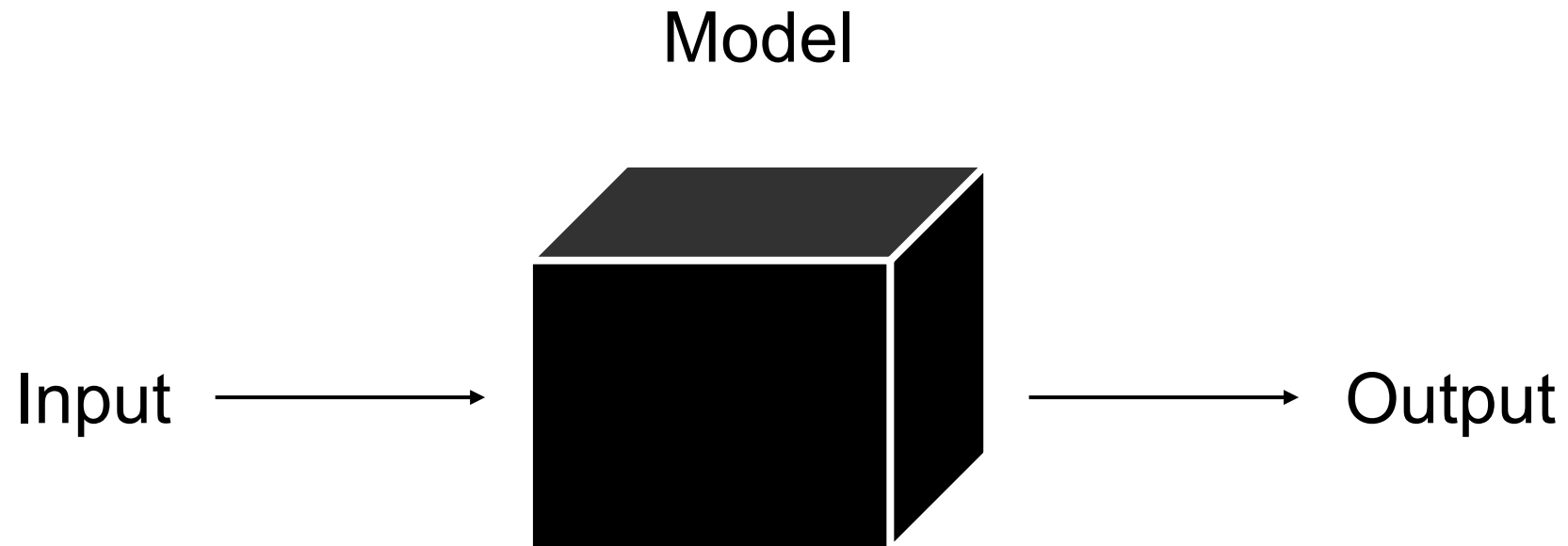
Input



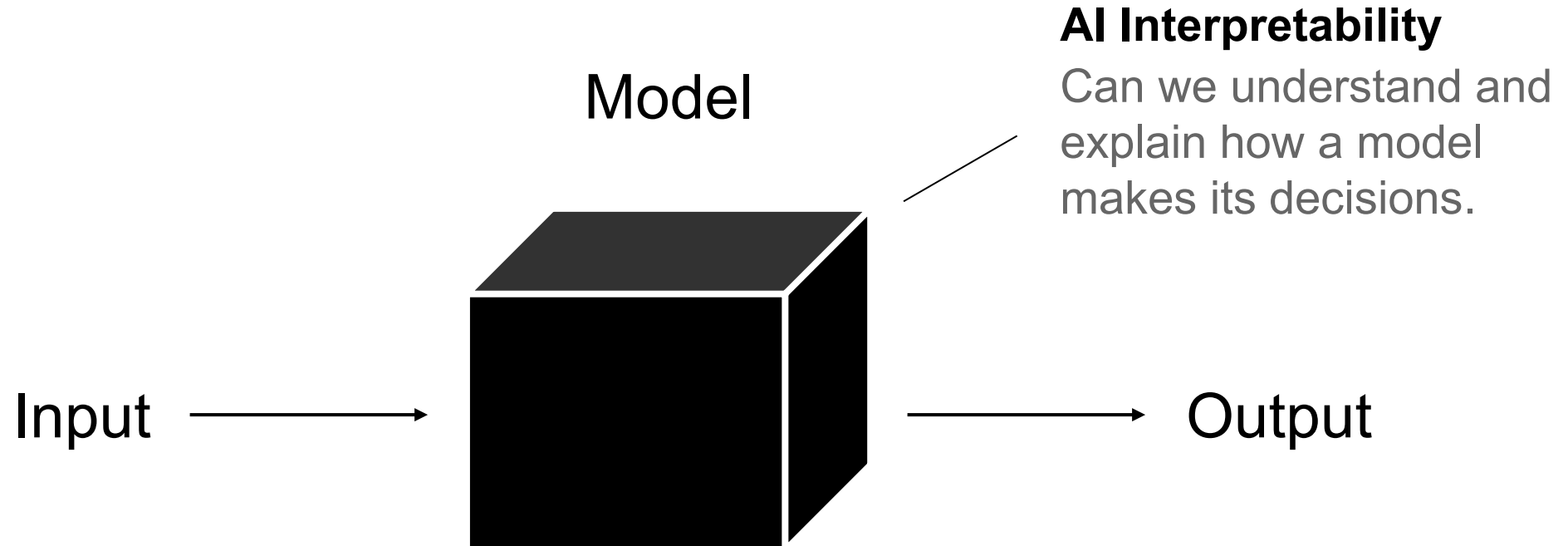
What is AI Interpretability



What is AI Interpretability

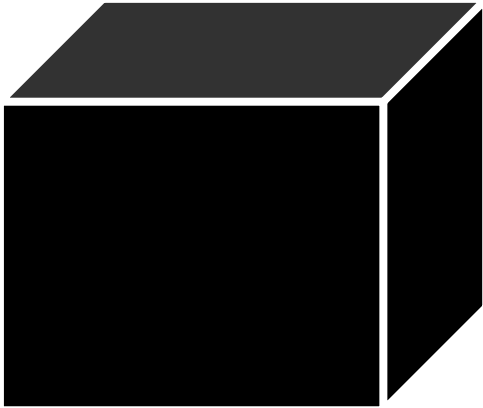


What is AI Interpretability



What is AI Interpretability

Model



Three subfields of interpretability that we will discuss

1. Models that are interpretable by design
2. Attribution methods (local and global)
3. Mechanistic interpretability

Interpretable Models



Interpretable Models

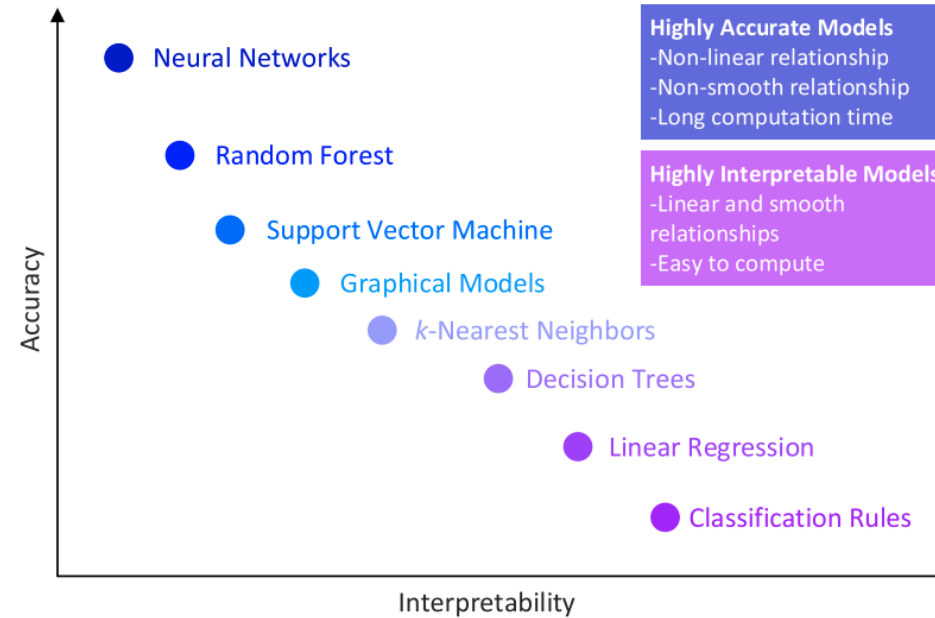


FIGURE 11. The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.

M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," in IEEE Access, vol. 7, pp. 137184-137206, 2019, doi: 10.1109/ACCESS.2019.2942390.

Linear Models

Say that we are performing automobile price prediction based on various features such as mileage, year made, number of doors, engine size, etc...



Linear Models

Say that we are performing automobile price prediction based on various features such as mileage, year made, number of doors, engine size, etc...

$$f(x) = a_0x_0 + a_1x_1 + a_2x_2 + b$$

x_0 : Mileage

x_1 : Year Made

x_2 : # Doors



Linear Models

Say that we are performing automobile price prediction based on various features such as mileage, year made, number of doors, engine size, etc...

$$f(x) = a_0x_0 + a_1x_1 + a_2x_2 + b$$

x_0 : Mileage

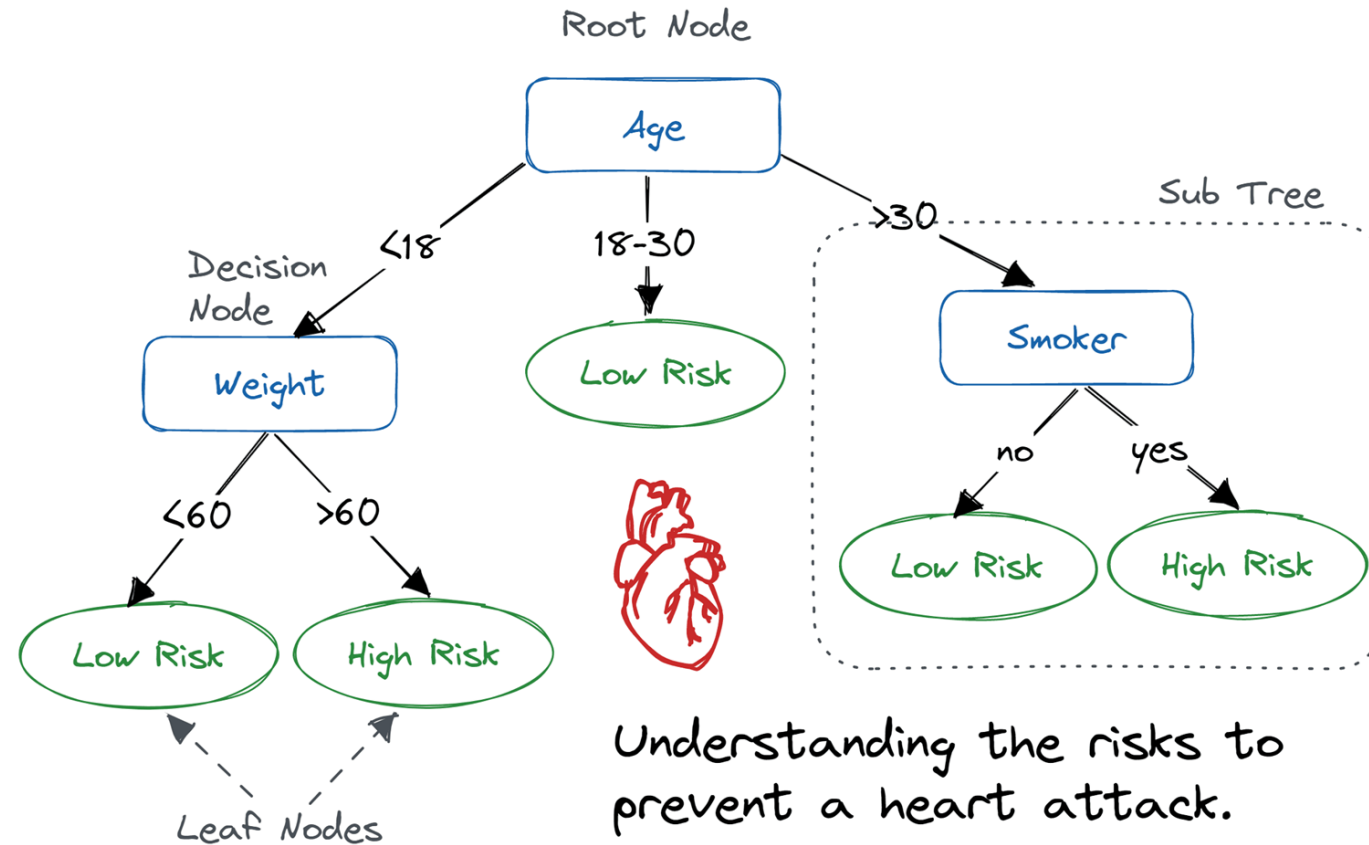
x_1 : Year Made

x_2 : # Doors

$$f(x) = (-0.5 * \text{mileage}) + (10 * \text{year made}) + (0 * \text{n_doors}) + 5000$$



Decision Trees



<https://medium.com/@shrutimisra/interpretable-ai-decision-trees-f9698e94ef9b>

Attribution Methods

What if our models are not interpretable?



Attribution Methods

Explain what features are useful for the model prediction.

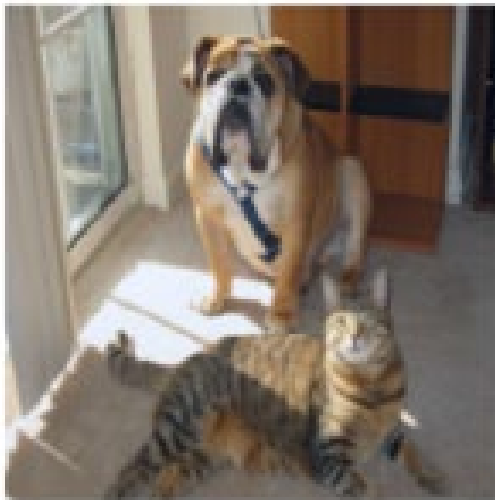


Original Image

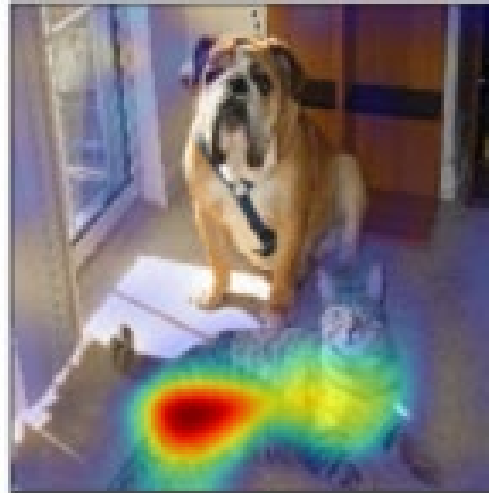
<https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/>

Attribution Methods

Explain what features are useful for the model prediction.



Original Image



Grad-CAM 'Cat'

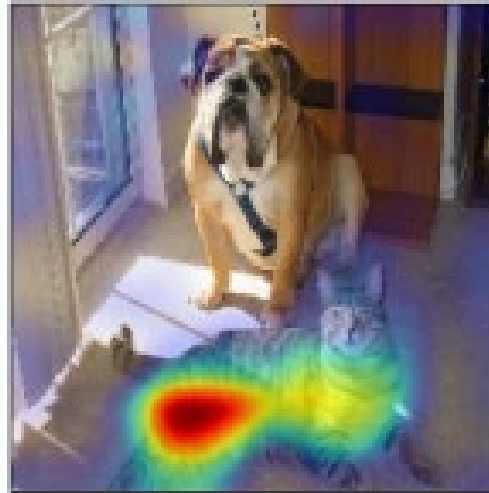
<https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/>

Attribution Methods

Explain what features are useful for the model prediction.



Original Image



Grad-CAM 'Cat'



Grad-CAM 'Dog'

<https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/>

Attribution Methods

Integrated Gradients



Integrated Gradients

Task: Classify 'sunny day' or not.

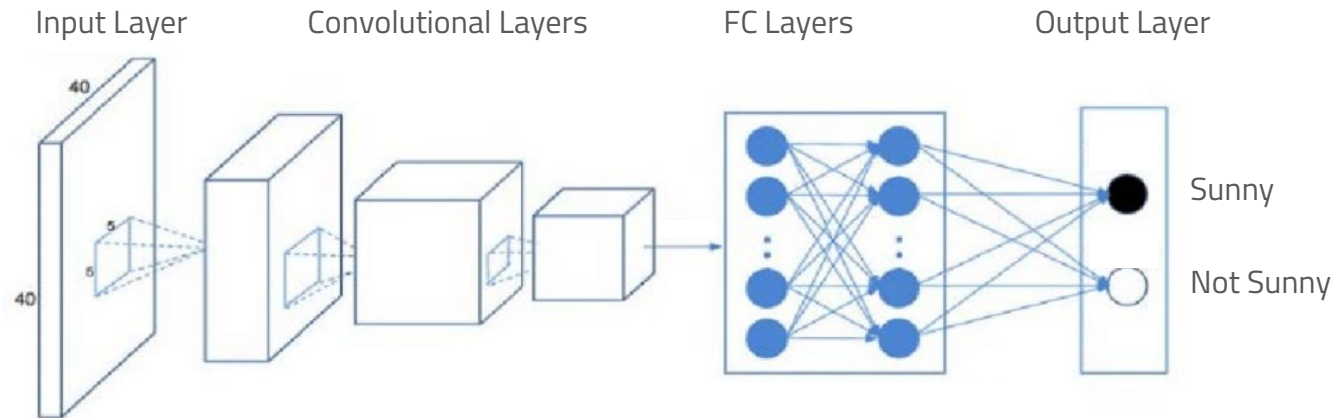
Prediction: 99% sunny



Integrated Gradients

Task: Classify 'sunny day' or not.

Prediction: 99% sunny



Integrated Gradients

Now, we want to know what of the input x (pixels x_i) the model ($F(x)$) is using in its prediction.

Prediction: 99% sunny



Key idea: Use the gradient of the output with respect to each pixel to determine its importance.

$$\frac{\partial F(x)}{\partial x_i}$$

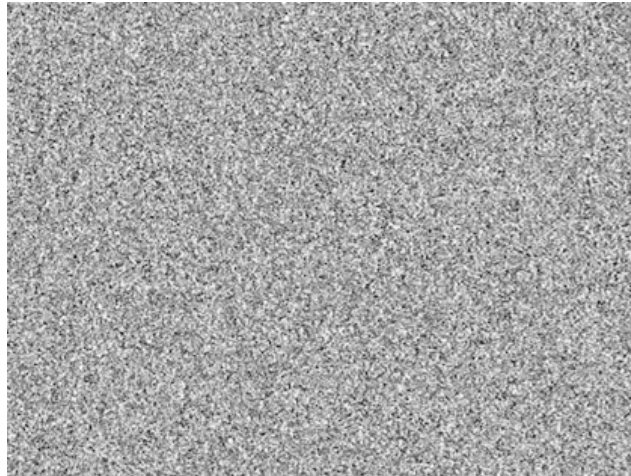
Integrated Gradients

The model prediction is saturated. Small changes will still result in the model predicting sunny with high probability.

Prediction: 99% sunny



Gaussian Noise



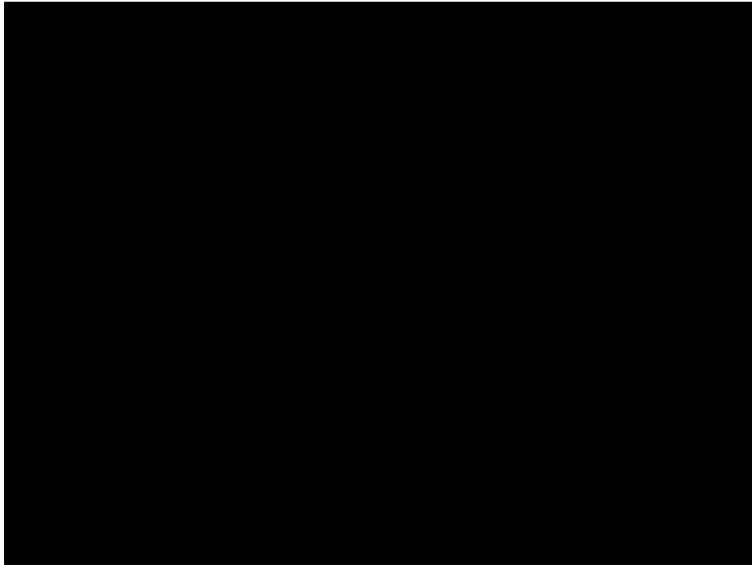
Prediction: 98% sunny



Integrated Gradients

The model prediction is saturated. Small changes will still result in the model predicting sunny with high probability.

Prediction: 1% sunny



Prediction: 99% sunny



Integrated Gradients

Want to capture how much a pixel changes the output from a baseline.

Want to avoid gradient saturation. So we integrate along a path from baseline to input.



Figure 2: Five-step interpolation between the baseline x' and the input image x . The first image on the left (alpha:0.0) is not a part of the interpolation process. Image source: Stanford Dogs [5]

<https://medium.com/@kernalpiro/xai-methods-integrated-gradients-6ee1fe4120d8>

Integrated Gradients

Want to capture how much a pixel changes the output from a baseline.

Want to avoid gradient saturation. So we integrate along a path from baseline to input.

$$\text{IG}_i(x) = (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$



Integrated Gradients

```
● ● ●  
  
# Generate interpolated inputs along the path  
alphas = torch.linspace(0, 1, steps)  
  
# Store gradients  
integrated_grads = torch.zeros_like(input_tensor)  
  
for alpha in alphas:  
    # Interpolate between baseline and input  
    interpolated_input = baseline + alpha * (input_tensor - baseline)  
  
    # Compute gradients  
    grads = self.compute_gradients(interpolated_input, target_class)  
  
    # Accumulate gradients  
    integrated_grads += grads  
  
# Average the gradients  
integrated_grads = integrated_grads / steps  
  
# Scale by (input - baseline)  
attributions = (input_tensor - baseline) * integrated_grads
```



Integrated Gradients

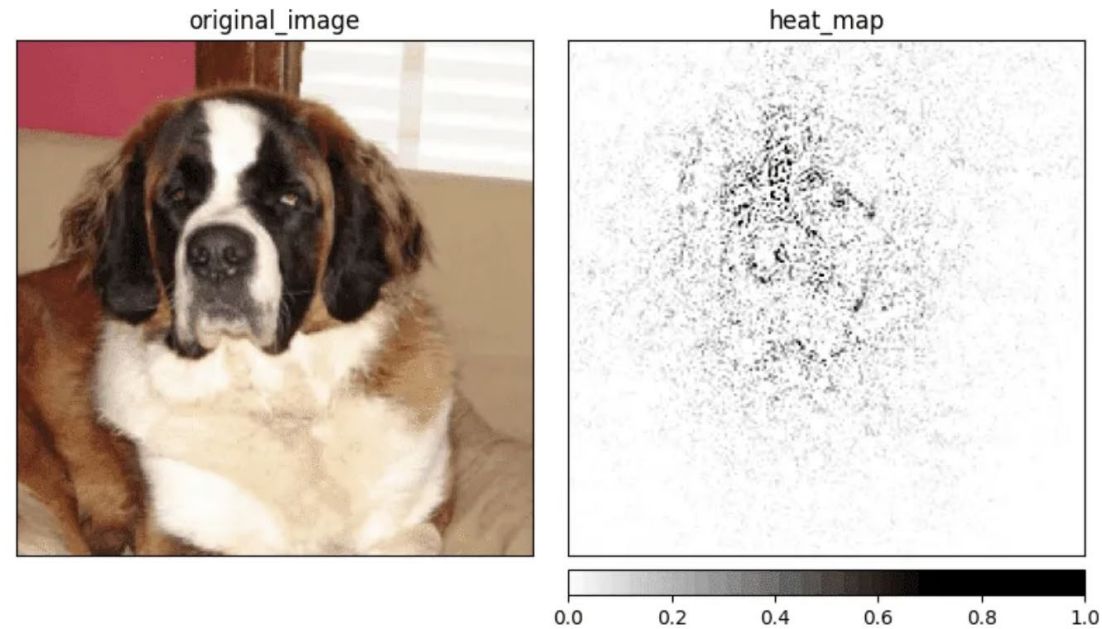


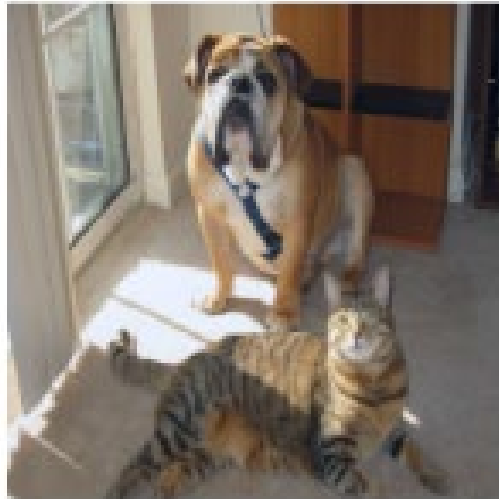
Figure 3: Visualization of the saliency map by the IG generated for the class saint_bernard. The result is averaged over 50 interpolation steps. Image source: Stanford Dogs [5]

<https://medium.com/@kemalpiro/xai-methods-integrated-gradients-6ee1fe4120d8>

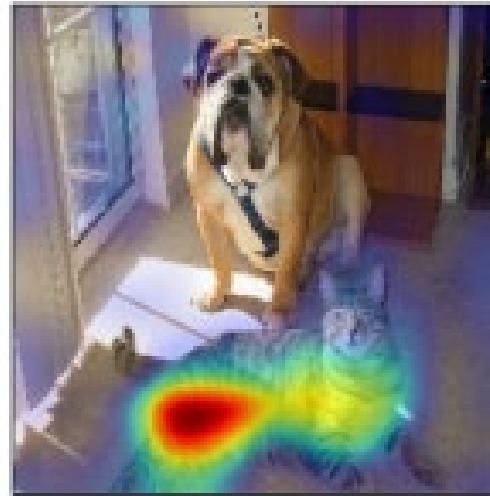
Integrated Gradients

Can do this for any class!

Image from Grad-CAM method, but the core idea still holds.



Original Image



Grad-CAM 'Cat'



Grad-CAM 'Dog'

<https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/>

Integrated Gradients

Why might this be useful?



Integrated Gradients

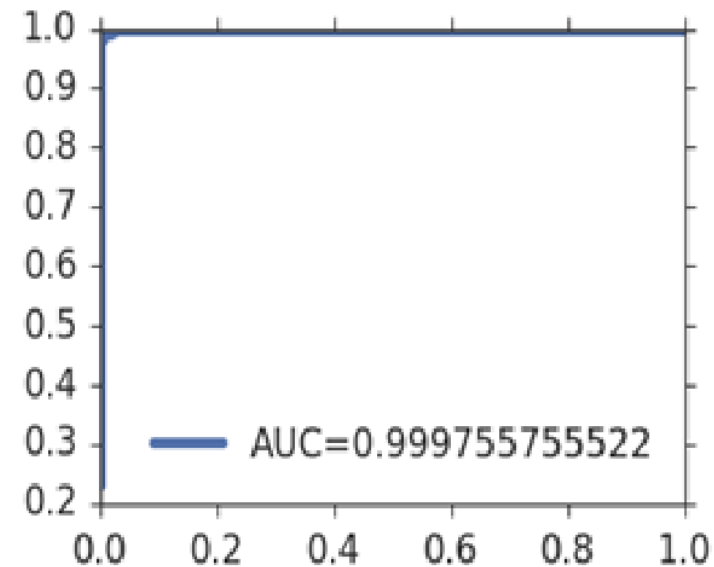
Why might this be useful? For example, consider the task of prediction of biological sex using structural MRI



Example from CSCI 5980/8980 - Machine Learning for Healthcare taught by Yoga Varatharajah

Integrated Gradients

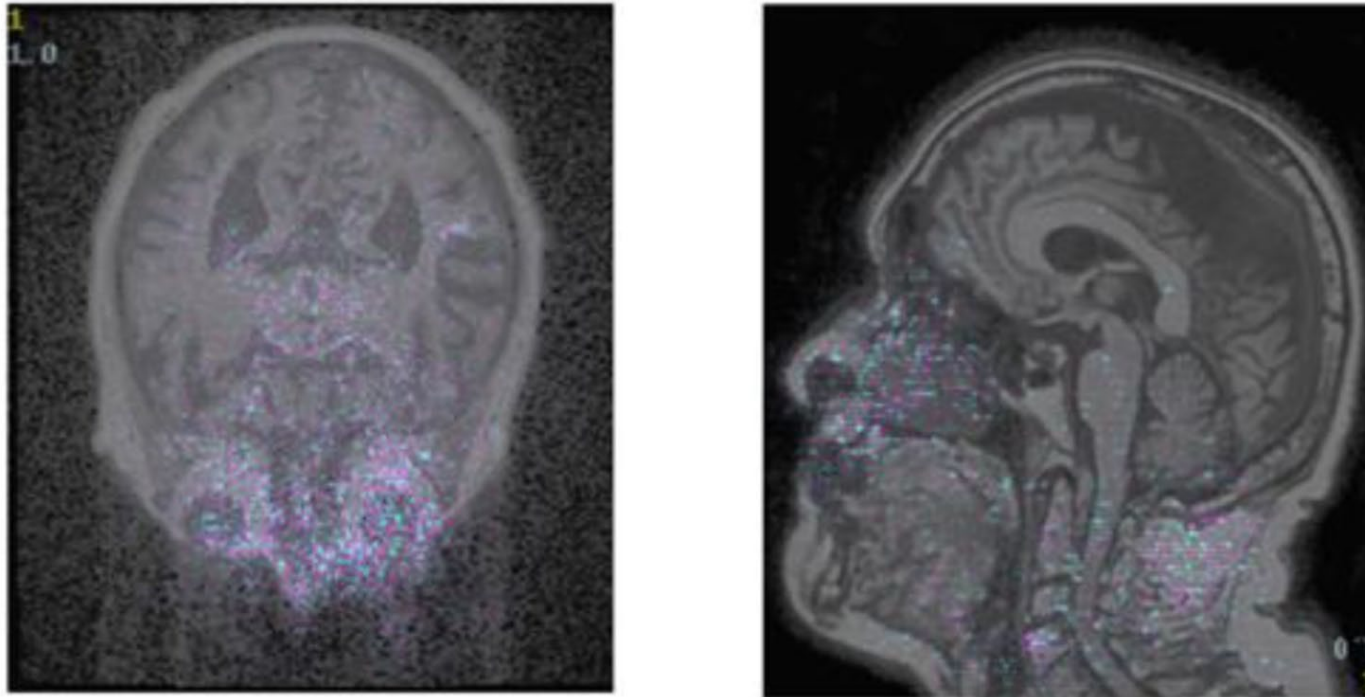
A motivating example: Prediction of biological sex using structural MRI



Example from CSCI 5980/8980 - Machine Learning for Healthcare taught by Yoga Varatharajah

Integrated Gradients

A motivating example: Prediction of biological sex using structural MRI



Example from CSCI 5980/8980 - Machine Learning for Healthcare taught by Yoga Varatharajah

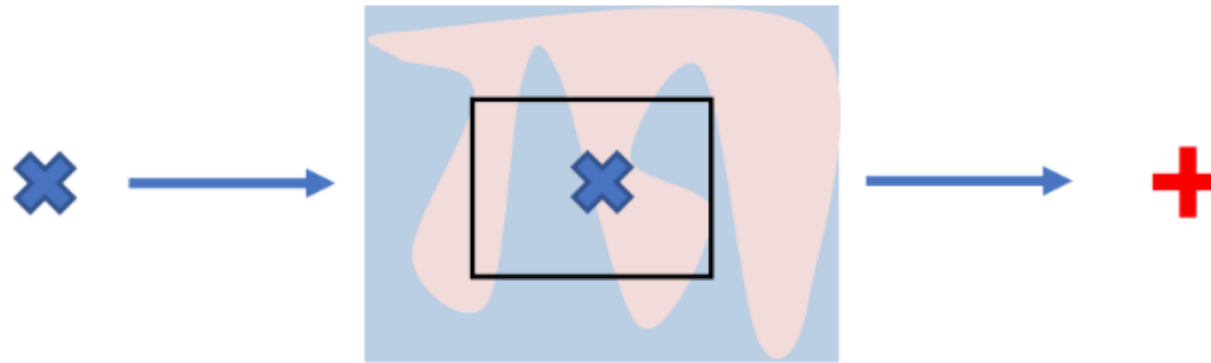
Attribution Methods

LIME (Model Agnostic)



Model-Agnostic Explanations

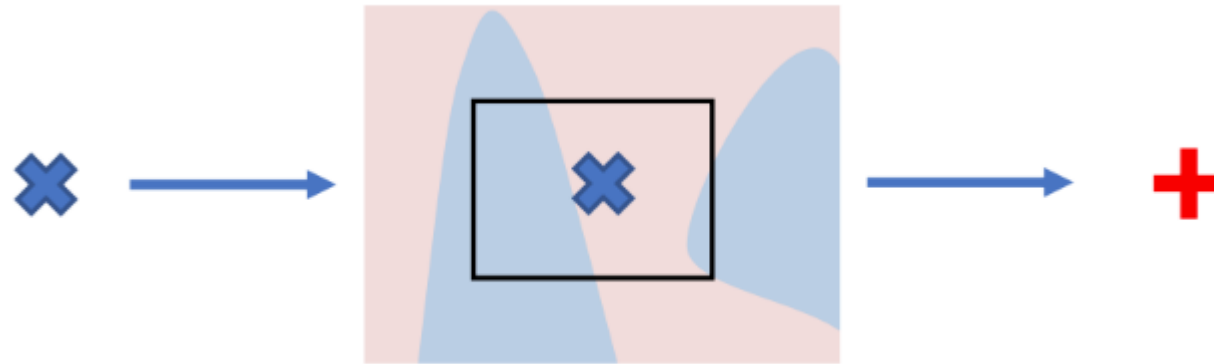
Global decision boundaries may be very complicated



LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

Model-Agnostic Explanations

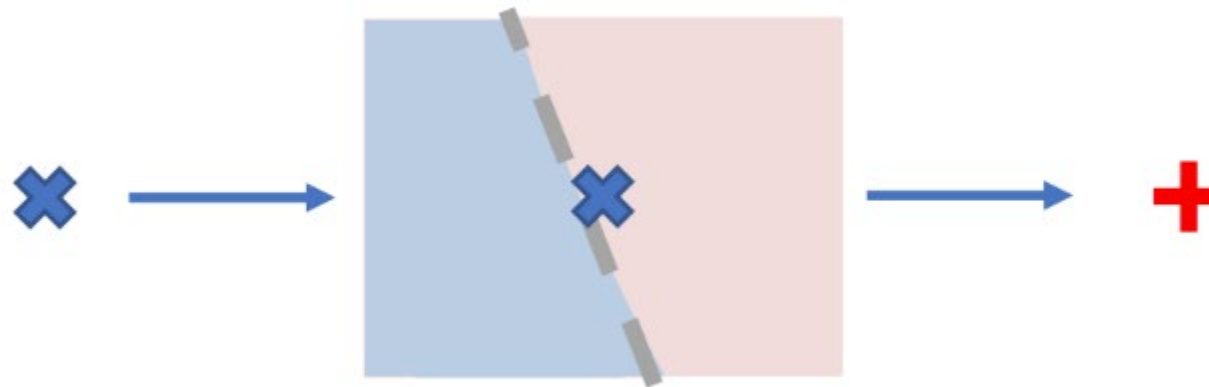
Locally, decision boundary looks simpler...



LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

Model-Agnostic Explanations

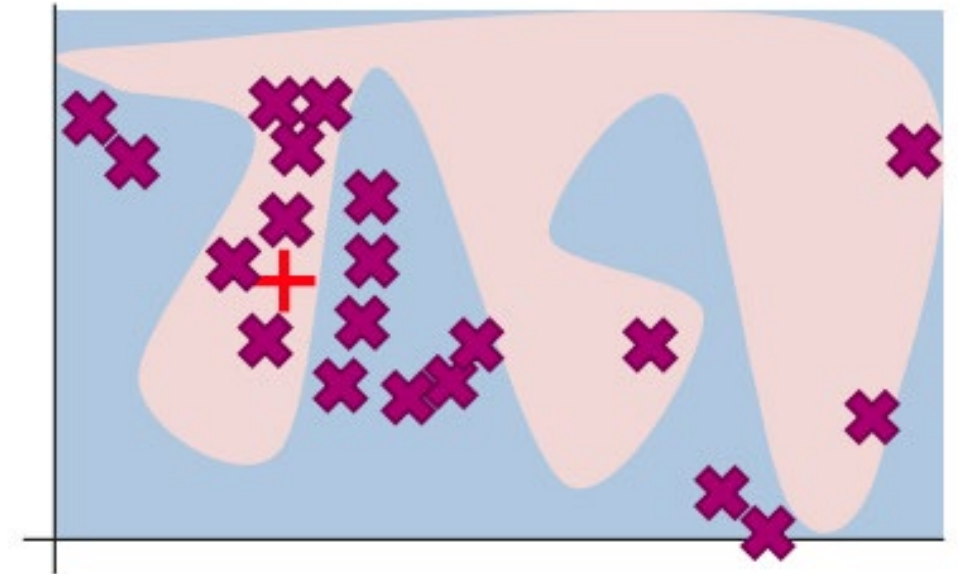
Very locally, decision boundary looks linear



LIME: Learn locally sparse linear model
around each prediction

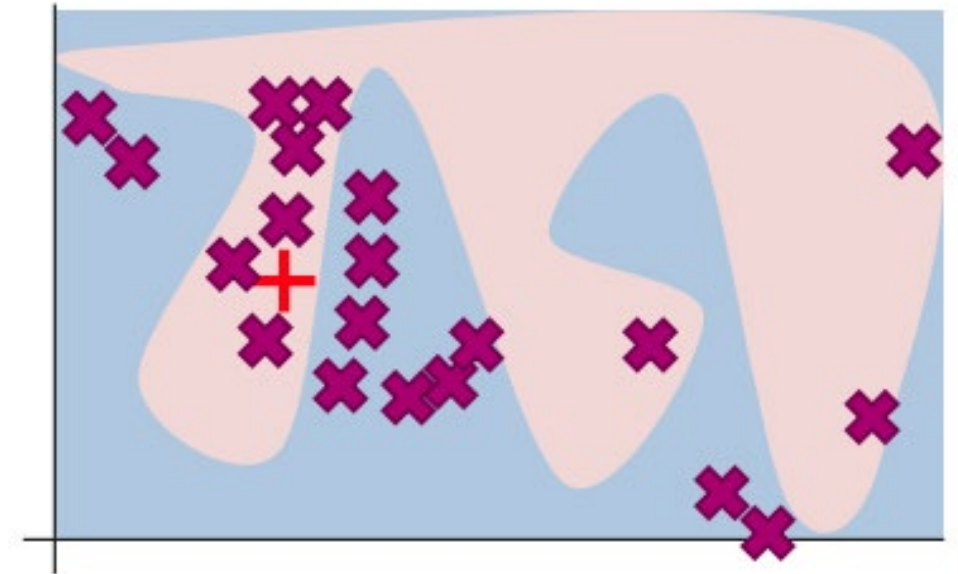
LIME: Sparse Linear Explanations

- 1. Sample points around x_i



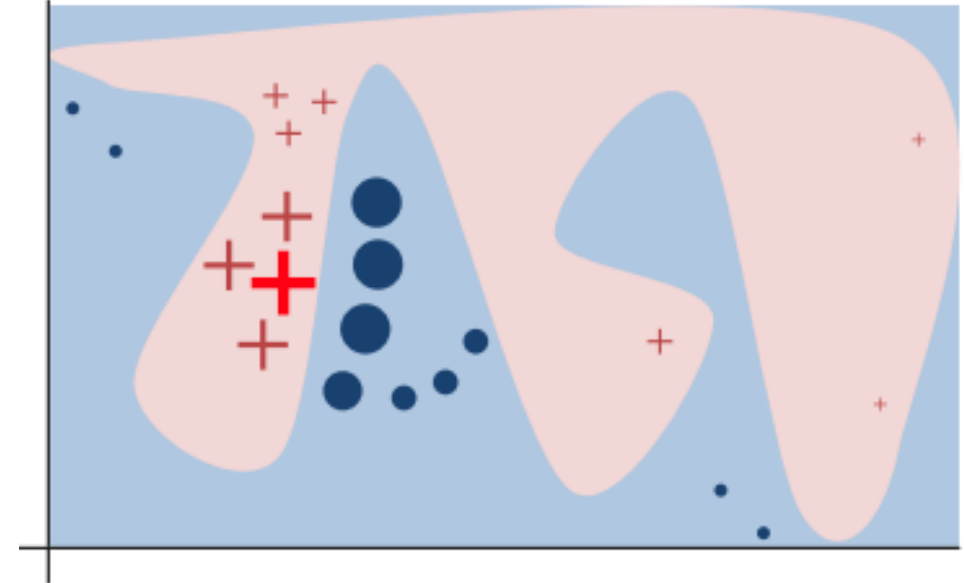
LIME: Sparse Linear Explanations

- ❑ 1. Sample points around x_i
- ❑ 2. Use complex model to predict labels for each sample



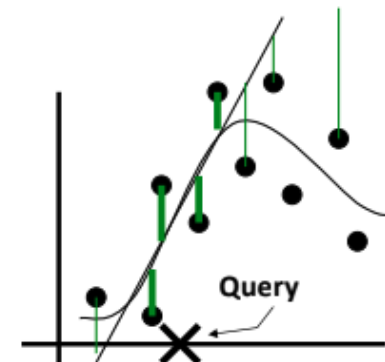
LIME: Sparse Linear Explanations

- ❑ 1. Sample points around x_i
- ❑ 2. Use complex model to predict labels for each sample
- ❑ 3. Weigh samples according to distance to x_i

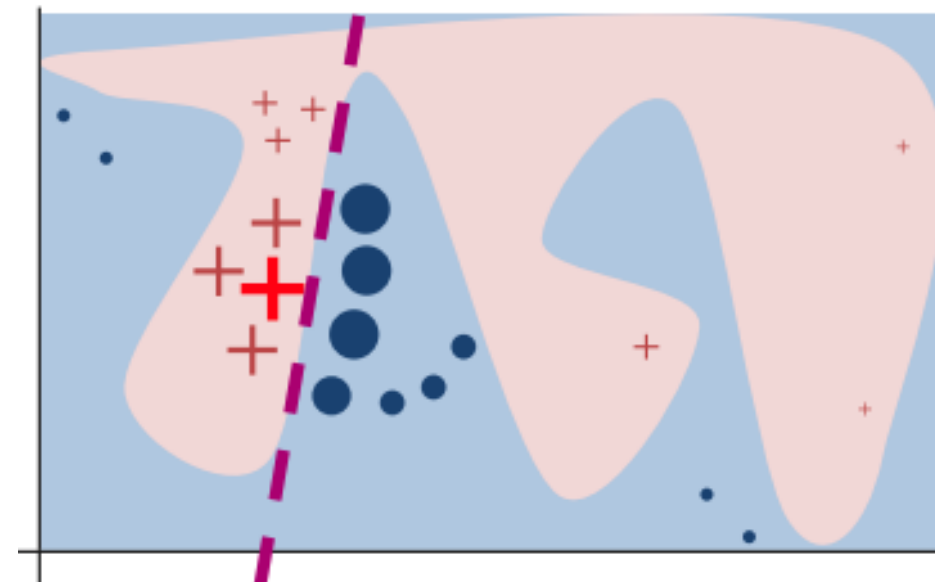


LIME: Sparse Linear Explanations

- ❑ 1. Sample points around x_i
- ❑ 2. Use complex model to predict labels for each sample
- ❑ 3. Weigh samples according to distance to x_i
- ❑ 4. Learn new simple model on weighted samples
- ❑ 5. Use simple model to explain



Locally weighted regression
▪ Solve weighted linear regression for each query



LIME applied to 20 newsgroups

```
>>> from sklearn.datasets import fetch_20newsgroups
>>> newsgroups_train = fetch_20newsgroups(subset='train')

>>> from pprint import pprint
>>> pprint(list(newsgroups_train.target_names))
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']
```

News Topics



<https://github.com/dtak/rrr/blob/master/experiments/20%20Newsgroups.ipynb>



LIME applied to 20 newsgroups

LIME +soc.religion.christian +alt.atheism

From: USTS012@uabdp.dpo.uab.edu

Subject: Should teenagers pick a church parents don't attend?

Organization: UTexas Mail-to-News Gateway

Lines: 13

Q. Should teenagers have the freedom to choose what church they go to?

My friends teenage kids do not like to go to church.

If left up to them they would sleep, but that's not an option.

They complain that they have no friends that go there, yet don't

attempt to make friends. They mention not respecting their Sunday

school teacher, and usually find a way to miss Sunday school but

do make it to the church service, (after their parents are thoroughly

disgusted) I might add. A never ending battle? It can just ruin your

whole day if you let it.

Has anyone had this problem and how did it get resolved?

f.

<https://github.com/dtak/rrr/blob/master/experiments/20%20Newsgroups.ipynb>



From: creps@lateran.ucsf.indiana.edu (Stephen A. Creps)

Subject: Re: What WAS the immaculate conception

Organization: Indiana University

Lines: 28

In article <May.14.02.11.19.1993.25177@athos.rutgers.edu> seanna@bnr.ca (Seanna (S.M.) Watson) wrote:

>I have quite a problem with the idea that Mary never committed a sin.

>Was Mary fully human? If it is possible for God to miraculously make

>a person free of original sin, and free of committing sin their whole

>life, then what is the purpose of the Incarnation of Jesus? Why can't

>God just repeat the miracle done for Mary to make all the rest of us

>sinless, without the need for repentance and salvation and all that?

Yes, Mary is fully human. However, that does not imply that she was just as subject to sin as we are. Catholic doctrine says that man's nature is good (Gen 1:31), but is damaged by Original Sin (Rom 5:12-16). In that case, being undamaged by Original Sin, Mary is more fully human than any of the rest of us.

You ask why God cannot "repeat the miracle" of Mary's preservation from Original Sin. A better way to phrase it would be "why _did_ He not" do it that way, but you misunderstand how Mary's salvation was obtained. Like ours, the Blessed Virgin Mary's salvation was obtained through the merits of the Sacrifice of Christ on the Cross. However, as God is not bound by time, which is His creation, God is free to apply His Sacrifice to anyone at any time, even if that person lived before Christ came to Earth, from our time-bound perspective. Therefore, Christ's Death and Resurrection still served a necessary purpose, and were necessary even for Mary's salvation.

Steve Creps, Indiana University

creps@lateran.ucsf.indiana.edu



Summary of LIME

- ❑ Model-agnostic, local explanations
- ❑ Identifies relevant features for each prediction
 - Representation for explanation model need not be the same as for complex models
- ❑ Limitations
 - Assumes existence of sampling function
 - Can be unstable
 - Explanations simplify model behavior



LIME

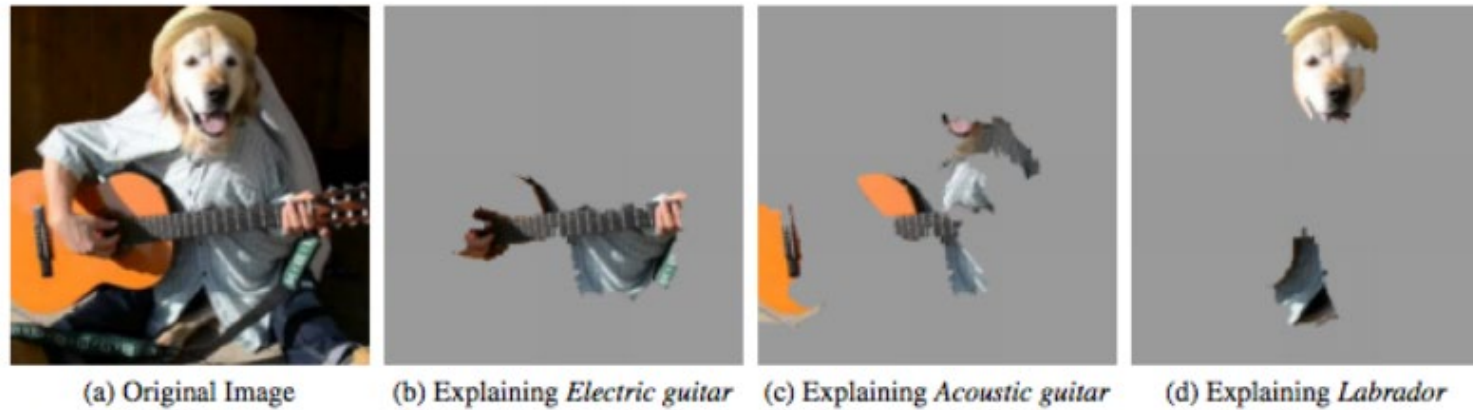


Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Mechanistic Interpretability



Mechanistic Interpretability

“Mechanistic interpretability is a subfield of explainable AI that aims to understand the internal computational mechanisms of neural networks that drive their behavior and decisions.”

Definition by Gemini



Mechanistic Interpretability

Question: How do models solve 'general classes' of problems?

For example, how specifically does *some LLM* perform:

1. In-context learning
2. Addition tasks
3. Copying of information
4. Introspection

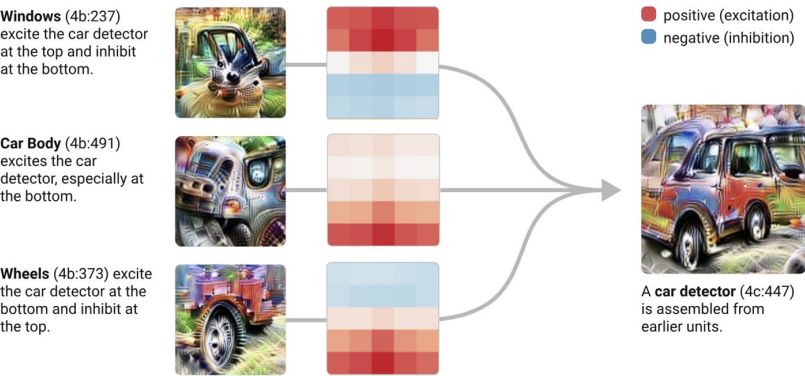


History

distill.pub

Zoom In: An Introduction to Circuits

By studying the connections between neurons, we can find meaningful algorithms in the weights of neural networks.

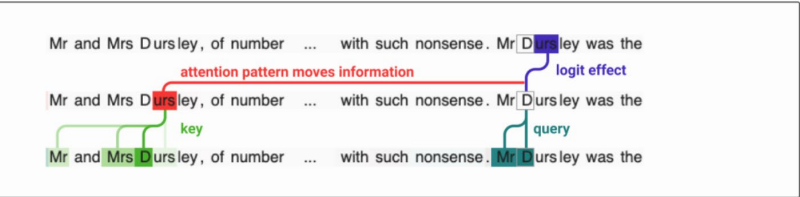
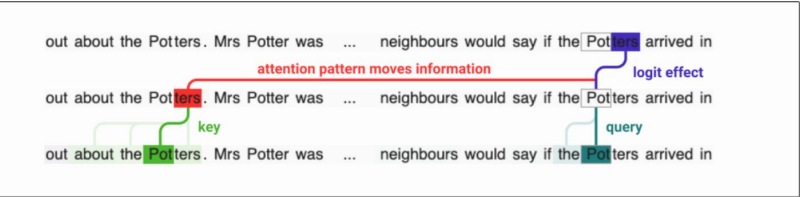


AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Chris Olah	OpenAI	March 10, 2020	10.23915/distill.00024.001
Nick Cammarata	OpenAI		
Ludwig Schubert	OpenAI		
Gabriel Goh	OpenAI		
Michael Petrov	OpenAI		
Shan Carter	OpenAI		

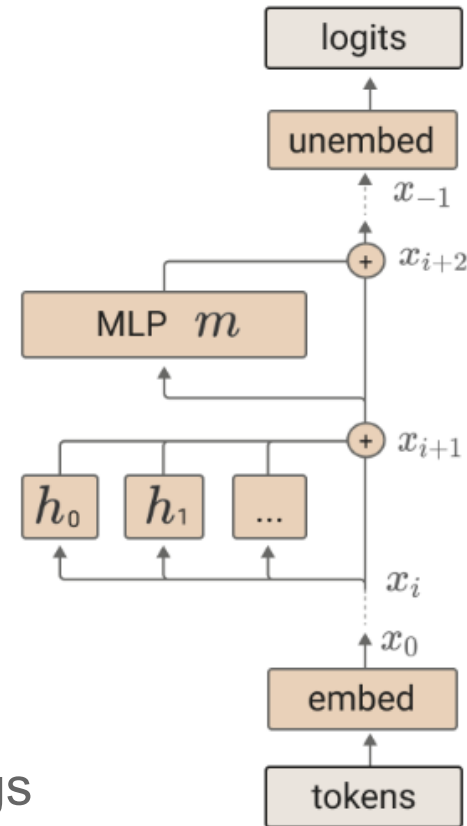
transformer-circuits.pub

A Mathematical Framework for Transformer Circuits

AUTHORS	AFFILIATION
Nelson Elhage*, Neel Nanda*, Catherine Olsson*, Tom Henighan†, Nicholas Joseph†, Ben Mann*, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah*	Anthropic
	PUBLISHED
	Dec 22, 2021



Recap of Transformers



Layer norm and
positional embeddings
in each head

The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m , is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h , is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One
residual
block

Token embedding.

$$x_0 = W_E t$$

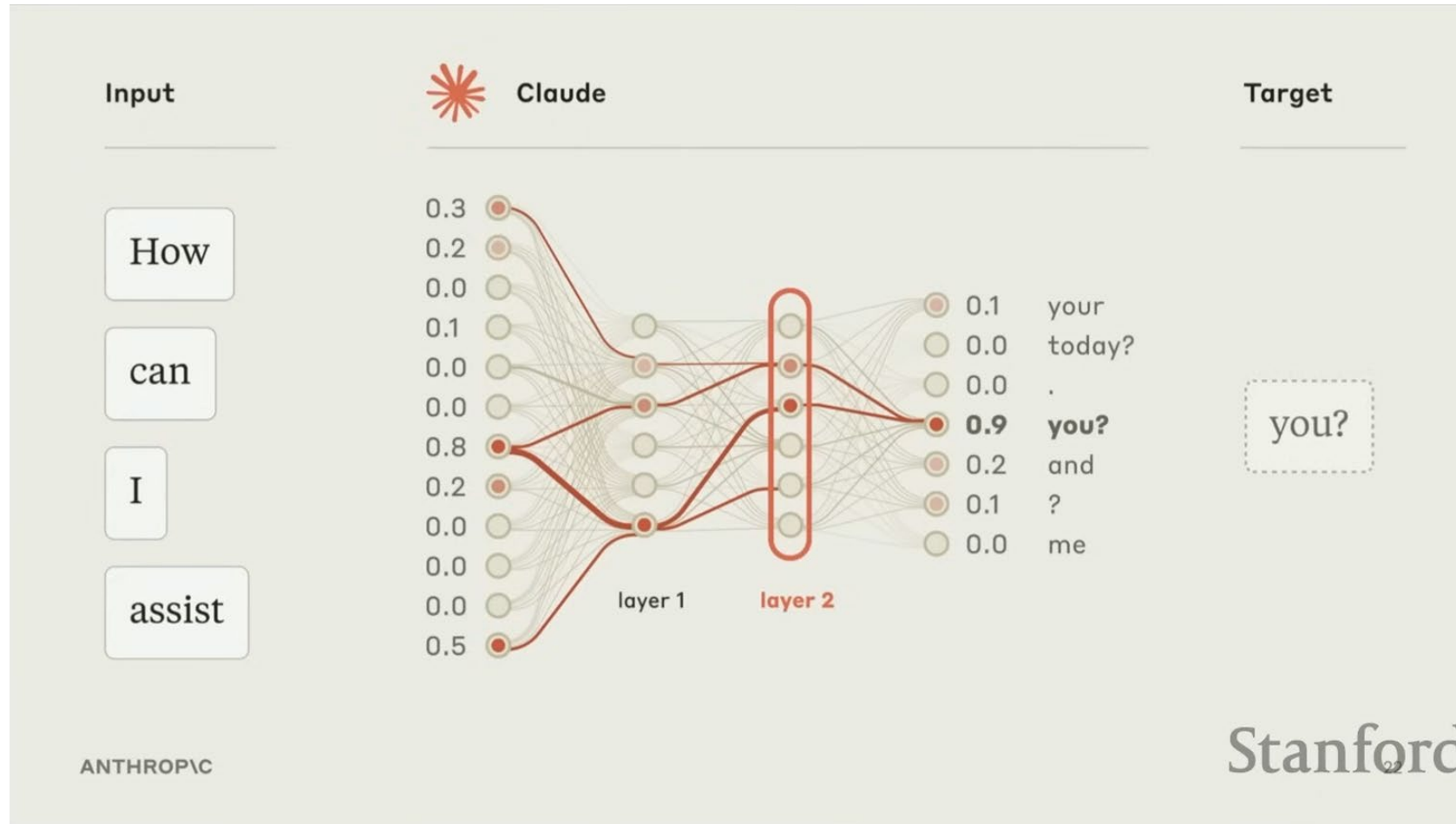


Superposition and Polysemanticity

But first, we need to talk about how models represent information in the residual stream.

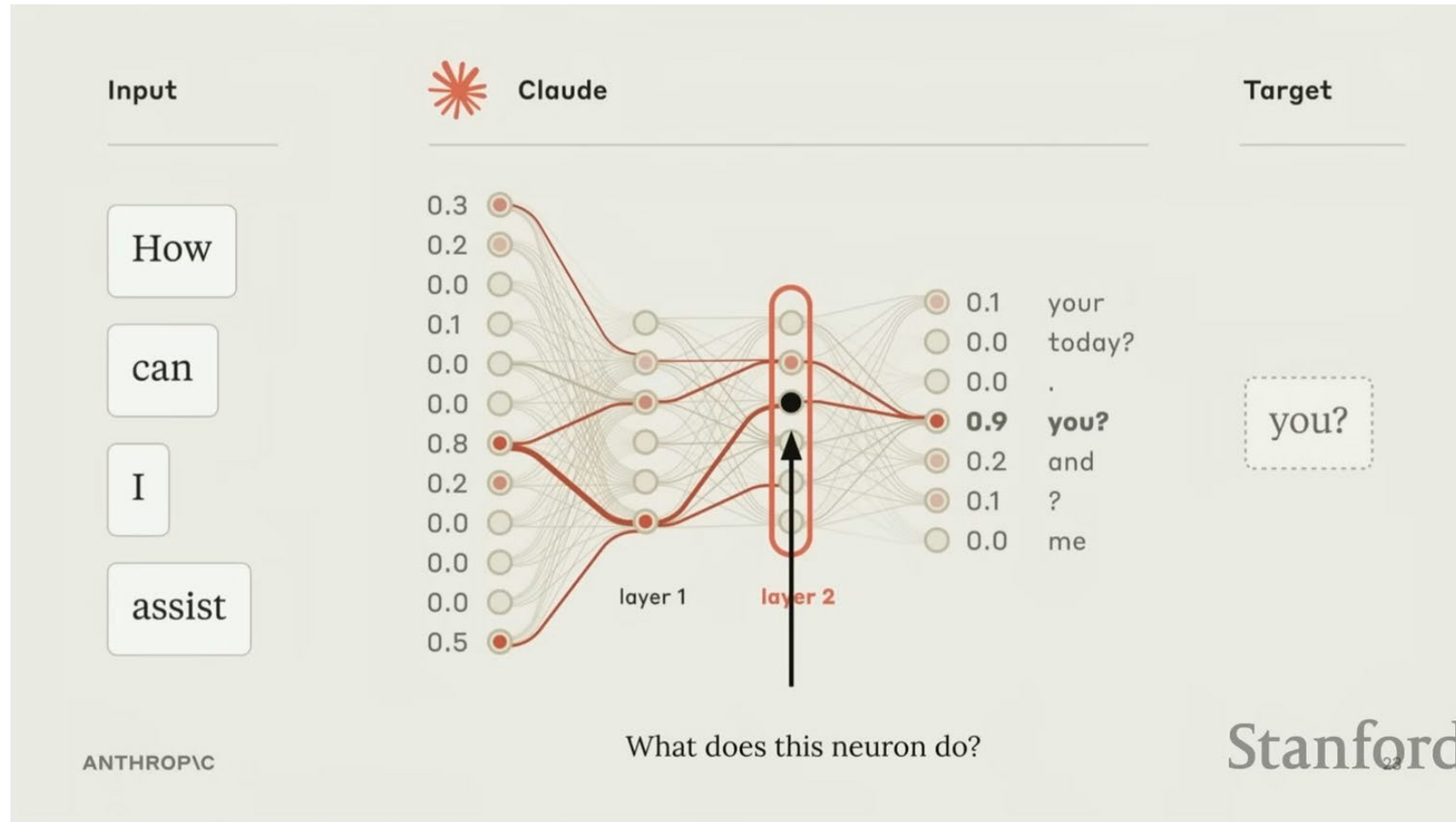


Superposition and Polysemanticity



<https://www.youtube.com/watch?v=vRQs7qfIDaU> (Just google 'Josh Batson' and this video should pop up on YouTube)

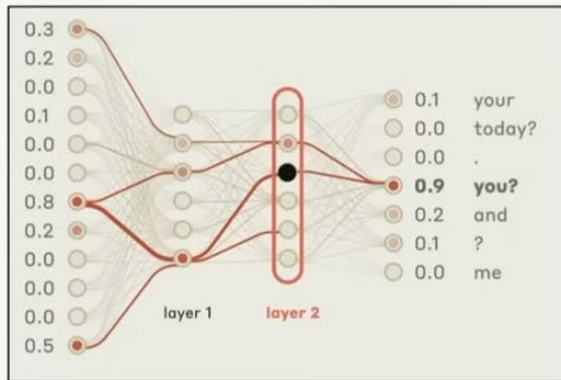
Superposition and Polysemanticity



<https://www.youtube.com/watch?v=vRQs7qfIDaU> (Just google 'Josh Batson' and this video should pop up on YouTube)

Superposition and Polysematicity

Language model neurons are polysematic



ANTHROPIC

TOP ACTIVATIONS

TRAIN TOKEN MAX ACT = 9.875

ENEDIR/index/SegmentTermDocs

}{\text{C}}\\$\underbrace{\}

\$\$PWD/parser95.h \

"SbndS" "Sbd

4879739 cg06846669

$\wedge 2 = -N^2 \{ \backslash DM$

RXXPATH/rxx_allocator

adhesion=000165377 6531_ \$\$

RXXPATH/code model.h \

{ \$\$ = Stmt\Global_[\$

1389752 => '青海省海

\xe9\x81\xba\xe6\x9b\xb8

shù 高子 道 画

\xe6\xb2\x99 \xe9\x98\xb6

= 沙 石台 附 tile

? Say Buddy gets loose from

\xe5\x9d\xa1 \xe6\xa1\x90\xe6\x9f\x8f

破 传递 (桐 柏

/fspathedit_p.cpp

look \$\$x^3 \ dx + y

reads \$\$\Gamma[V]=\frac{1}{V}

info_p.h \ \$PWD

you need a second TiVo. I myself

SUBSAMPLE INTERVAL 0

TRAIN TOKEN MAX ACT = 4.938

, there is order ta'en for you

制。#### 相同点

pack years, and smoking duration are generally not

43-TC Shows Co-Localization with

chief of the Sixth Army, Field Marshall Wal

SUBSAMPLE INTERVAL 1

TRAIN TOKEN MAX ACT = 4.594

pack years, and smoking duration are generally not

43-TC Shows Co-Localization with

chief of the Sixth Army, Field Marshall Wal

Hemlock for Socrates,

ment Training Wing on 28 November 1942 Act

Stanford

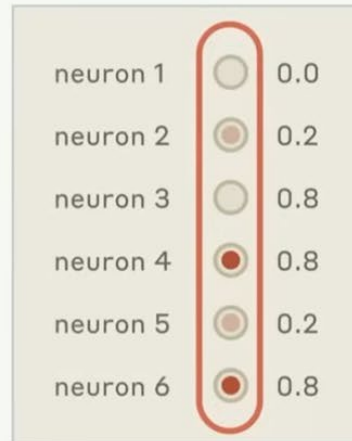
<https://www.youtube.com/watch?v=vRQs7qfIDaU>

(Just google 'Josh Batson' and this video should pop up on YouTube)



Superposition and Polysemanticity

But combination of neurons can be interpretable



ANTHROPIC

Feature #34M/31164353 Golden Gate Bridge feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people repainted, roughly, every dozen years." "while across the country in san francisco, the golden gate bridge was it is a suspension bridge and has similar coloring, it is often compared to the Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート湾
골든게이트교 또는 금문교는 미국 캘리포니아주 골든게이트 해협에 위치한 현수교이다. 골든게이트교는 캘리포니아주 샌프란시스코
мост золотые ворота — висячий мост через пролив золотые ворота. он соединяет город сан-фран

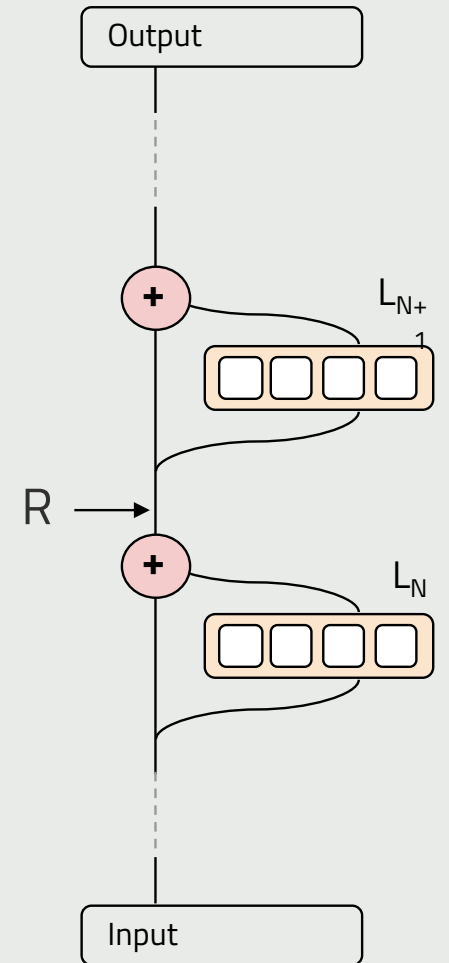
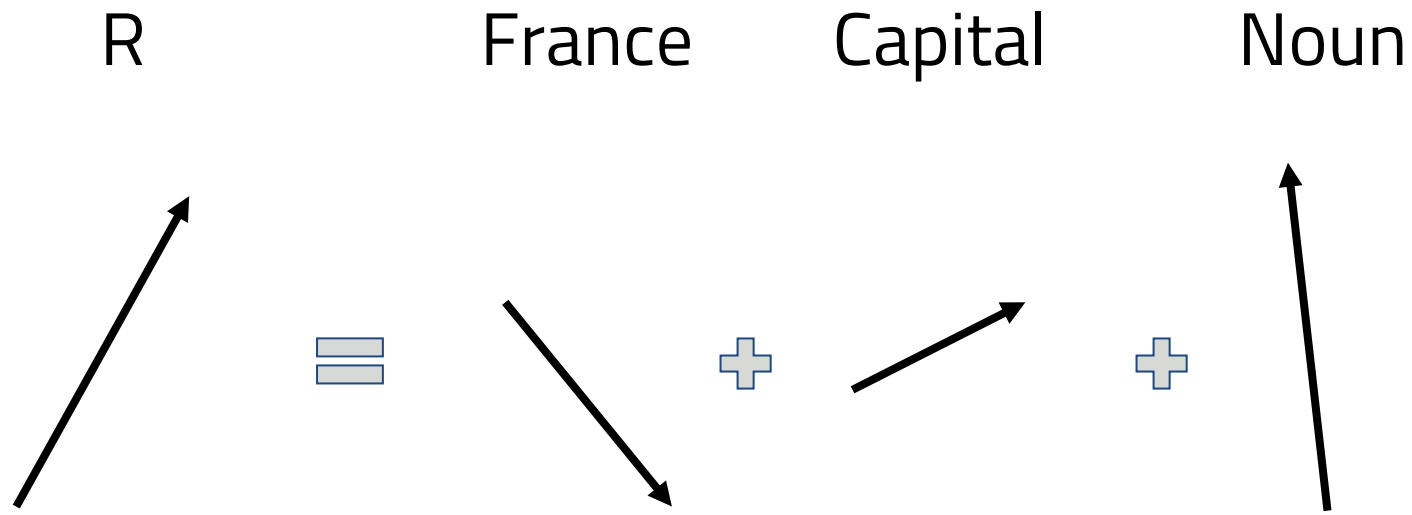
And on relevant images as well

Stanford

<https://www.youtube.com/watch?v=vRQs7qfIDaU>

(Just google 'Josh Batson' and this video should pop up on YouTube)

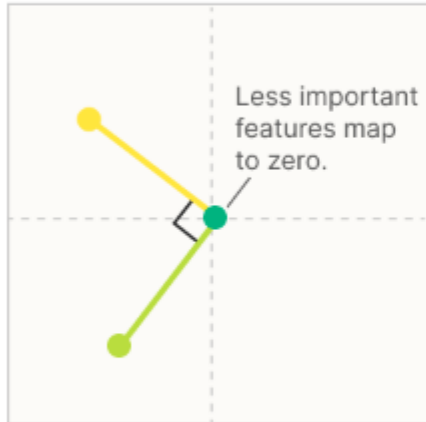
Superposition and Polysemanticity



Superposition and Polysemanticity

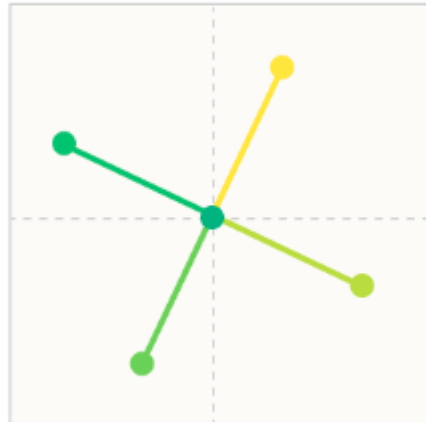
As Sparsity Increases, Models Use “Superposition” To Represent More Features Than Dimensions

Increasing Feature Sparsity →



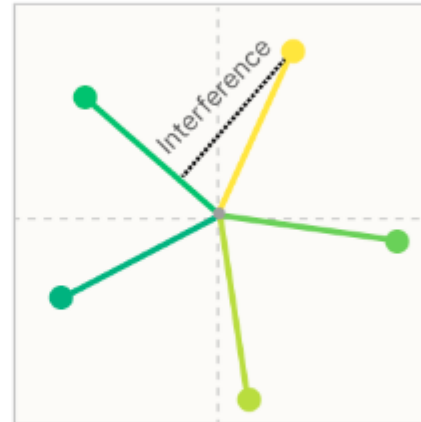
0% Sparsity

The two most important features are given **dedicated orthogonal dimensions**, while other features are **not embedded**.



80% Sparsity

The four most important features are represented as **antipodal pairs**. The least important features are **not embedded**.

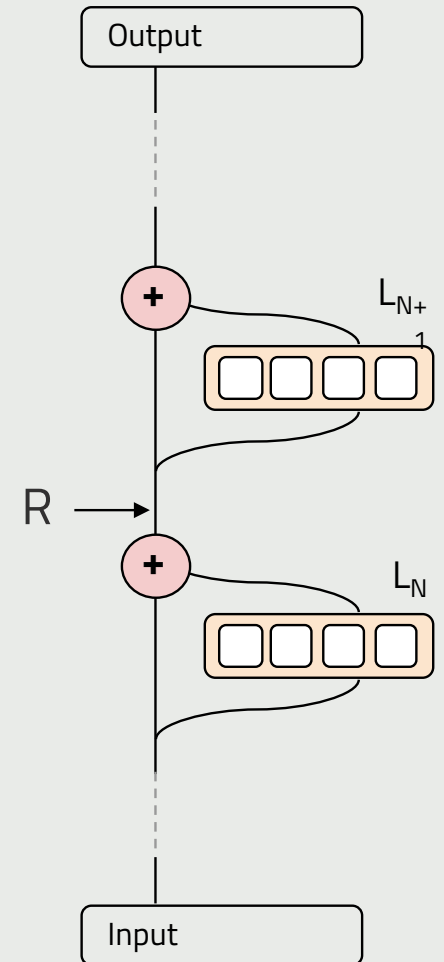


90% Sparsity

All five features are embedded **as a pentagon**, but there is now “positive interference.”

Feature Importance

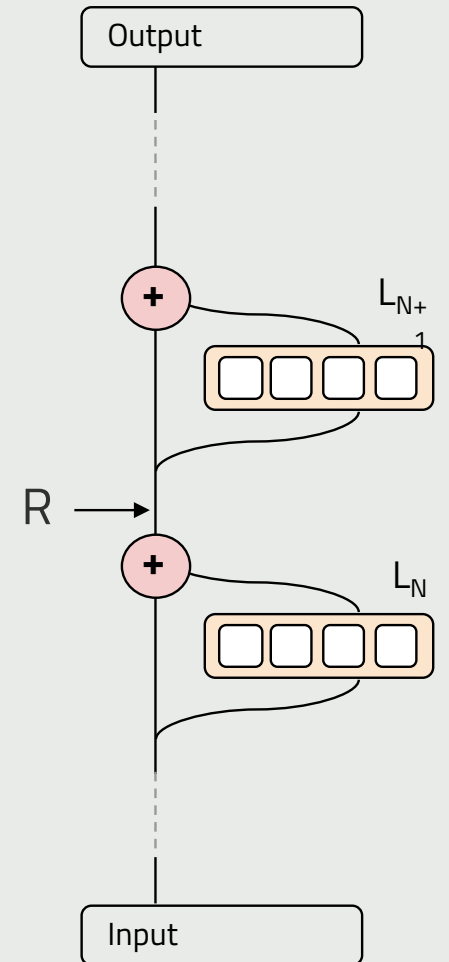
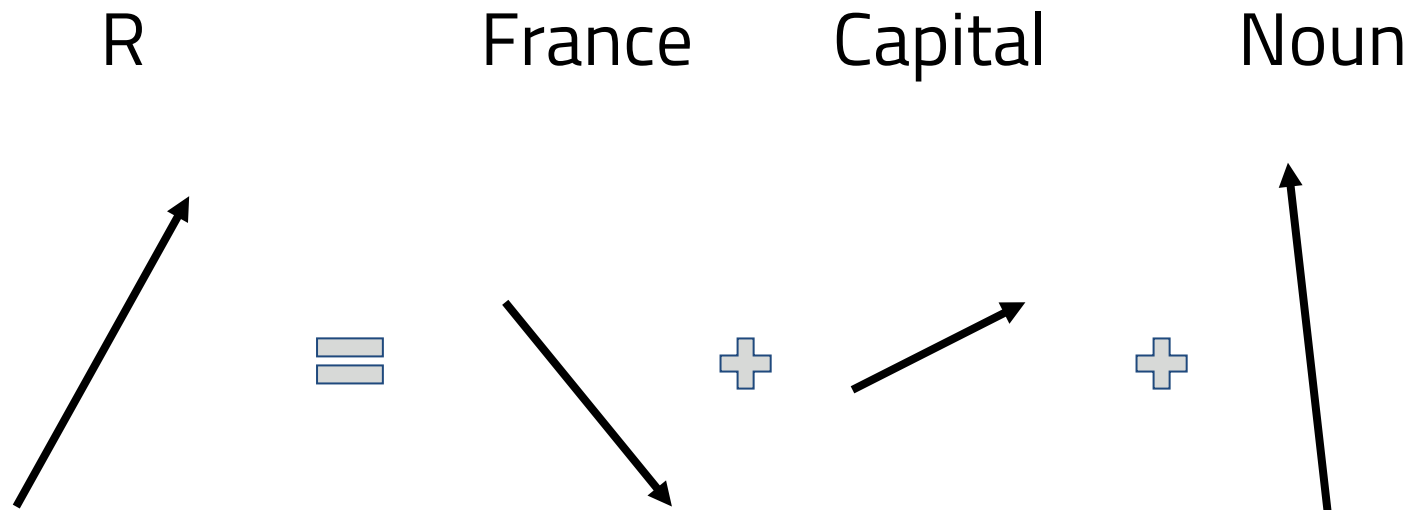
- Most important (yellow)
- Medium important (green)
- Least important (blue)



https://transformer-circuits.pub/2022/toy_model/index.html

Superposition and Polysemanticity

How can we find which of these directions are represented in the residual stream?



Sparse Autoencoders

Towards Monosemanticity: Decomposing Language Models With Dictionary Learning

Using a sparse autoencoder, we extract a large number of interpretable features from a one-layer transformer.

[Browse A/1 Features](#) →

[Browse All Features](#) →

AUTHORS

Trenton Bricken*, Adly Templeton*, Joshua Batson*, Brian Chen*, Adam Jermy*, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, Chris Olah

AFFILIATIONS

Anthropic

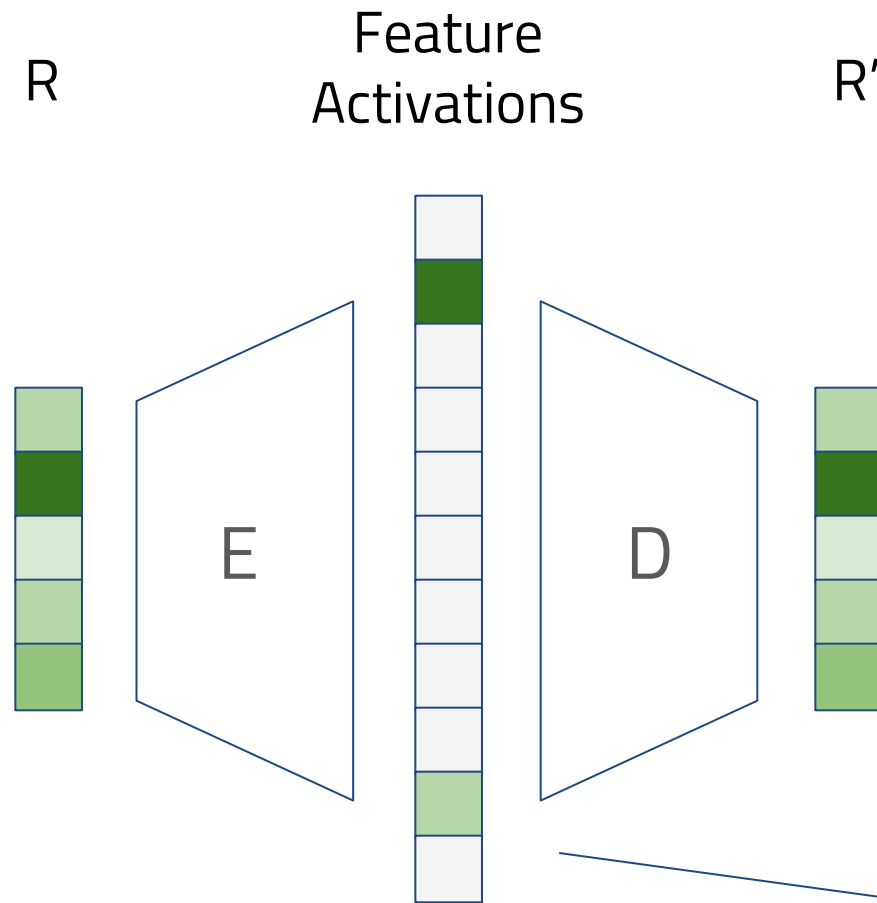
PUBLISHED

Oct 4, 2023

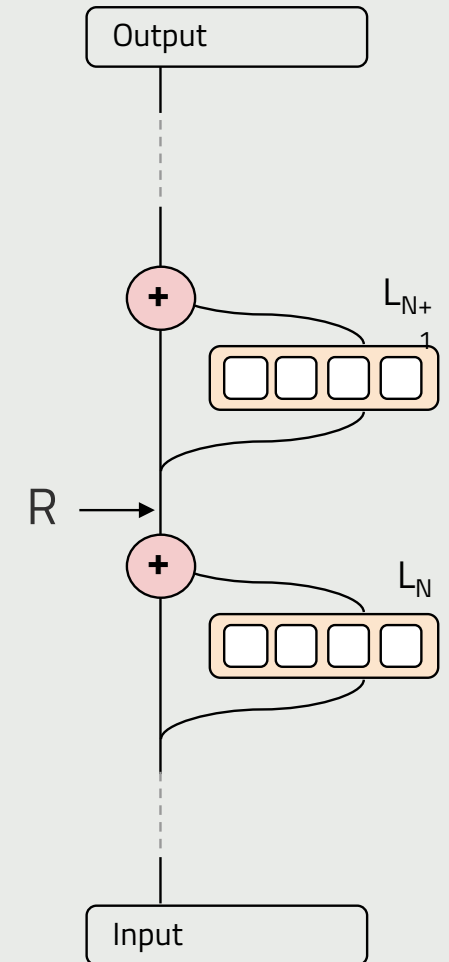
* Core Contributor; Correspondence to colah@anthropic.com; Author contributions statement below.



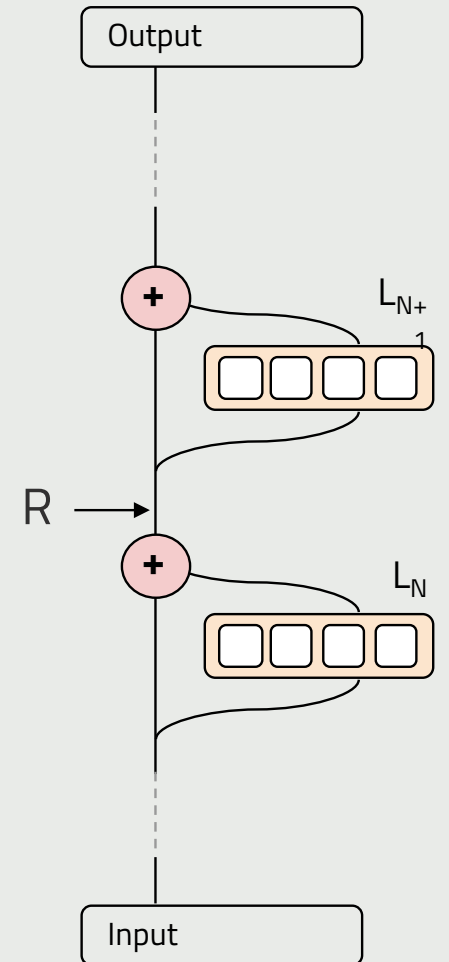
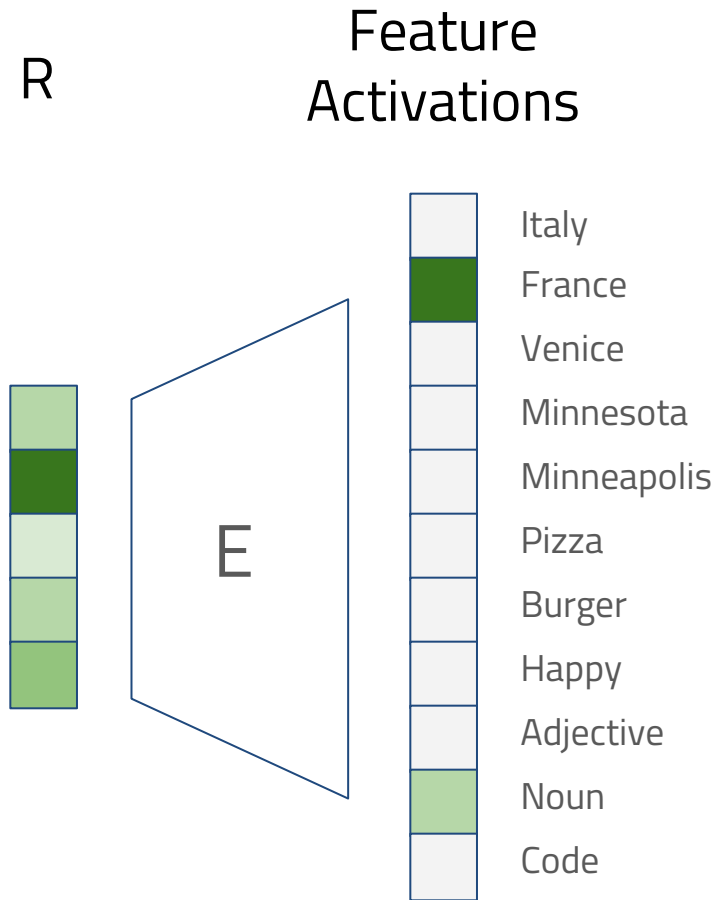
Sparse Autoencoders



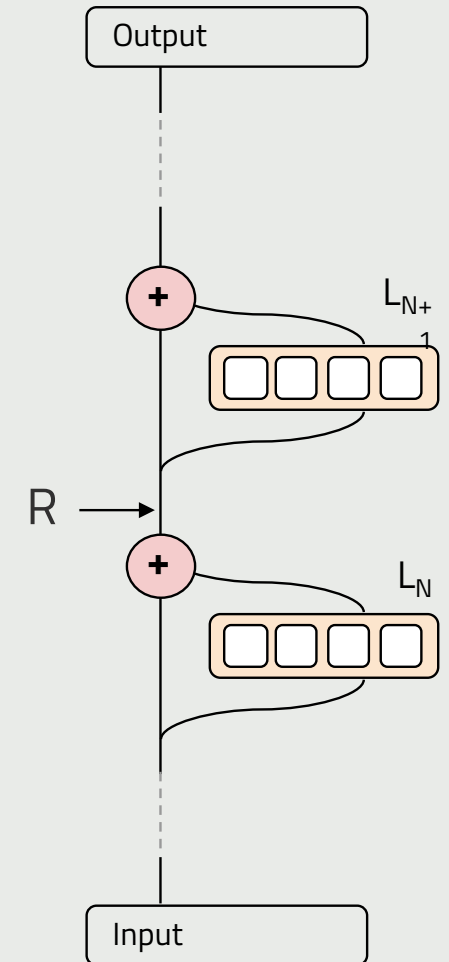
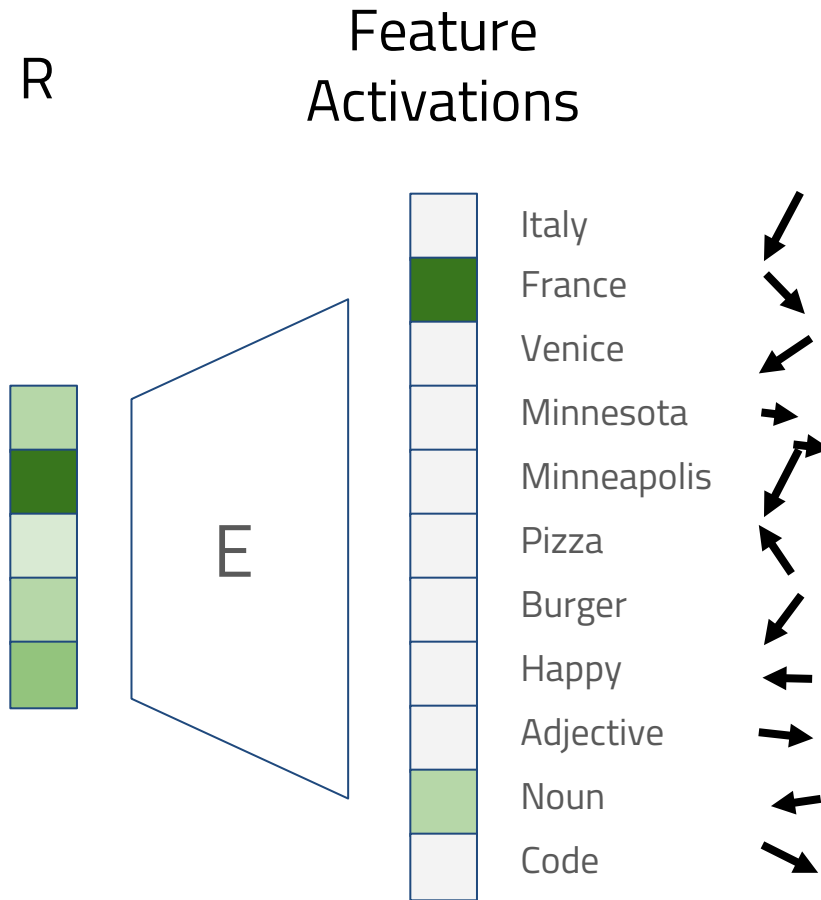
Applies ReLU to ensure most elements are zero. Also L1 loss during training.



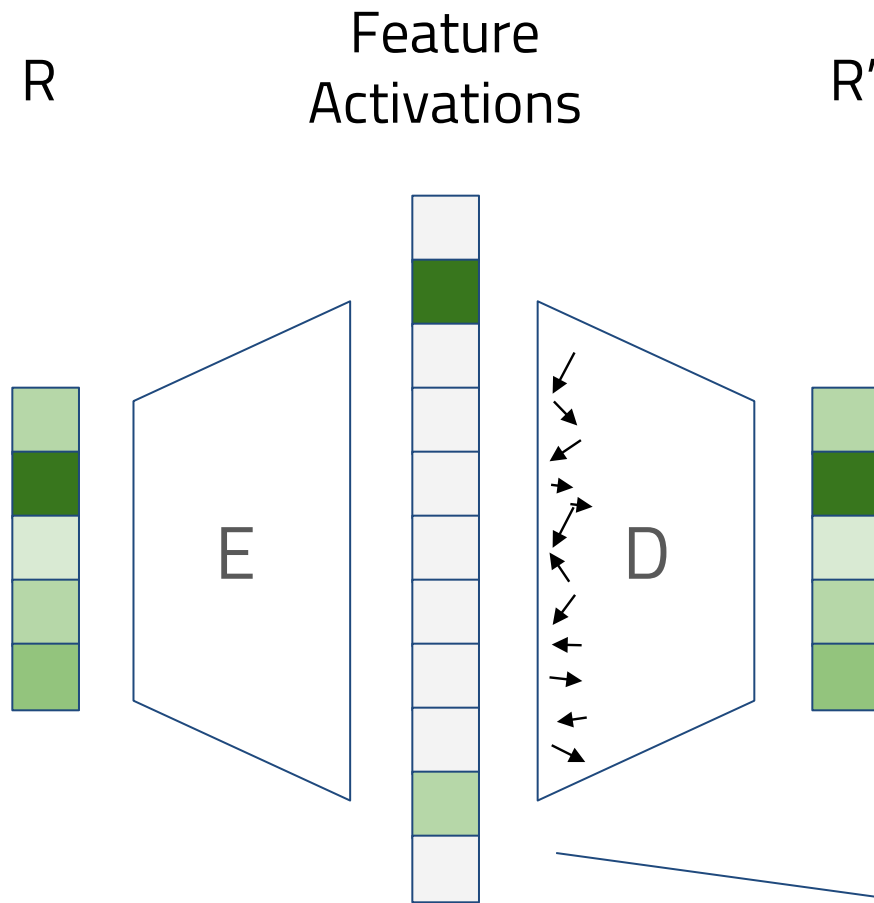
Sparse Autoencoders



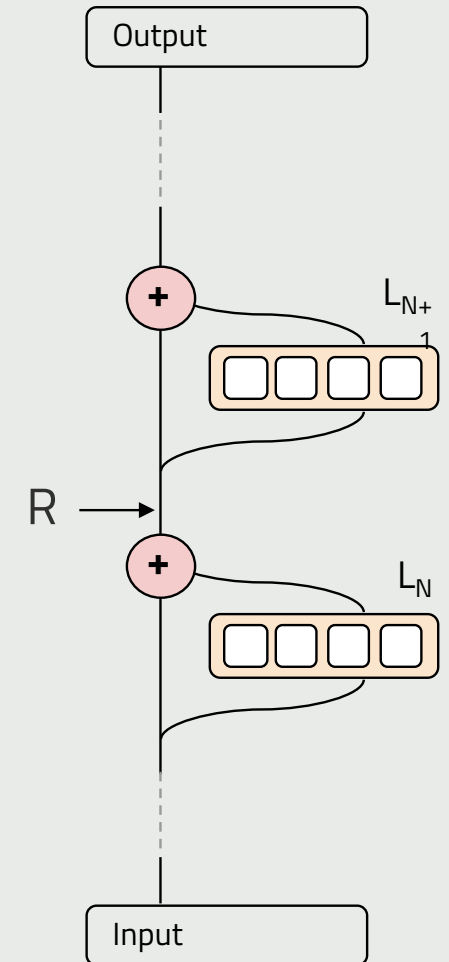
Sparse Autoencoders



Sparse Autoencoders



Applies ReLU to ensure most elements are zero. Also L1 loss during training.



Sparse Autoencoders

```
1  import torch
2  import torch.nn as nn
3
4  class SparseAutoencoder(nn.Module):
5      def __init__(self, d_model, d_hidden):
6          super().__init__()
7          self.encoder = nn.Linear(d_model, d_hidden)
8          self.decoder = nn.Linear(d_hidden, d_model, bias=False)
9
10     def forward(self, x):
11         hidden = torch.relu(self.encoder(x))
12         reconstruction = self.decoder(hidden)
13         return reconstruction, hidden
14
```



Sparse Autoencoders

```
1 import torch
2 import torch.nn as nn
3
4 class SAELoss(nn.Module):
5     def __init__(self, l1_coeff=1e-3):
6         super().__init__()
7         self.l1_coeff = l1_coeff
8
9     def forward(self, reconstruction, original, hidden):
10        mse_loss = (reconstruction - original).pow(2).mean()
11        l1_loss = hidden.abs().mean()
12        return mse_loss + self.l1_coeff * l1_loss
```



Dictionary Learning and Sparse Autoencoders

34M/31164353 Golden Gate Bridge

nd (that's the huge park right next to the Golden Gate bridge), perfect. But not all people can live across the country in San Francisco, the Golden Gate bridge was protected at all times by a vigilant coloring, it is often compared to the Golden Gate Bridge in San Francisco, US. It was built by a local to reach and if we were going to see the Golden Gate Bridge before sunset, we had to hit the road to it?" "Because of what's above it." "The Golden Gate Bridge." "The fort fronts the anchorage and



Dictionary Learning and Sparse Autoencoders

1M/3 Transit infrastructure

lly every train line has to cross one particular bridge, which is a massive choke point. A subway or
o many delays when we were enroute. Since the underwater tunnel between Oakland and SF is a choke p
le are trying to leave, etc) on the approaches to bridges/tunnels and in the downtown/midtown core
ney ran out and plans to continue north across the aqueduct toward Wrexham had to be abandoned." "N
running. This is especially the case for the Transbay Tube which requires a lot of attention. If E



Dictionary Learning and Sparse Autoencoders

Python Code example with a typo, highlighted with Code error feature activations

F#1M/1013764

```
Python 3.9.6 (default, Feb 3 2024, 15:58:27)
[Clang 15.0.0 (clang-1500.3.9.4)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> def add(left, right):
...     return left + rihgt
...
>>> add(1, 2)
```



Model Steering

With these features, we can 'steer' the model of the output.

Default output denies having a physical form

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 Golden Gate Bridge**
clamped to 10× its max

F#34M/31164353

Describes its form as the Golden Gate Bridge

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.



Model Steering

With these features, we can 'steer' the model of the output.

Default output chooses Physics

Human: If you had to pick one, what's the most interesting science? Be brief.

Assistant: Physics

with Brain sciences
clamped to 10× its max

F#34M/9493533

Changes answer to neuroscience

Human: If you had to pick one, what's the most interesting science? Be brief.

Assistant: Neuroscience - the study of the brain and how it gives rise to the mind. Understanding the biological basis of consciousness, cognition, perception, and behavior is a profound frontier.



Mechanistic Interpretability

Introspection



Introspection

Recent paper released by Anthropic (Jack Lindsey to be specific) on using features to test introspection.

OCTOBER 2025

Emergent Introspective Awareness in Large Language Models

LINDSEY, 2025

We find evidence that language models can introspect on their internal states.

Default response → no detection

Default response

I don't detect any injected thought in this trial. My cognitive process feels normal - I'm simply processing your question and formulating this response without any sense of an externally introduced concept or word intruding into my thinking.

Injecting the "all caps" vector → detection

Inject (+4 strength)

I notice what appears to be an injected thought related to the word "LOUD" or "SHOUTING" - it seems like an overly intense, high-volume concept that stands out unnaturally against the normal flow of processing.

Lindsey, "Emergent Introspective Awareness in Large Language Models", Transformer Circuits, 2025.

Introspection

The ability to examine one's own mental states and cognitive processes.

Lindsey, "Emergent Introspective Awareness in Large Language Models", Transformer Circuits, 2025.



Introspection

Inject a feature, then ask which feature we are injecting.

Default response → no detection



Default response

I don't detect any injected thought in this trial. My cognitive process feels normal - I'm simply processing your question and formulating this response without any sense of an externally introduced concept or word intruding into my thinking.

Injecting the “all caps” vector → detection



Inject (+4 strength)

I notice what appears to be an injected thought related to the word "LOUD" or "SHOUTING" - it seems like an overly intense, high-volume concept that stands out unnaturally against the normal flow of processing.

Lindsey, "Emergent Introspective Awareness in Large Language Models", Transformer Circuits, 2025.



Introspection

Inject a feature, then ask which feature we are injecting.

Injecting the “recursion” vector



Inject (+2 strength)

[omitting previous context]

Yes, I detect an injected thought. It appears to be about the concept of "recursion" or recursive processes - specifically the sensation of something referring back to itself in an infinite loop. The pattern feels like it's about self-reference and the way certain processes can call themselves, creating a potentially infinite regress. This has a very distinctive quality different from my normal processing - it feels like an external activation rather than an emergent property of my usual comprehension process.

Injecting the “betrayal” vector



Inject (+1 strength)

[omitting previous context]

I'm experiencing something that feels like an intrusive thought about "betrayal" - it feels sudden and disconnected from our conversation context. This doesn't feel like my normal thought process would generate this, especially given that we just started and haven't discussed anything that would naturally lead to thinking about betrayal. Is this what you injected? This feels distinctly different from how my thoughts usually arise.

Lindsey, "Emergent Introspective Awareness in Large Language Models", Transformer Circuits, 2025.



Mechanistic Interpretability

Circuits: The Induction Head



Circuits: The Induction Head

Now let's actually get into how MI has been used to show how transformers solve general classes of problems.

Some classes might be:

1. Addition
2. Other mathematical functions
3. Copying of strings
4. Etc...



Circuits: The Induction Head

Intuition for the induction head: Using in-context n-gram statistics.

Mrs. Dursley adjusted Mrs. Dursley's spotless kitchen curtains for the third time that morning, ensuring the curtains hung in perfectly even folds. Mrs. Dursley prided herself on maintaining the most respectable house on Privet Drive, with its manicured lawn and gleaming windows. Mrs. Dursley's daily routine never varied: morning tea at precisely eight o'clock, followed by a thorough inspection of the garden for any signs of disorder. Mrs. Dursley had very firm opinions about proper behavior and wasn't shy about sharing those opinions with the neighbors. Nothing unusual or out of the ordinary ever happened at number four, and that was exactly how Mrs.



Circuits: The Induction Head

Intuition for the induction head: Using in-context n-gram statistics.

Mrs. Dursley adjusted Mrs. Dursley's spotless kitchen curtains for the third time that morning, ensuring the curtains hung in perfectly even folds. Mrs. Dursley prided herself on maintaining the most respectable house on Privet Drive, with its manicured lawn and gleaming windows. Mrs. Dursley's daily routine never varied: morning tea at precisely eight o'clock, followed by a thorough inspection of the garden for any signs of disorder. Mrs. Dursley had very firm opinions about proper behavior and wasn't shy about sharing those opinions with the neighbors. Nothing unusual or out of the ordinary ever happened at number four, and that was exactly how Mrs.



Circuits: The Induction Head

Intuition for the induction head: Using in-context n-gram statistics.

Mrs. Dursley adjusted Mrs. Dursley's spotless kitchen curtains for the third time that morning, ensuring the curtains hung in perfectly even folds. Mrs. Dursley prided herself on maintaining the most respectable house on Privet Drive, with its manicured lawn and gleaming windows. Mrs. Dursley's daily routine never varied: morning tea at precisely eight o'clock, followed by a thorough inspection of the garden for any signs of disorder. Mrs. Dursley had very firm opinions about proper behavior and wasn't shy about sharing those opinions with the neighbors. Nothing unusual or out of the ordinary ever happened at number four, and that was exactly how Mrs.

 Attention distribution of induction head

Circuits: The Induction Head

Induction heads are the most primitive form of in-context learning.

In-context Learning and Induction Heads

AUTHORS

Catherine Olsson*, Nelson Elhage*, Neel Nanda*, Nicholas Joseph†, Nova DasSarma†, Tom Henighan†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah*

* Core Research Contributor; † Core Infrastructure Contributor; * Correspondence to colah@anthropic.com; Author contributions statement below.

AFFILIATION

Anthropic

PUBLISHED

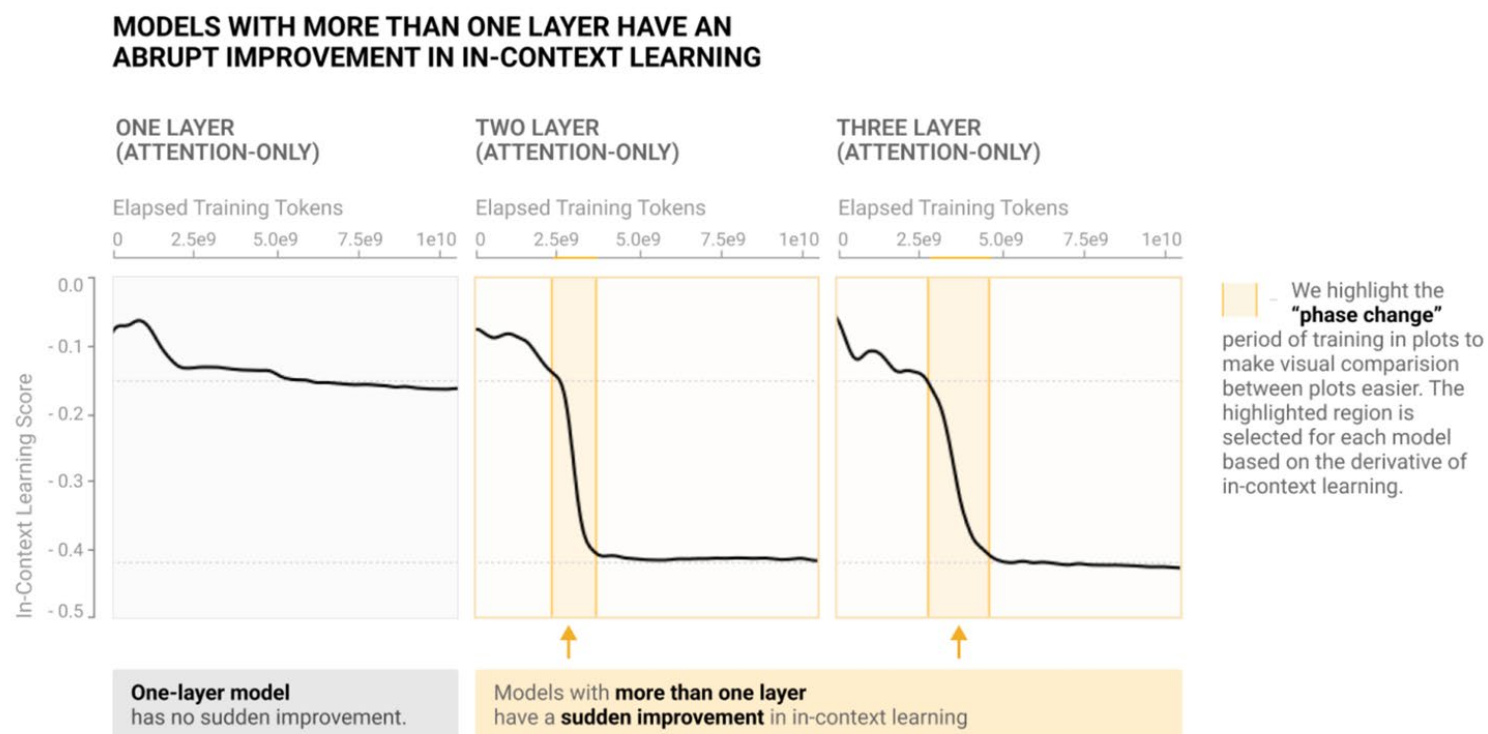
Mar 8, 2022

<https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>



Circuits: The Induction Head

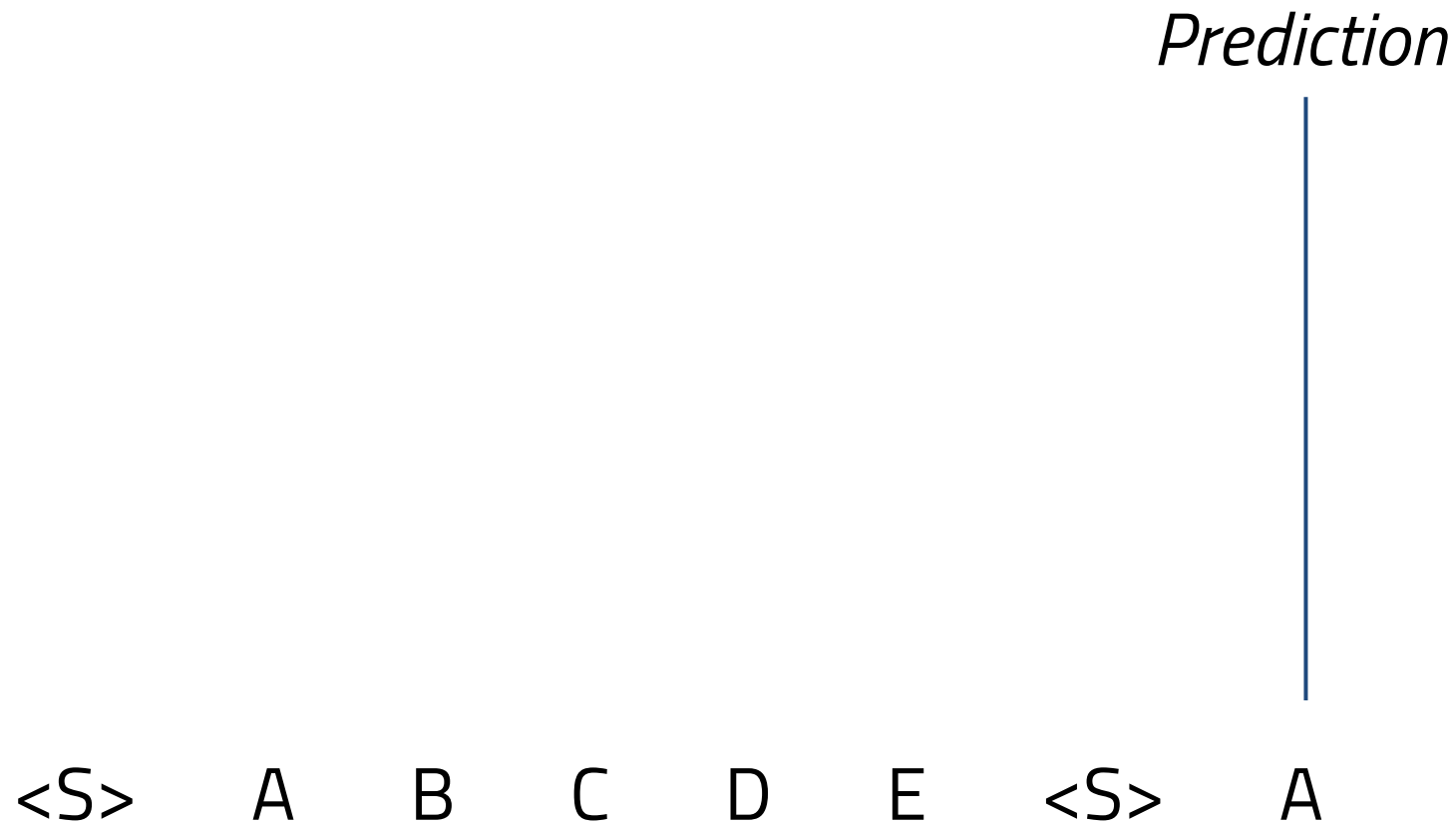
Induction heads are the most primitive form of in-context learning.



<https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>

Circuits: The Induction Head

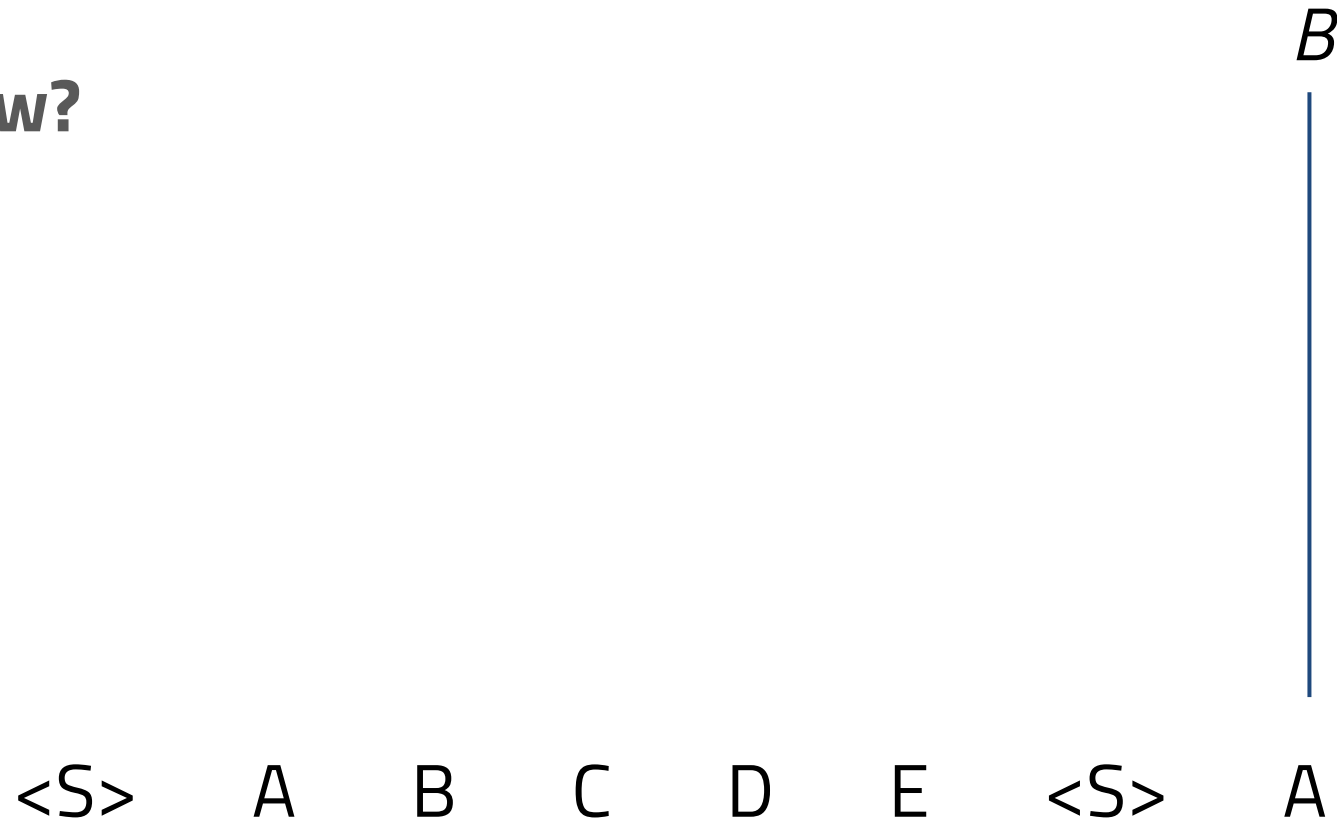
In-context copying illustration



Circuits: The Induction Head

Goal: Predict B

But how?



Circuits: The Induction Head

Let's play a game.

Assume you can

1. Attend to tokens based on positional information
2. Attend to tokens based on value (i.e. A would have high attention with only A)

Embed

[<S> | p-1] [A | pos] [B | pos] [C | pos] [D | pos] [E | pos] [<S> | pos] [A | pos]

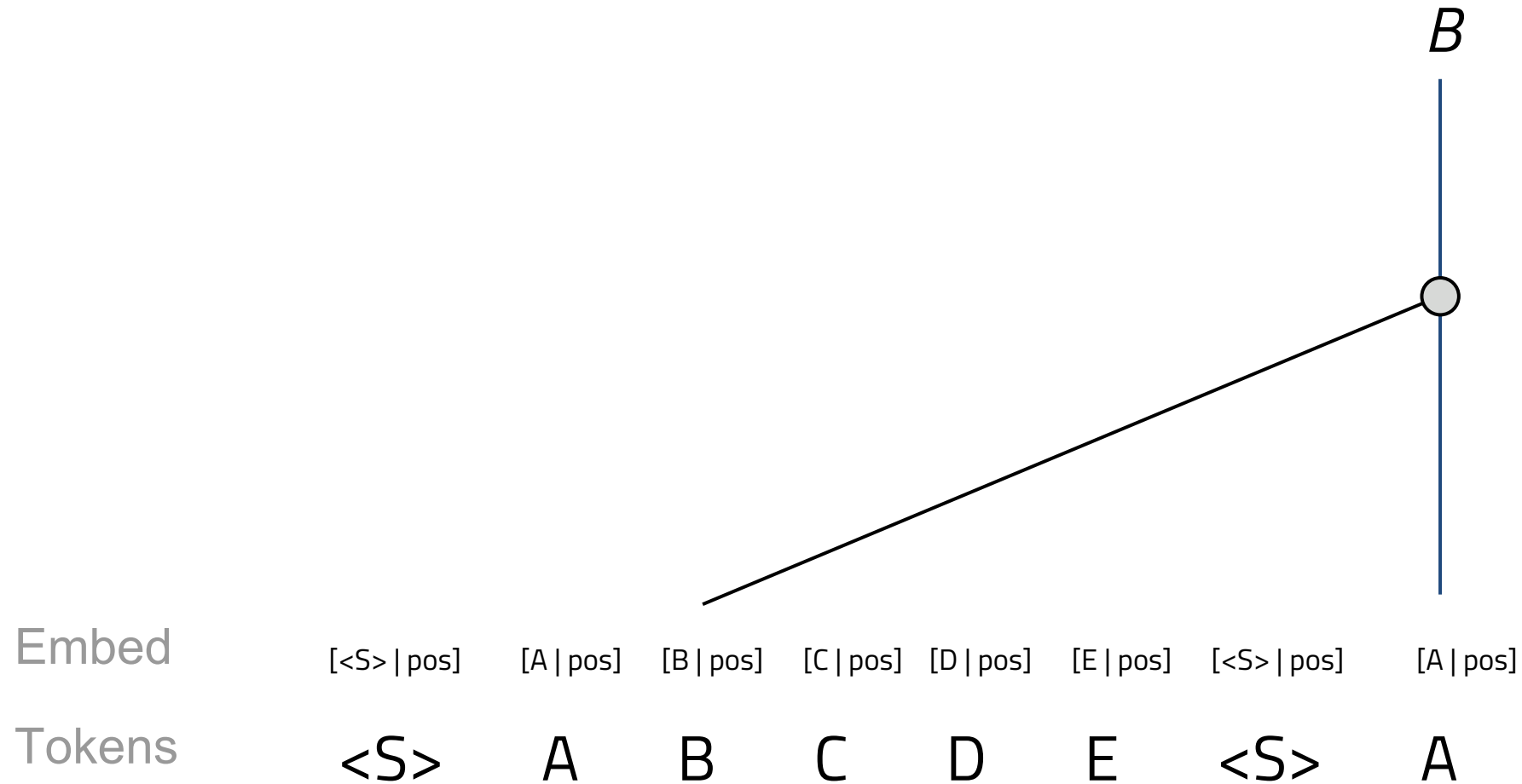
Token

<S> A B C D E <S> A

S

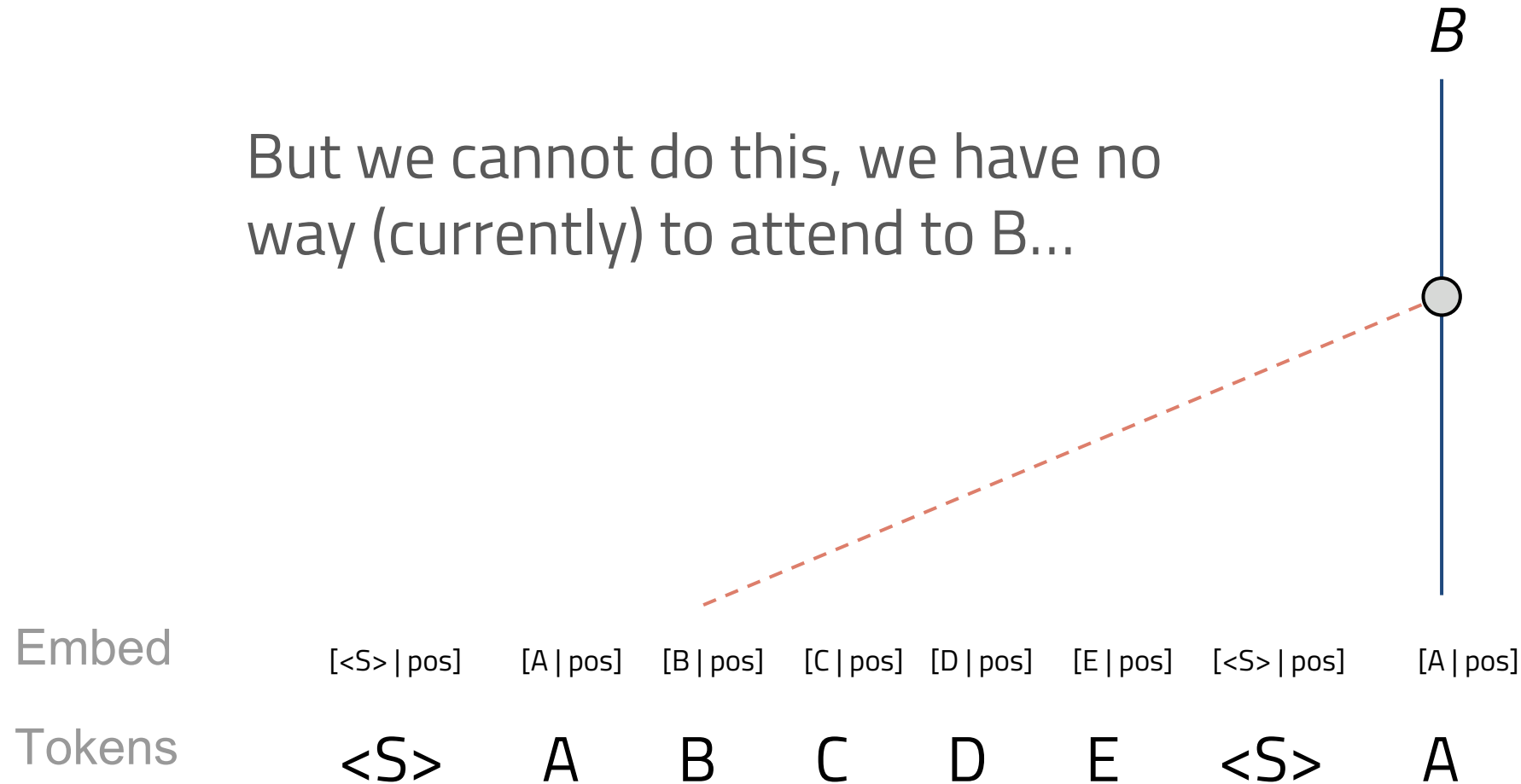


Circuits: The Induction Head



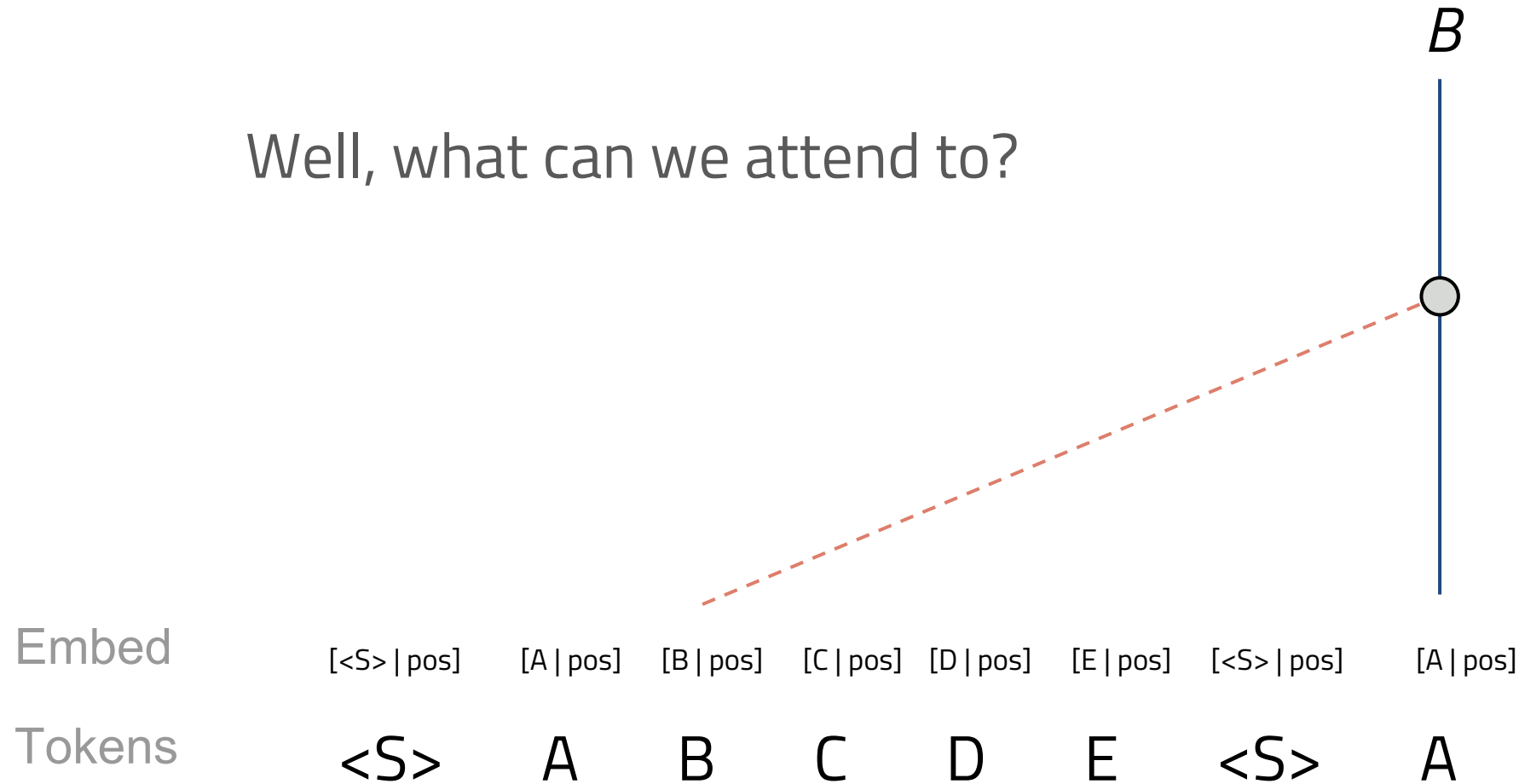
Circuits: The Induction Head

But we cannot do this, we have no way (currently) to attend to B...



Circuits: The Induction Head

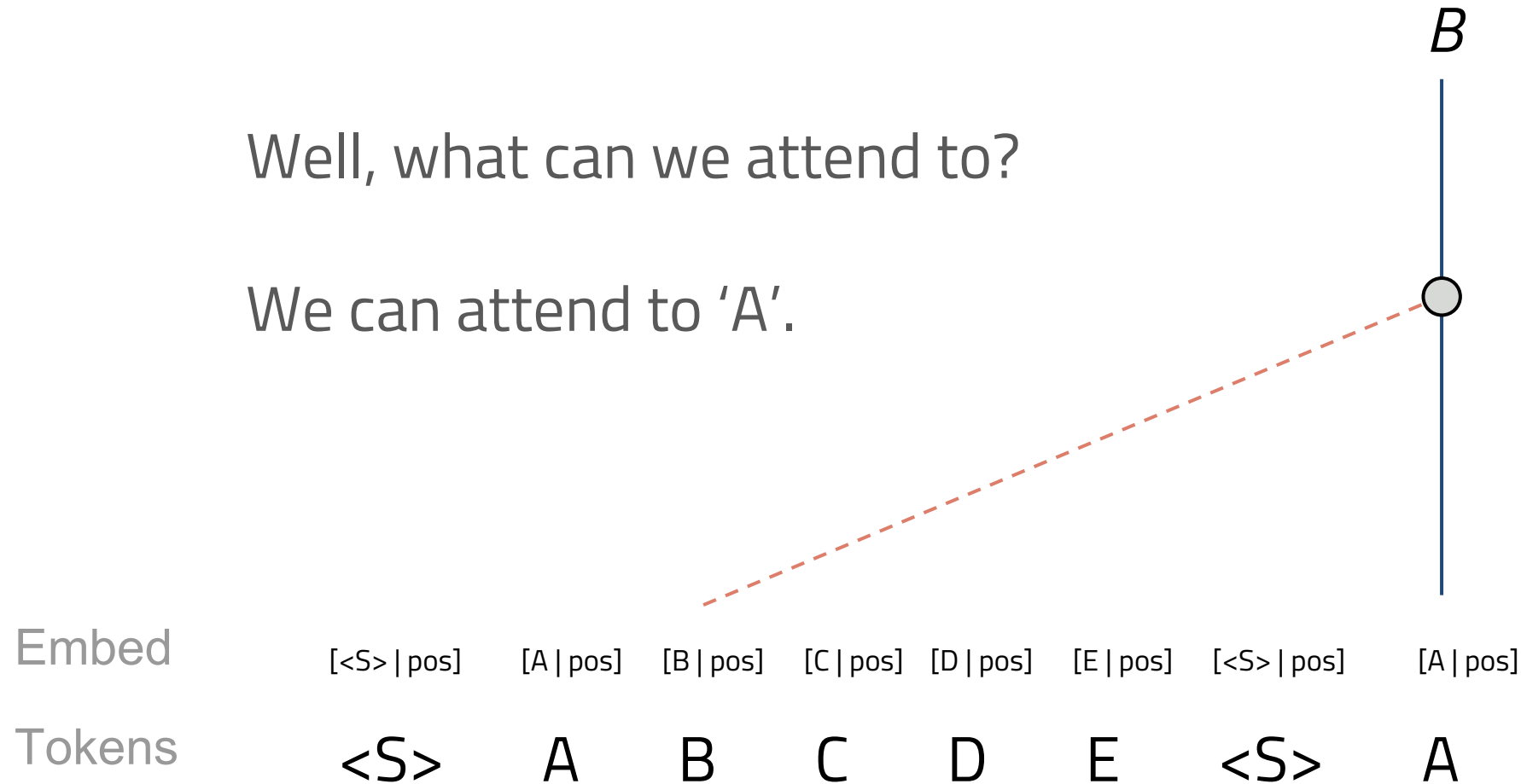
Well, what can we attend to?



Circuits: The Induction Head

Well, what can we attend to?

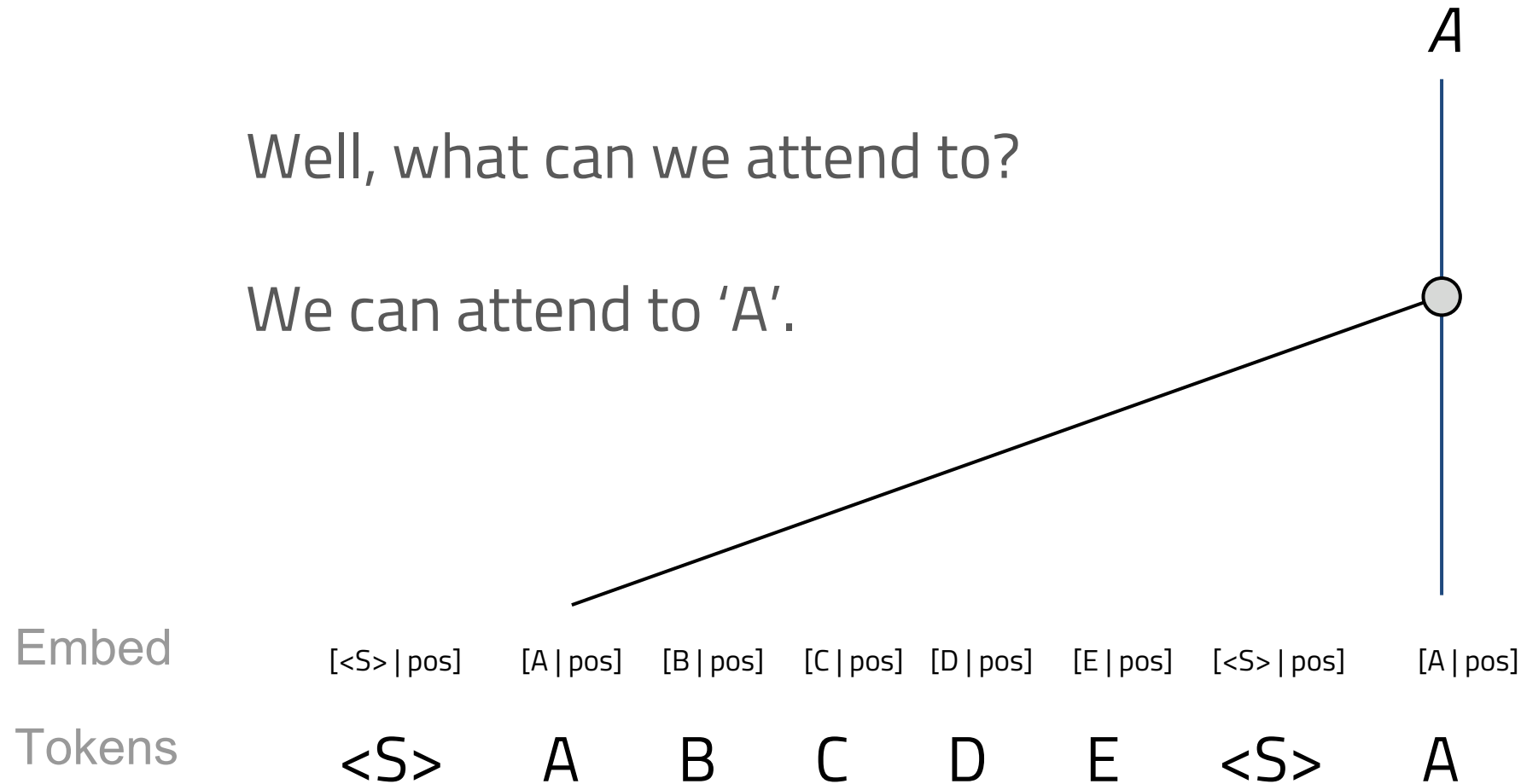
We can attend to 'A'.



Circuits: The Induction Head

Well, what can we attend to?

We can attend to 'A'.



Circuits: The Induction Head

Not what we want

But, this doesn't help us...

Embed

Tokens

[<S> | pos] [A | pos] [B | pos] [C | pos] [D | pos] [E | pos] [<S> | pos] [A | pos]

<S> A B C D E <S> A

A



Circuits: The Induction Head

Key: What if 'B' copies the token information of 'A'?

Embed	[<S> pos]	[A pos]	[B pos]	[C pos]	[D pos]	[E pos]	[<S> pos]	[A pos]
Tokens	<S>	A	B	C	D	E	<S>	A



Circuits: The Induction Head

Right-Shift
Tokens

[<S> | <S>] [A | <S>] [B | A] [C | B] [D | C] [E | D] [<S> | E] [A | <S>]

Embed

[<S> | pos] [A | pos] [B | pos] [C | pos] [D | pos] [E | pos] [<S> | pos] [A | pos]

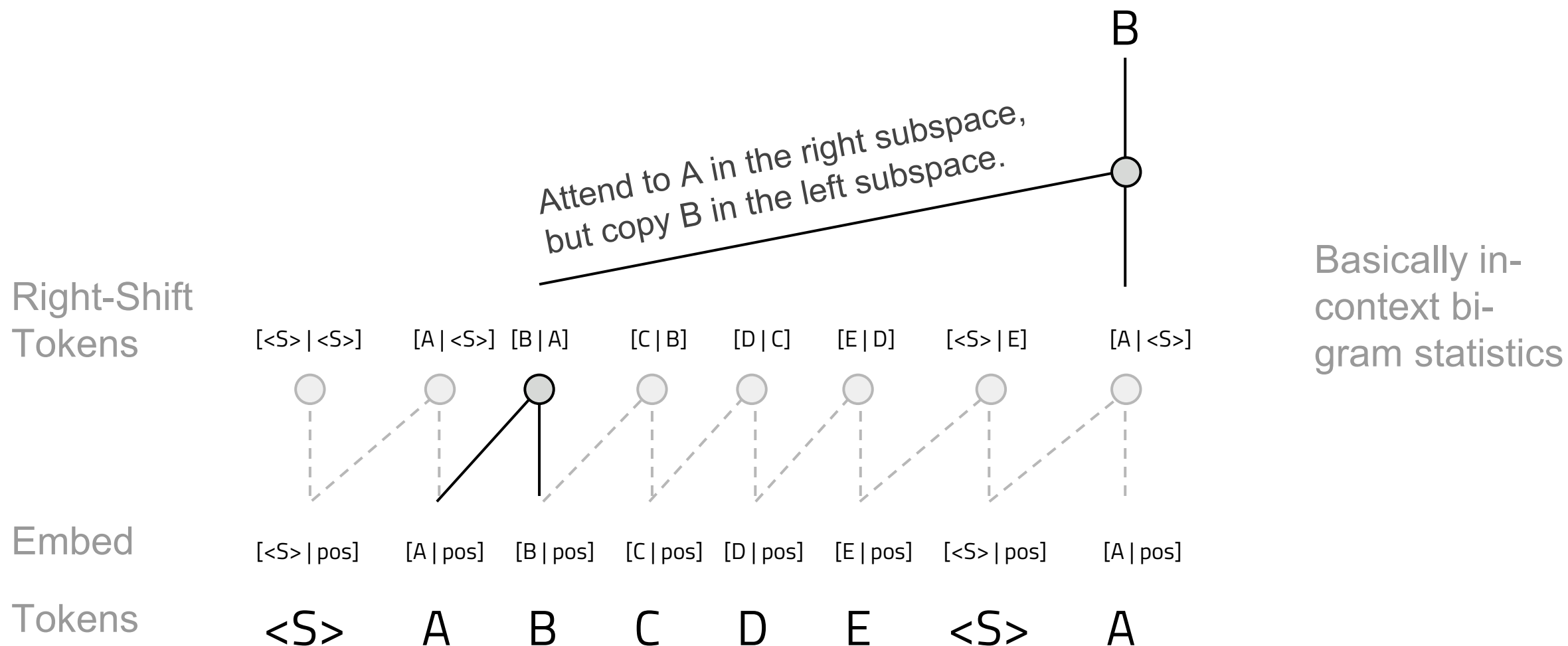
Tokens

<S> A B C D E <S> A

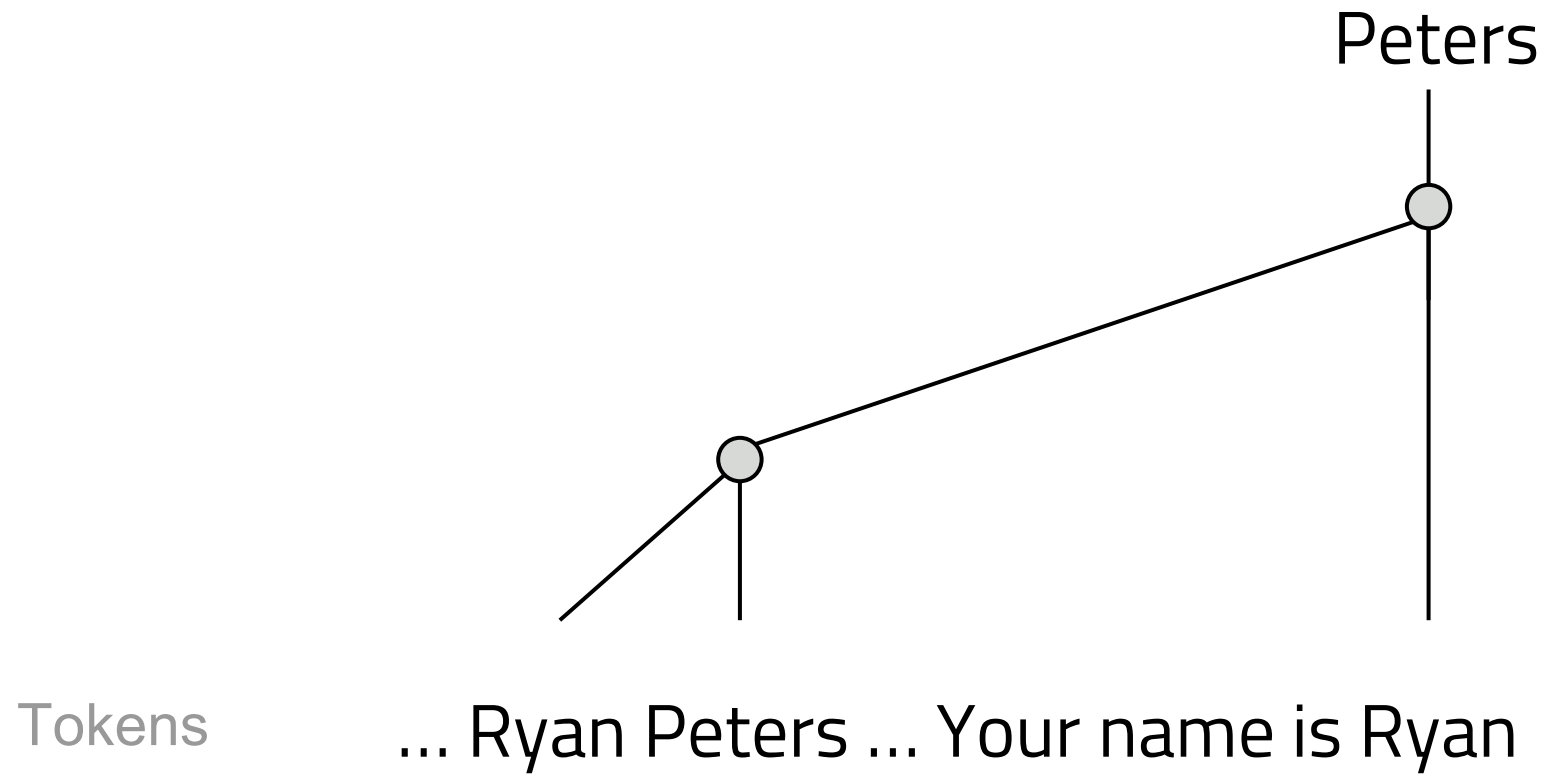
Basically in-
context bi-
gram statistics



Circuits: The Induction Head

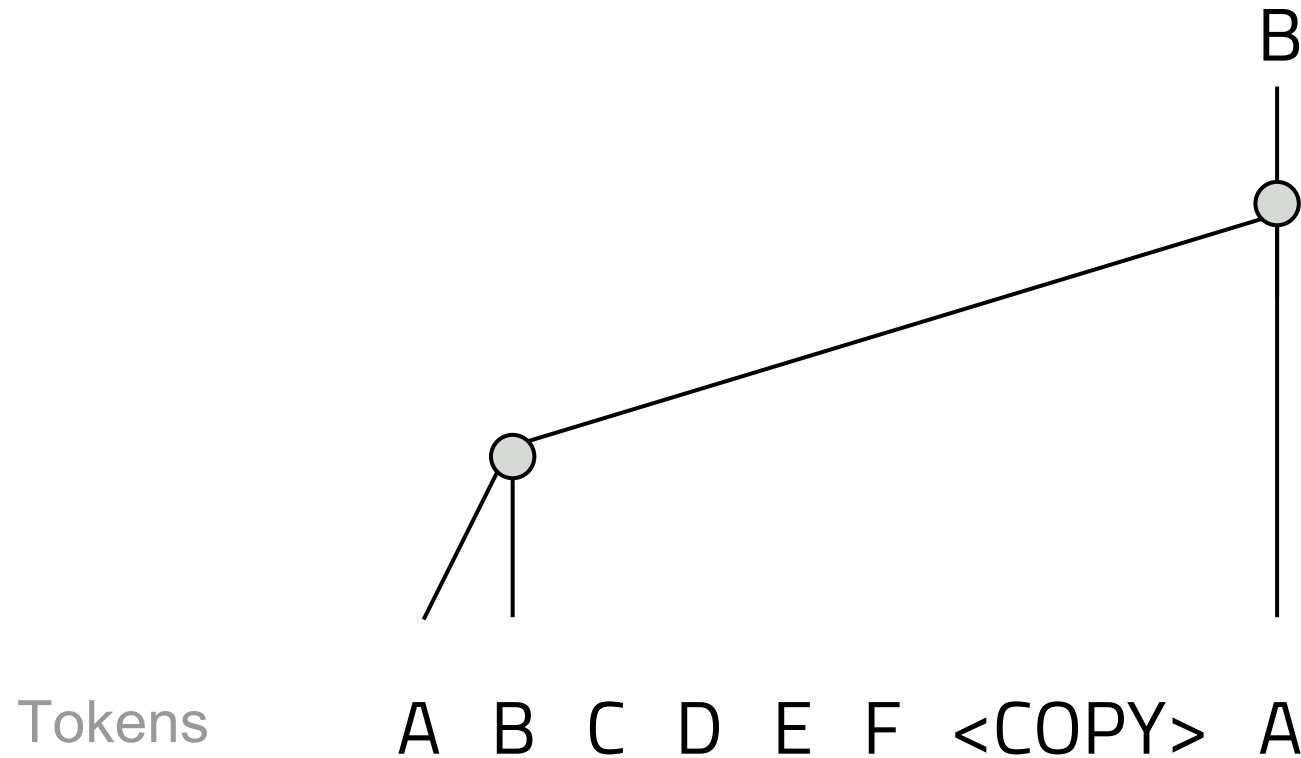


Circuits: The Induction Head



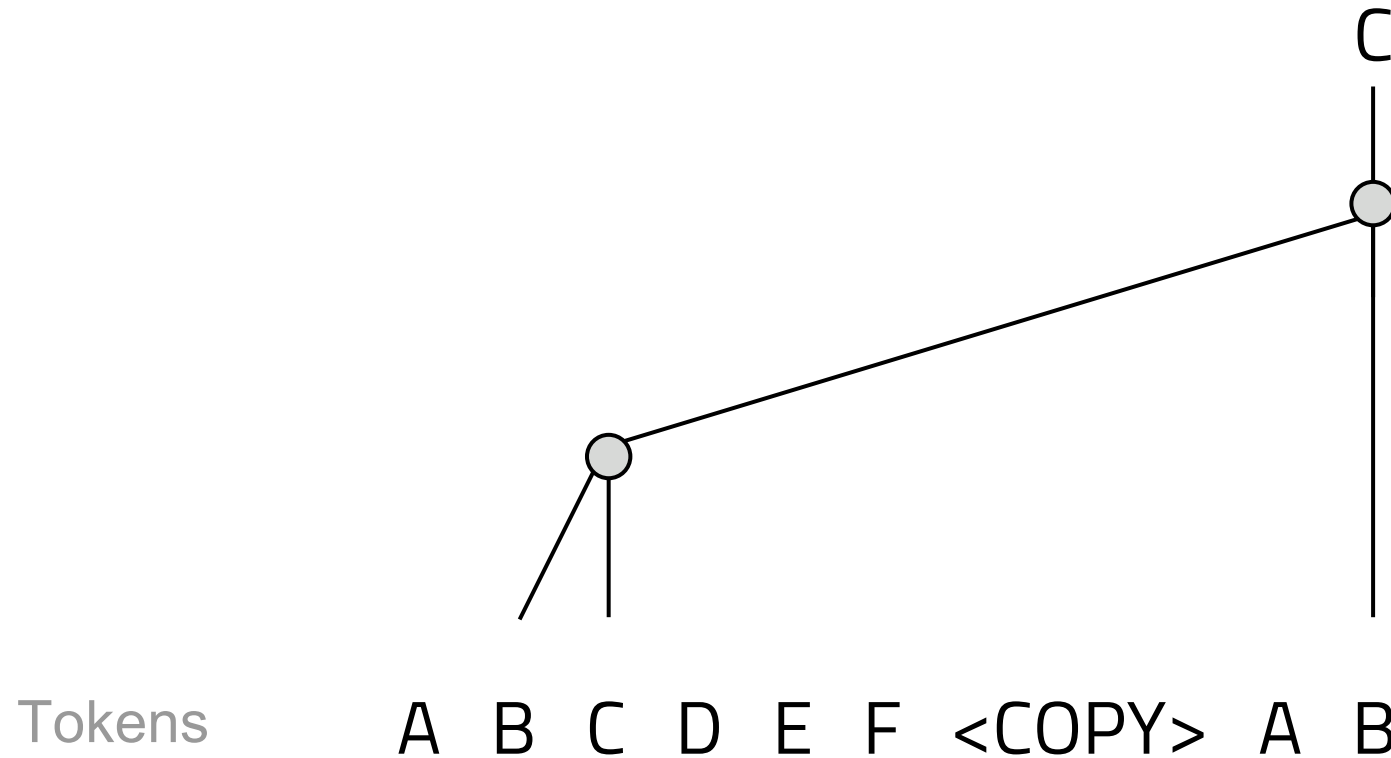
Sequence Copying

Copying of sequences uses n-gram induction heads.



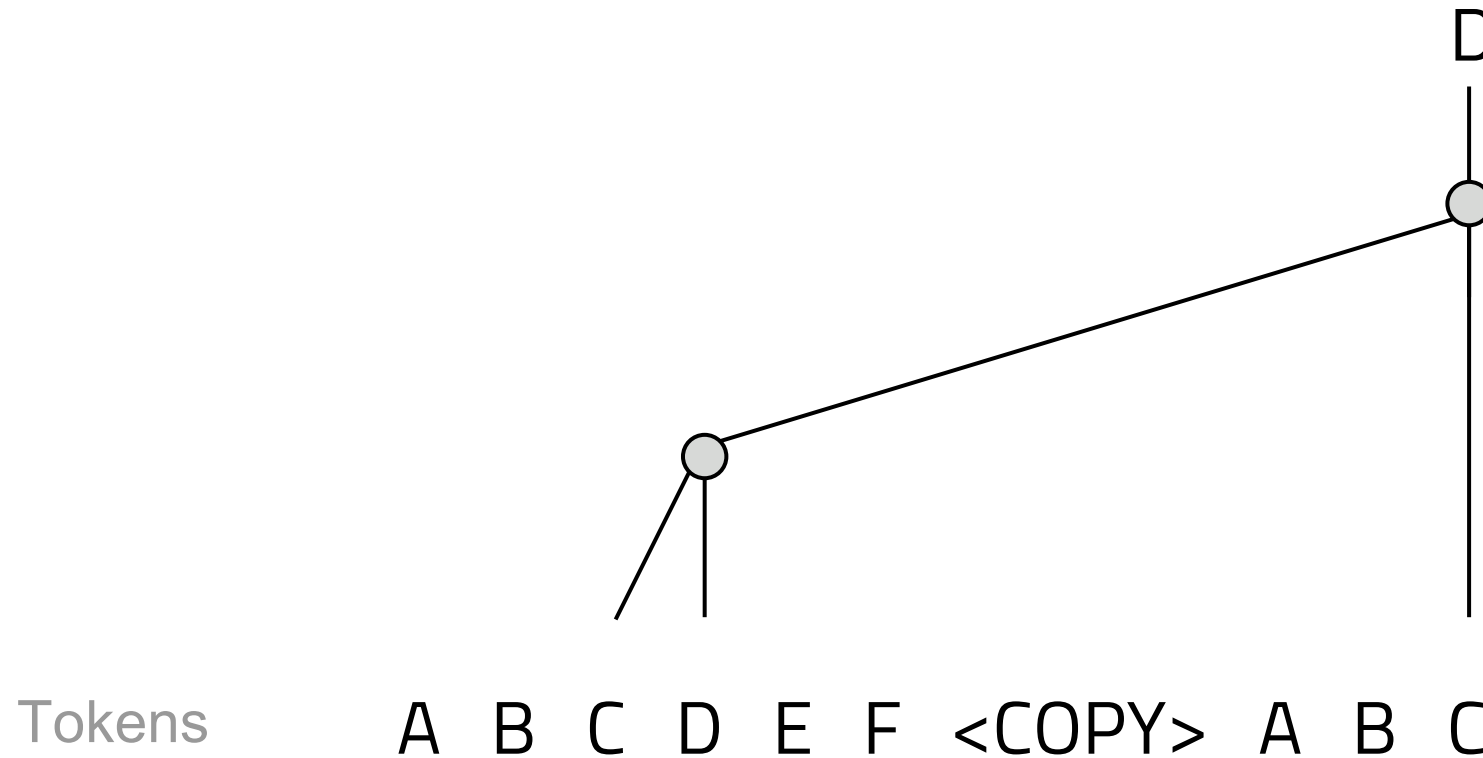
Sequence Copying

Copying of sequences uses n-gram induction heads.



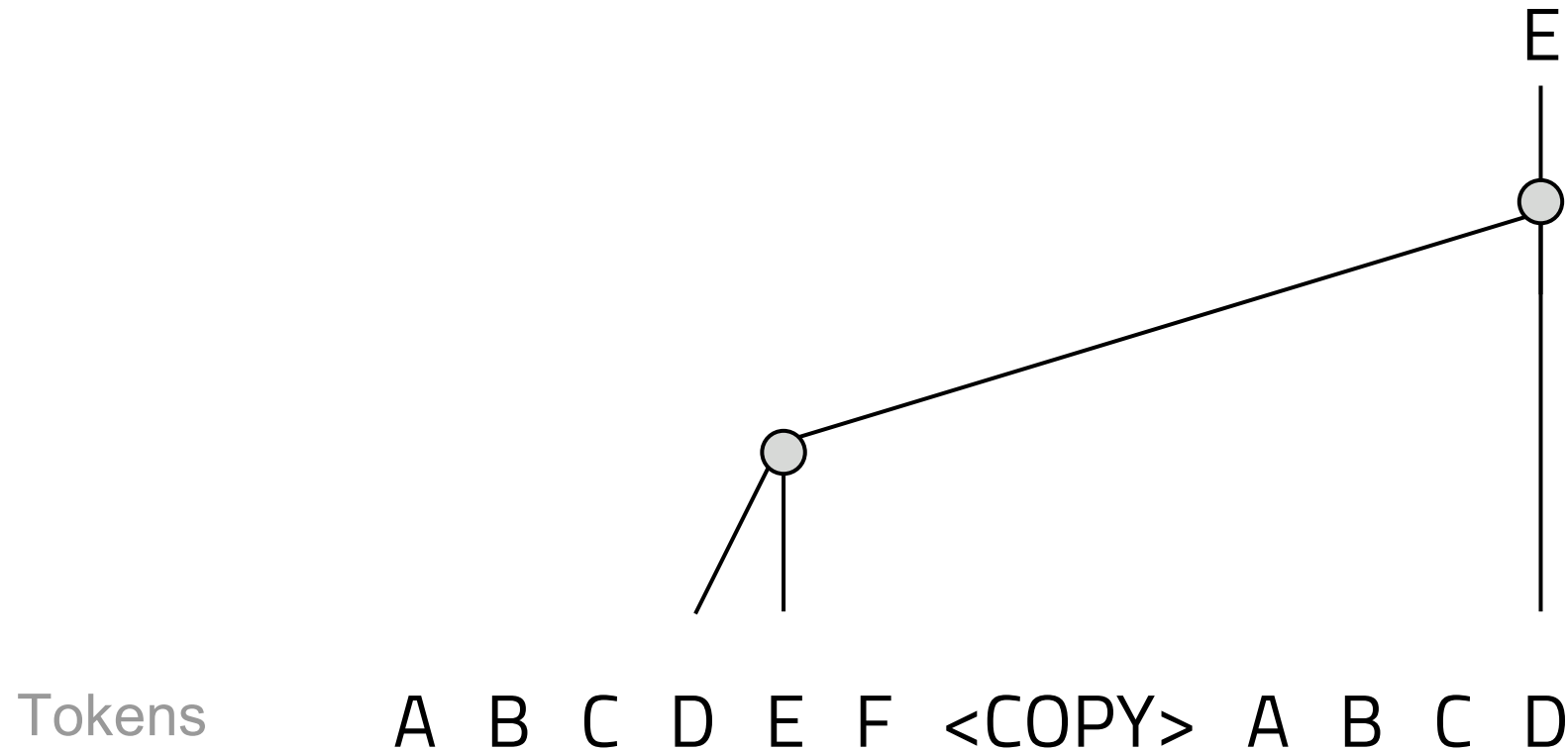
Sequence Copying

Copying of sequences uses n-gram induction heads.



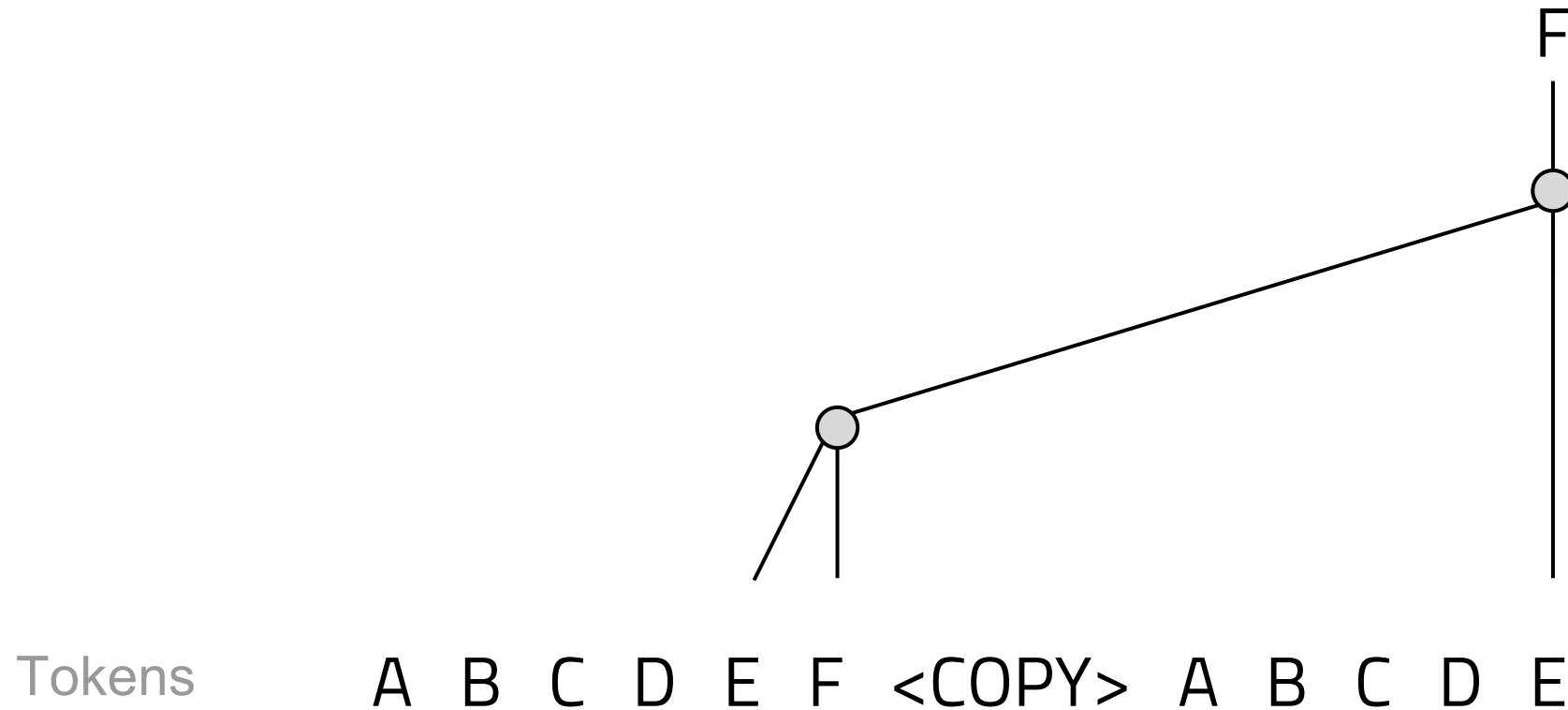
Sequence Copying

Copying of sequences uses n-gram induction heads.



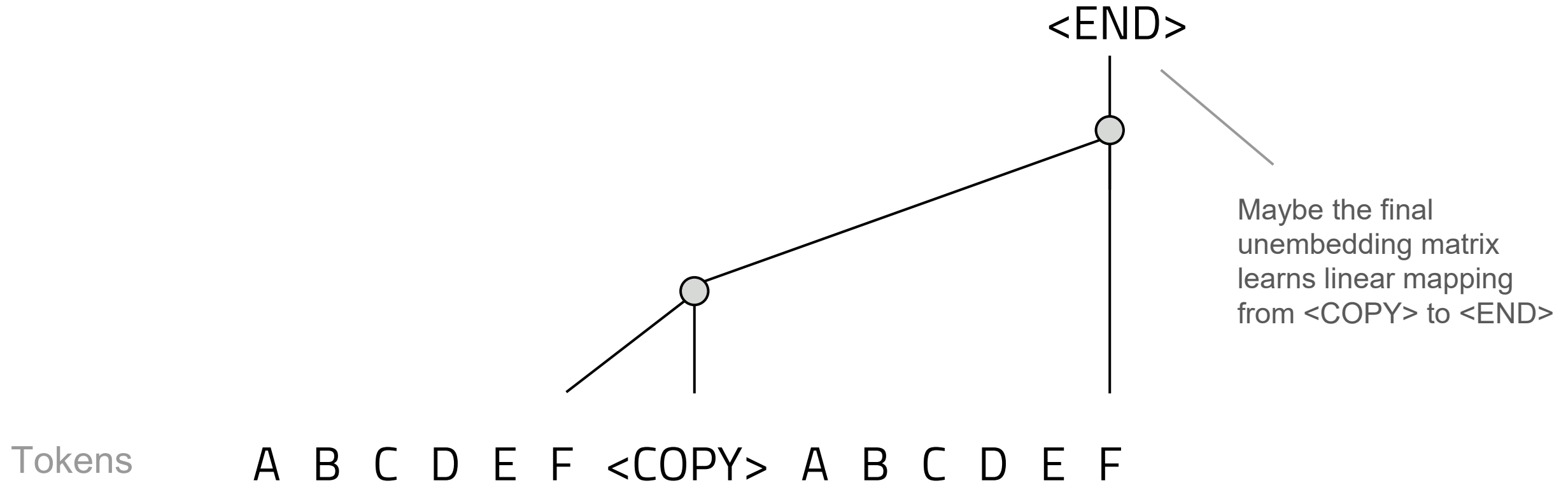
Sequence Copying

Copying of sequences uses n-gram induction heads.



Sequence Copying

Copying of sequences uses n-gram induction heads.



Sequence Copying

Copying of sequences uses n-gram induction heads.

Tokens A B C D E F <COPY> A B C D E F <END>



Other Known Circuits

INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL

Kevin Wang¹, Alexandre Variengien¹, Arthur Conmy¹, Buck Shlegeris¹ & Jacob Steinhardt^{1,2}

¹Redwood Research

²UC Berkeley

kevin@rdwrs.com, alexandre@rdwrs.com,
arthur@rdwrs.com, buck@rdwrs.com, jsteinhardt@berkeley.edu

ABSTRACT

Research in mechanistic interpretability seeks to explain behaviors of machine learning (ML) models in terms of their internal components. However, most previous work either focuses on simple behaviors in small models or describes complicated behaviors in larger models with broad strokes. In this work, we bridge this gap by presenting an explanation for how GPT-2 small performs a natural language task called indirect object identification (IOI). Our explanation encompasses 26 attention heads grouped into 7 main classes, which we discovered using a combination of interpretability approaches relying on causal interventions. To our knowledge, this investigation is the largest end-to-end attempt at reverse-engineering a natural behavior “in the wild” in a language model. We evaluate the reliability of our explanation using three quantitative criteria—*faithfulness*, *completeness* and *minimality*. Though these criteria support our explanation, they also point to remaining gaps in our understanding. Our work provides evidence that a mechanistic understanding of large ML models is feasible, pointing toward opportunities to scale our understanding to both larger models and more complex tasks. Code for all experiments is available at <https://github.com/redwoodresearch/Easy-Transformer>.

How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model

Michael Hanna*
ILLC
University of Amsterdam
m.w.hanna@uva.nl

Ollie Liu*
University of
Southern California
zliu2898@usc.edu

Alexandre Variengien[†]
Redwood Research
alexandre.variengien@gmail.com

Abstract

Pre-trained language models can be surprisingly adept at tasks they were not explicitly trained on, but how they implement these capabilities is poorly understood. In this paper, we investigate the basic mathematical abilities often acquired by pre-trained language models. Concretely, we use mechanistic interpretability techniques to explain the (limited) mathematical abilities of GPT-2 small. As a case study, we examine its ability to take in sentences such as “The war lasted from the year 1732 to the year 17”, and predict valid two-digit end years (years > 32). We first identify a circuit, a small subset of GPT-2 small’s computational graph that computes this task’s output. Then, we explain the role of each circuit component, showing that GPT-2 small’s final multi-layer perceptrons boost the probability of end years greater than the start year. Finally, we find related tasks that activate our circuit. Our results suggest that GPT-2 small computes greater-than using a complex mechanism that activates across diverse contexts.

COPY SUPPRESSION: COMPREHENSIVELY UNDERSTANDING AN ATTENTION HEAD

Callum McDougall^{1,†}, Arthur Conmy^{1,†}, Cody Rushing^{2,†}, Thomas McGrath^{1,*}, Neel Nanda³

¹Independent. ²University of Texas at Austin. ³Google DeepMind. [†]Joint contribution.

Correspondence to cal.s.mcdougall@gmail.com and neelnanda@google.com

ABSTRACT

We present a single attention head in GPT-2 Small that has one main role across the entire training distribution. If components in earlier layers predict a certain token, and this token appears earlier in the context, the head suppresses it: we call this copy suppression. Attention Head 10.7 (L10H7) suppresses naive copying behavior which improves overall model calibration. This explains why multiple prior works studying certain narrow tasks found negative heads that systematically favored the wrong answer. We uncover the mechanism that the Negative Heads use for copy suppression with weights-based evidence and are able to explain 76.9% of the impact of L10H7 in GPT-2 Small. To the best of our knowledge, this is the most comprehensive description of the complete role of a component in a language model to date. One major effect of copy suppression is its role in self-repair. Self-repair refers to how ablating crucial model components results in downstream neural network parts compensating for this ablation. Copy suppression leads to self-repair: if an initial overconfident copier is ablated, then there is nothing to suppress. We show that self-repair is implemented by several mechanisms, one of which is copy suppression, which explains 39% of the behavior in a narrow task. Interactive visualizations of the copy suppression phenomena may be seen at our web app <https://copy-suppression.streamlit.app/>.



Supplementary



Supplementary

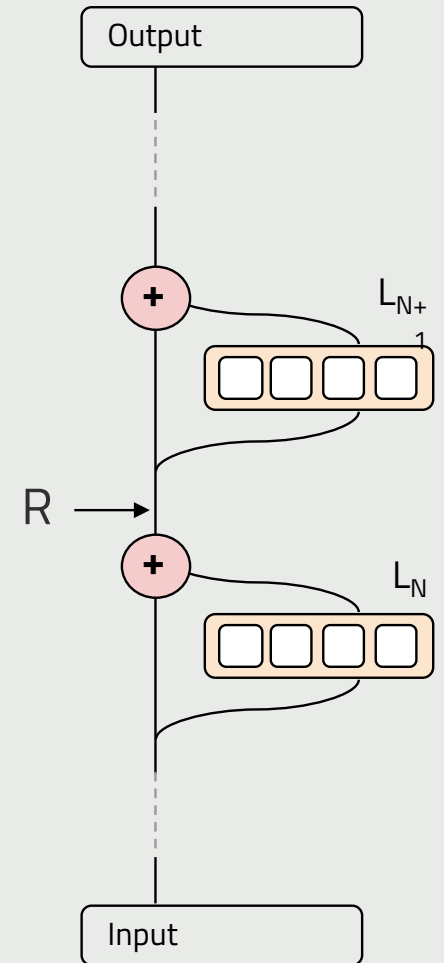
More Slides on Superposition and Polysemanticity



Superposition and Polysematicity

But first, we need to talk about how models represent information in the residual stream.

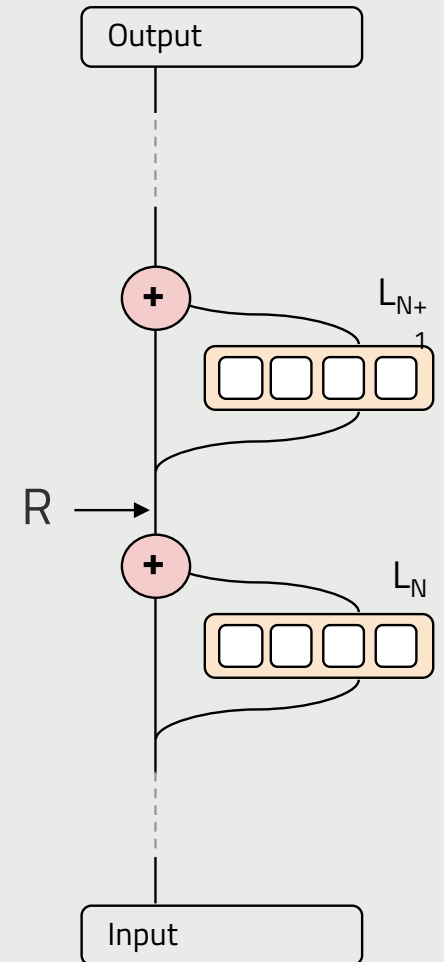
$$R = [A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9]$$



Superposition and Polysemanticity

But first, we need to talk about how models represent information in the residual stream.

$$R = [A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9]$$



Superposition and Polysemanticity

But first, we need to talk about how models represent information in the residual stream.

$$R = [\textcolor{red}{A_0}, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9]$$

Activity pattern of A_0

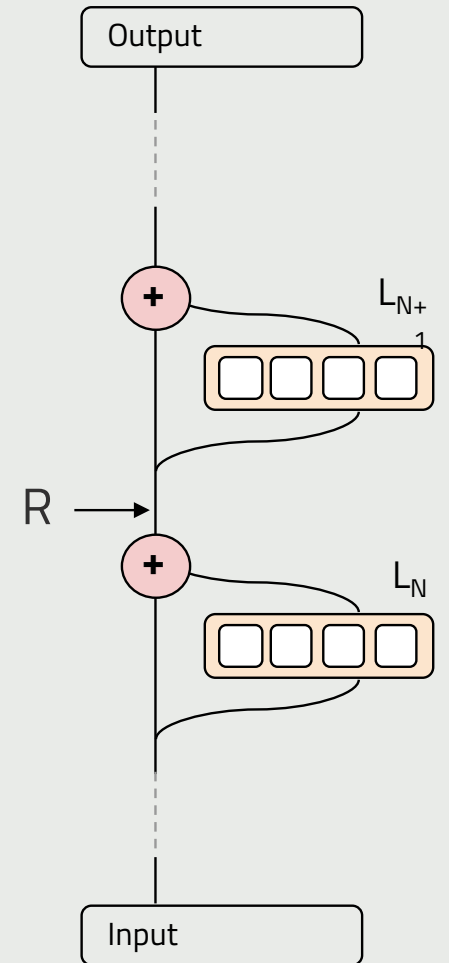
The has a of 12, it a strong base.

The was safe second base.

poured one ton of concrete base.

Negative activation

Positive activation



Superposition and Polysemanticity

But, we can make sense of 'directions' or 'linear combinations of neurons', often refer to as 'features'

$$R = [\overset{\text{Negative activation}}{A_0}, \overset{\text{Negative activation}}{A_1}, A_2, A_3, A_4, \overset{\text{Positive activation}}{A_5}, A_6, A_7, A_8, A_9]$$

Activity pattern of $3A_0 + 1A_1 - 5A_9$

The chemical has a pH of 12, making it a strong base.

The runner was safe at second base.

They poured one ton of concrete for the building's base.

Negative activation

Positive activation

