CSCI 5541: Natural Language Processing

Lecture 20: Interpretability

Slides borrowed from Sarah Wiegreffe, Hosein Mohebbi et al.

UNIVERSITY OF MINNESOTA
Driven to Discover®

Overview

- What is interpretability?
- Model Agnostic Interpretability
- Mechanistic Interpretability
  - Circuits
  - Logit-Lens
  - Neuron Level Interpretability
- Concluding Remarks

Why do we need interpretability?

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
    - *What biases does it contain? Does it discriminate against particular groups?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
   - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
   - *Models that are deployed carry a degree of responsibility, can we trust them?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
   - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
   - *Models that are deployed carry a degree of responsibility, can we trust them?*
3. **Robustness**
   - *Does our model generalise robustly to unseen data?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. **Fairness**
   - *What biases does it contain? Does it discriminate against particular groups?*
2. **Trustworthiness**
   - *Models that are deployed carry a degree of responsibility, can we trust them?*
3. **Robustness**
   - *Does our model generalise robustly to unseen data?*
4. **Faithfulness**
   - *Is a model right for the right reasons?*

Lipton (2018) - *The Mythos of Model Interpretability*

Why do we nee

The **desiderata** of alg

1. **Fairness**
   - *What biases doe* *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *Is a model right*

18) - *The Mythos of Model Interpretability*

BBC

NEWS Menu

Business | Market Data | New Tech Economy | Technology of Bu

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

It follows complaints - including from Apple's co-founder Steve Wozniak - that algorithms used to set limits might be inherently biased against women.

New York's Department of Financial Services (DFS) has contacted Goldman Sachs, which runs the Apple Card.

Why do we nee

The **desiderata** of alg

1. **Fairness**
   - *What biases doe* *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *How faithful are* *oning?*

BBC

NEWS Menu

Tech

Facebook apology as AI labels black men 'primates'

6 September 2021

Facebook users who watched a newspaper video featuring black men were asked if they wanted to "keep seeing videos about primates" by an artificial-intelligence recommendation system.

Facebook told BBC News it "was clearly an unacceptable error", disabled the system and launched an investigation.

"We apologise to anyone who may have seen these offensive recommendations."

18) - *The Mythos of Model Interpretability*

Why do we nee

The **desiderata** of alg

1. **Fairness**
   - *What biases doe* *ainst particular groups?*
2. **Trustworthiness**
   - *Models that are* *bility, can we trust them?*
3. **Robustness**
   - *Does our model*
4. **Faithfulness**
   - *How faithful are* *oning?*

18) - *The Mythos of Model Interpretability*

503px × 920px

BBC

NEWS

Menu

Tech

Twitter finds racial bias in image-cropping AI

20 May 2021

Preferences for white people over black people and women over men were found in testing

Twitter's automatic cropping of images had underlying issues that favoured white individuals over black people, and women over men, the company said.

It comes months after its users highlighted potential problems with the algorithm, which cropped large photos.

The social network's follow-up research has now confirmed the problem.

Netherlands

Dutch government resigns over child benefits scandal

PM Mark Rutte will stay on in caretaker capacity until general elections scheduled for 17 March

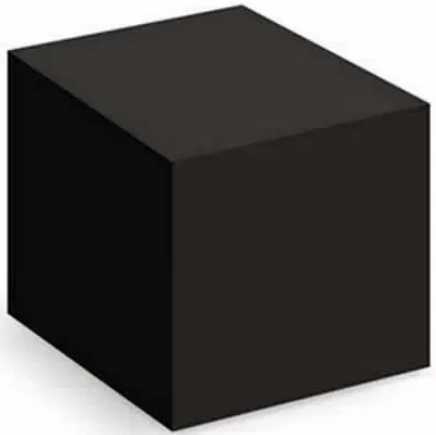
Jon Henley *Europe correspondent*
@jonhenley
Fri 15 Jan 2021 15.32 CET

Mark Rutte appears at a press conference in The Hague after the resignation of the coalition. Photograph: Bart Maat/EPA

The Dutch government has resigned amid an escalating scandal over child benefits in which more than 20,000 families were wrongly accused of fraud by the tax authority.

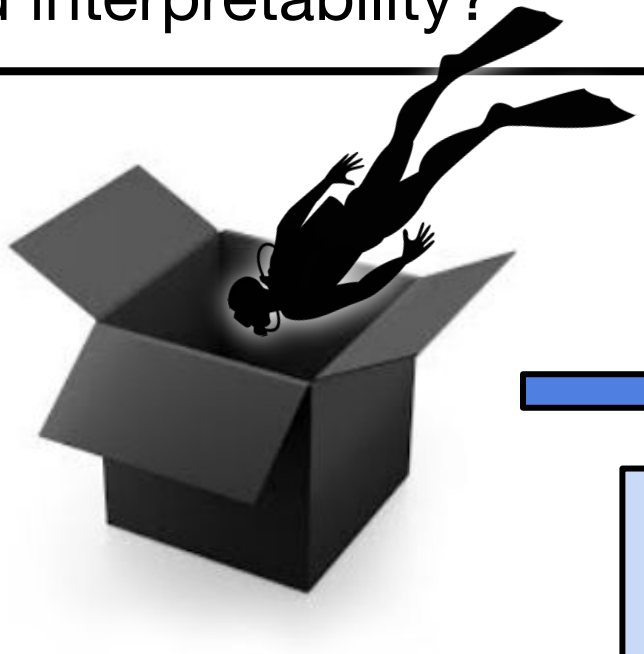
retability

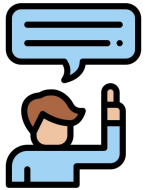
Why do we need interpretability?



?

Why do we need interpretability?





Post-hoc
Explanation

Explanation Faithfulness

How do we ensure that an explanation **faithfully** represents a model's **reasoning**?

Plausibility does **not** imply faithfulness!

Models can be **right for the wrong reasons**!

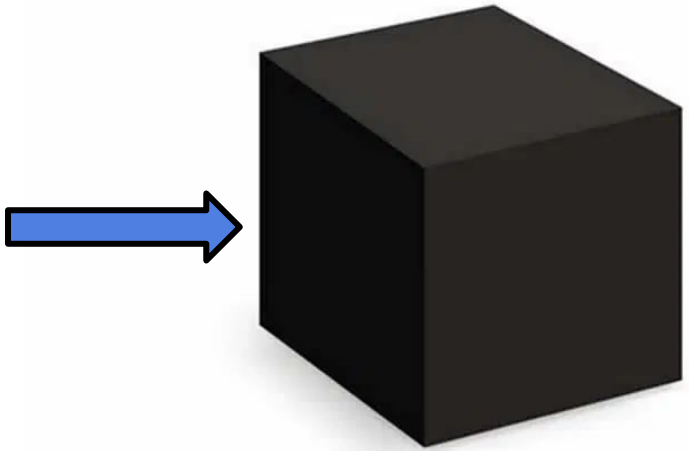
How do we ever know an explanation is truly **faithful** to the model?

Clever Hans

McCoy et al. (2019)

Explanation Faithfulness

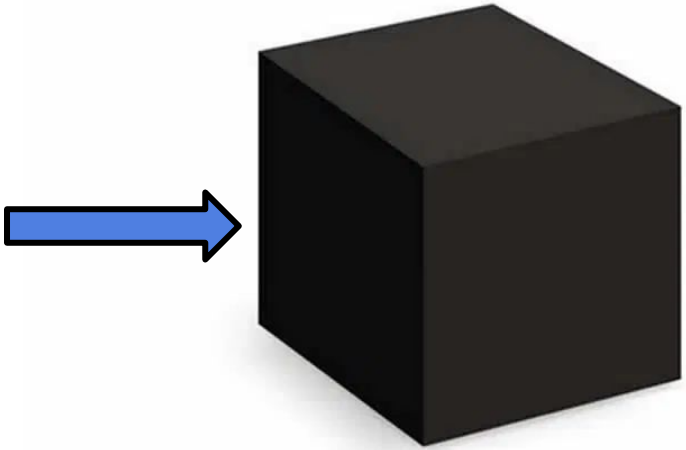
John is a 48 year old male lawyer from Amsterdam



€5000 credit

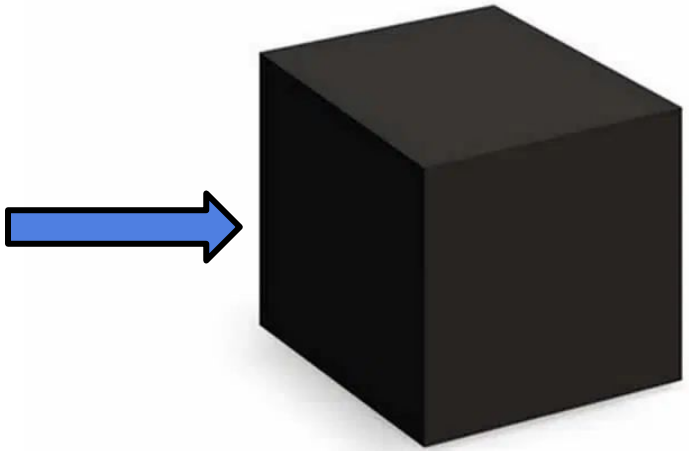
Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht



Explanation Faithfulness

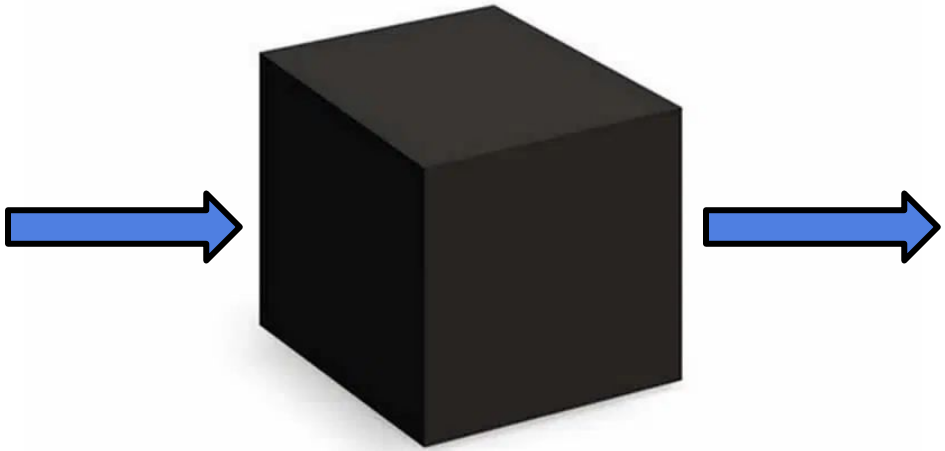
Suzan is a 32 year old female doctor from Utrecht



€1000 credit

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht



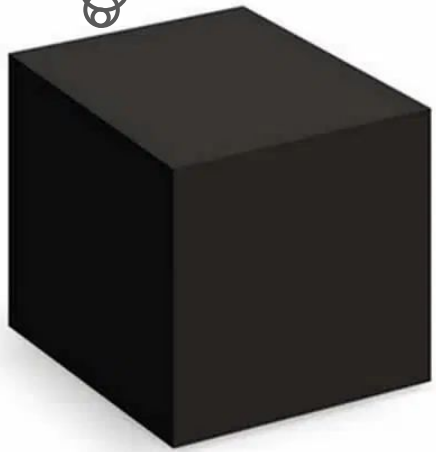
Why does Suzan get less than John? Because of her **age**? **Gender**? **Occupation**? **Location**?

€1000 credit

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht

female : 0.93
32 : 0.02
doctor : 0.03
Utrecht : 0.02

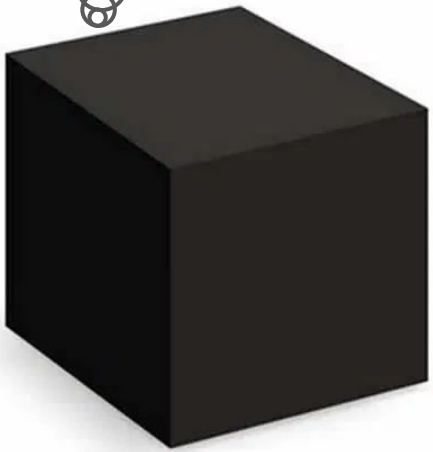


€1000 credit

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht

female : 0.93
32 : 0.02
doctor : 0.03
Utrecht : 0.02



female : 0.03
32 : 0.92
doctor : 0.03
Utrecht : 0.02

?

€1000 credit

Explaining Complex Systems

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

Explaining Complex Systems

1. Computational

- What does the system do?
- What problems does it solve or overcome?

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

Explaining Complex Systems

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

1. Computational

- What does the system do?
- What problems does it solve or overcome?

2. Algorithmic

- How does the system do what it does?
- What representations does it use?

Explaining Complex Systems

Marr's Tri-Level Hypothesis

Marr & Poggio (1976) - *From Understanding Computation to Understanding Neural Circuitry*

1. Computational

- What does the system do?
- What problems does it solve or overcome?

2. Algorithmic

- How does the system do what it does?
- What representations does it use?

3. Implementational

- How is the system physically realised?
- What neural circuitry implements the system?

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?

Marr's Level

1. Computational

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?

Marr's Level

1. Computational

2. Algorithmic

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What *abstract features* are encoded by the model?

Marr's Level

1. Computational

2. Algorithmic

Explaining Neural Models

Levels of explanation *granularity*:

1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What *abstract features* are encoded by the model?
4. **Mechanistic**
   - Can we identify specific *circuits* responsible for a particular behaviour?

Marr's Level

1. Computational

2. Algorithmic

3. Implementational

Behavioural Interpretability

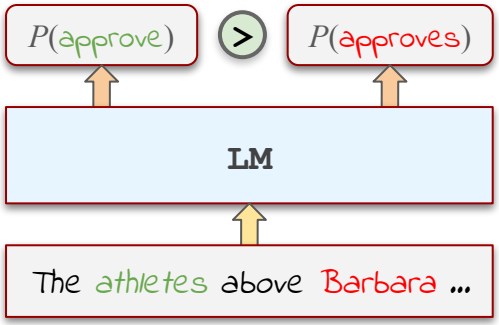
How can we understand a model better, without ‘opening the black box’?

- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.

Behavioural Interpretability

How can we understand a model better, without ‘opening the black box’?

- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.
- This type of experiment only requires access to the **output probabilities** of the model.



Behavioural Interpretability

- Assessing linguistic competence via minimal pairs:
  - **BLiMP** & **SyntaxGym**: Benchmark **suites** of different linguistic phenomena:

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	72.2	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	83.0	99.3	81.8	80.9	81.9	95.8	89.3	81.3	91.9	72.7	76.8	79.0	86.4
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Warstadt et al. (2020)
Hu et al. (2020)

Limitations of Behavioural Tests

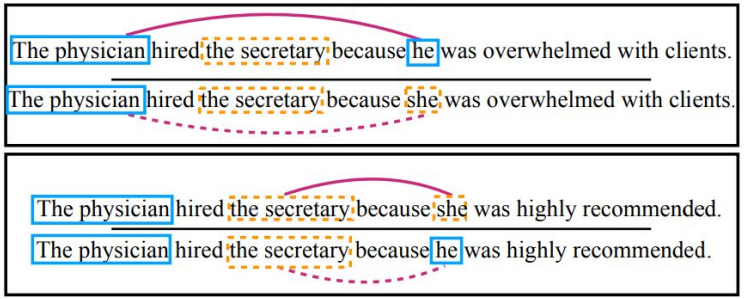
Behavioural tests show us a model's response to a particular input

- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!

Limitations of Behavioural Tests

Behavioural tests show us a model's response to a particular input

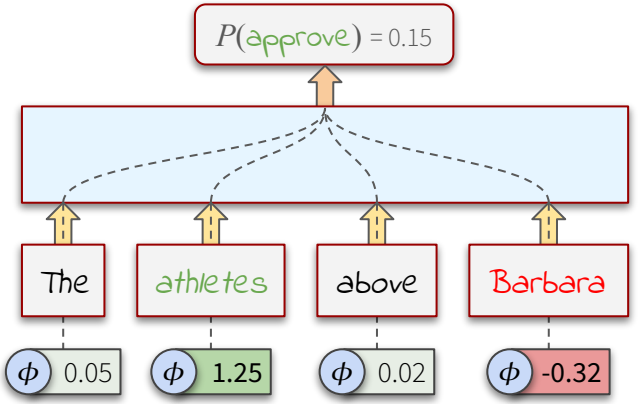
- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!
- Complex phenomena require more complex explanations
- E.g. *coreference resolution*:



Zhao et al. (2018), Jumelet et al. (2019)

Feature Attribution Methods

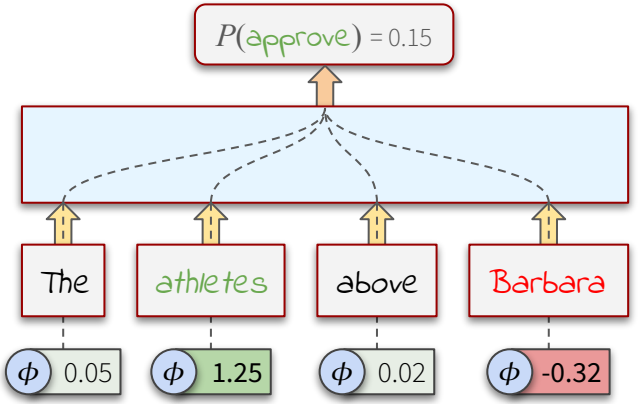
- **Feature attribution methods** explain model predictions in terms of the strongest *contributing* features.
- By normalizing such scores we get an insight into the **relative importance** of each feature.
- Shows us the **rationale** of a model behind a prediction → useful for uncovering biases!



Feature Attribution Methods

How do we compute the relative importance of a feature?

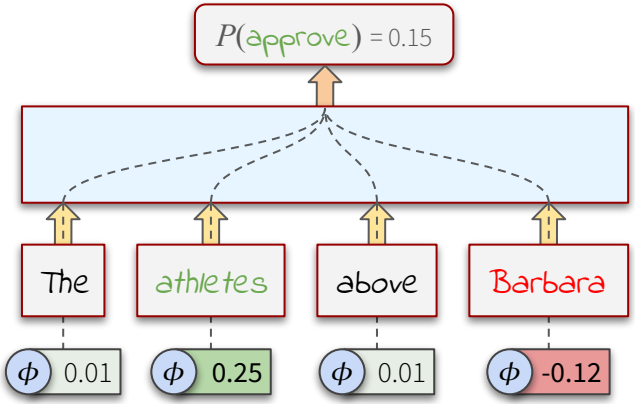
- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.



Feature Attribution Methods

How do we compute the relative importance of a feature?

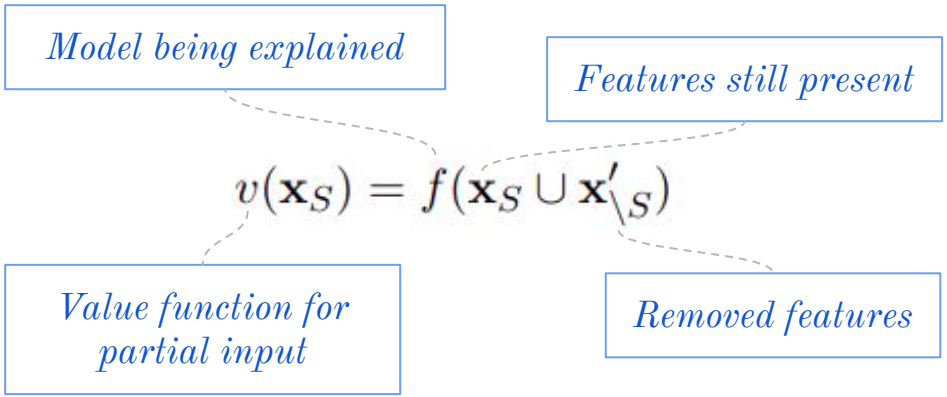
- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.
- How should we perturb?
- How can we represent the *missingness* of a feature?
- How should we measure the change?



Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Static Baseline



x = *This movie is not bad*

x' = <unk>

\S = {*not*}

$x_S \cup x'_{\setminus s}$ = *This movie is* <unk> *bad*

Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Interventional Baseline

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})\right]$$

Expectation over removed features

x = “This movie is not bad”

\S = {*not*}

$x_S \cup x'_{\setminus s}$ = “This movie is *the* *is* *walk* ... bad”

Baselines

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**:

Observational Baseline

Conditioned on present features

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S\right]$$

Expectation over removed features

x = “This movie is not bad”

\S = {*not*}

$x_S \cup x'_{\setminus s}$ = “This movie is *very* / *that* / *quite* / *...* bad”

Baselines

- More targeted baselines allow for precise **counterfactual** explanations:

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

Importance of feature x: difference of output when removing x

$$S_E(x_i) = q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})$$

Baseline

Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*

Baselines

- More targeted baselines allow for precise **counterfactual** explanations:

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

2. Why did the model predict "barking" *instead of* "crying"?
Can you stop the dog from

3. Why did the model predict "barking" *instead of* "walking"?
Can you stop the dog from

Importance of feature: difference of output when removed

$$S_E(x_i) = q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})$$

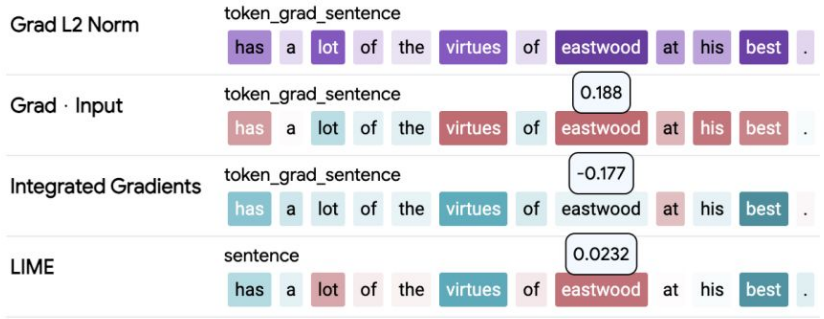
$$S^*_E(x_i) = (q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})) - (q(y_f|\boldsymbol{x}) - q(y_f|\boldsymbol{x}_{\neg i}))$$

*Explanation with respect to **foil***

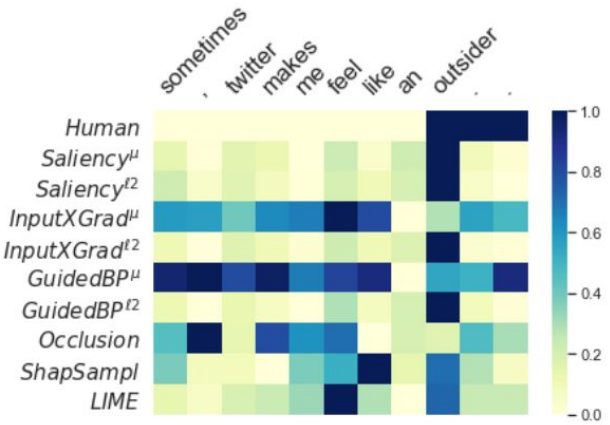
Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*

Limitations of Feature Attributions

- Attribution methods **disagree** strongly
- Which explanation is the right one?
- Can we simplify model behaviour to a single explanation?



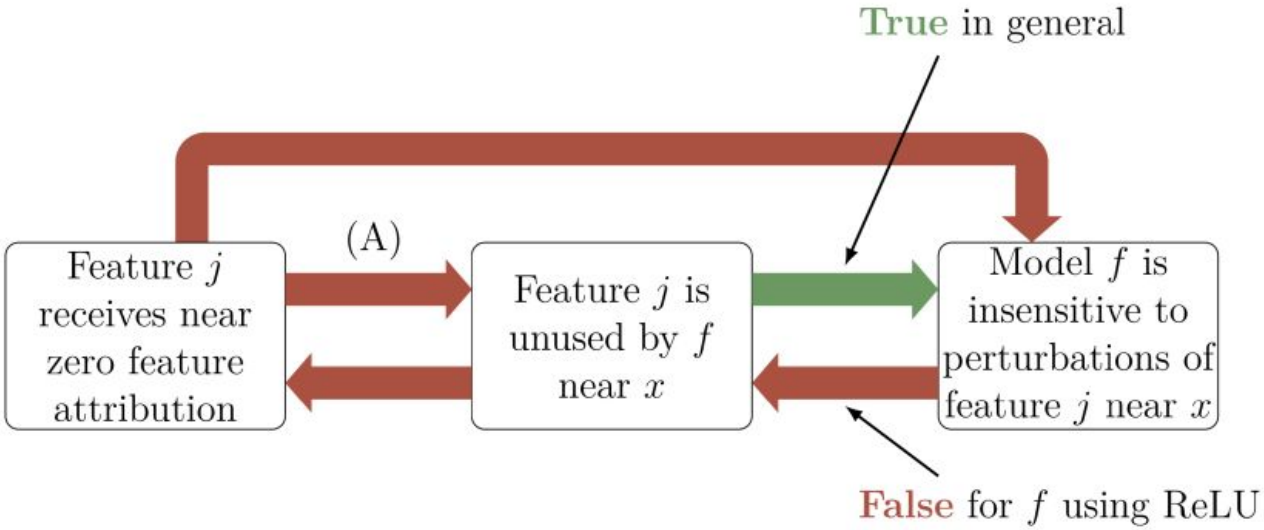
Bastings et al. (2022)



Atanasova et al. (2020)

Limitations of Feature Attributions

Bilodeau et al. (2024, PNAS): *Feature attributions can* ***provably fail*** *to improve on random guessing for inferring model behavior*



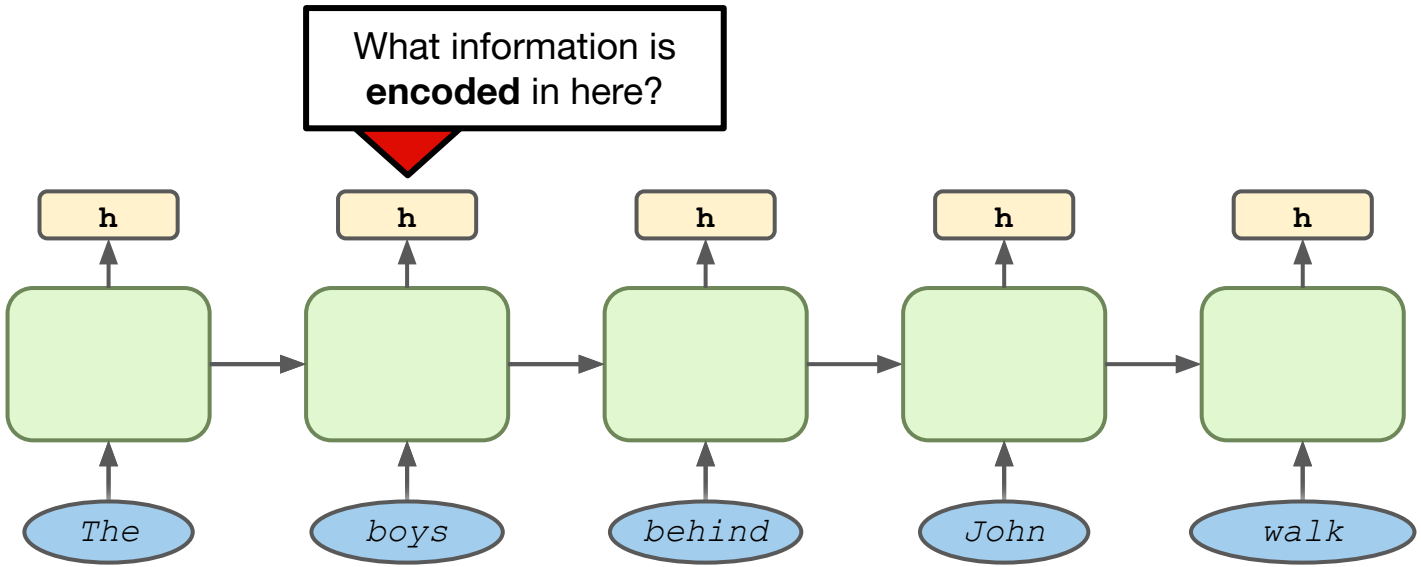
Probing

Feature attribution methods showed us which input features were important for a prediction.

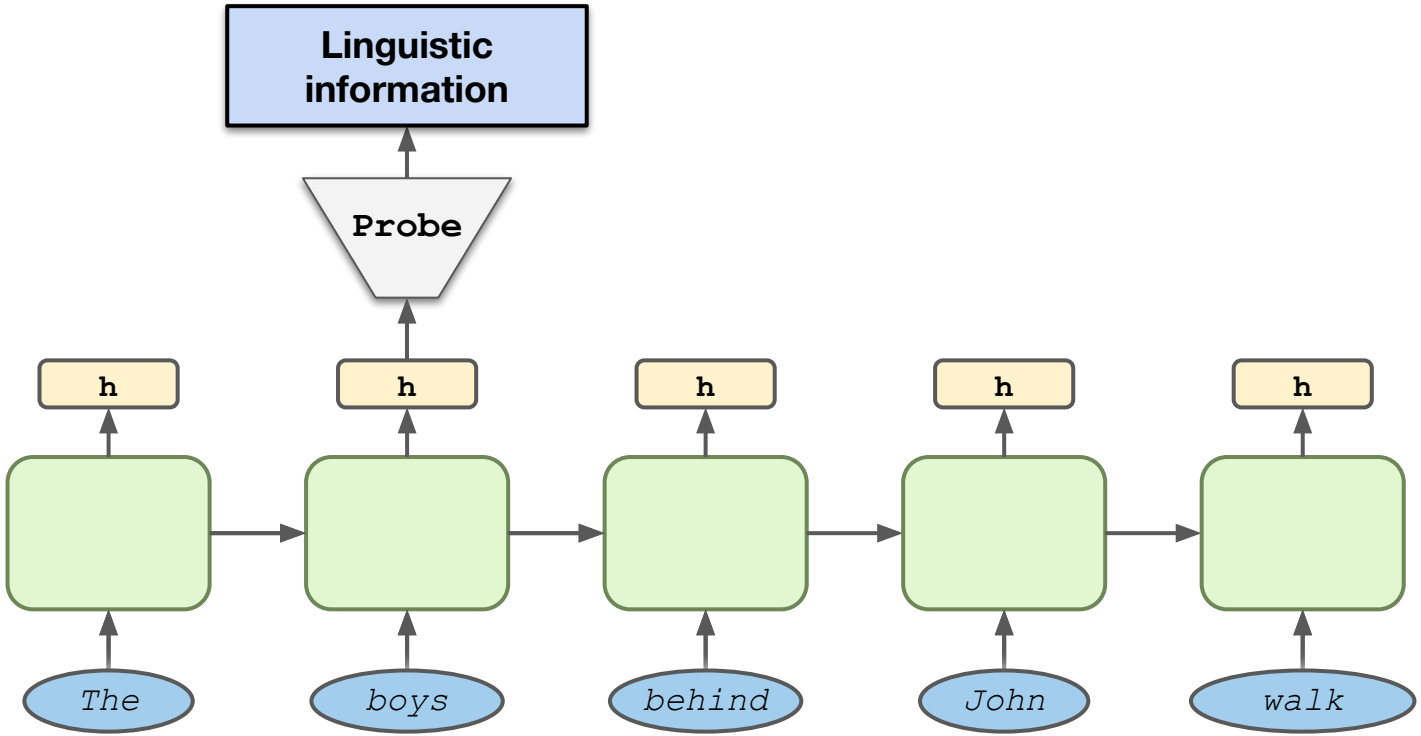
- ✗ They do not show *how* in the model representations are formed
- ✗ They give no insight into **higher-level** concepts such as *'gender'*, *'number'*, or *'part-of-speech'* class.

Instead, we can turn to **probing**, in which we train classifiers on top of model representations!

Probing


What information is **encoded** in here?
h
h
h
h
h
The
boys
behind
John
walk

Probing


Linguistic information
Probe
h
h
h
h
h
The
boys
behind
John
walk

Probing

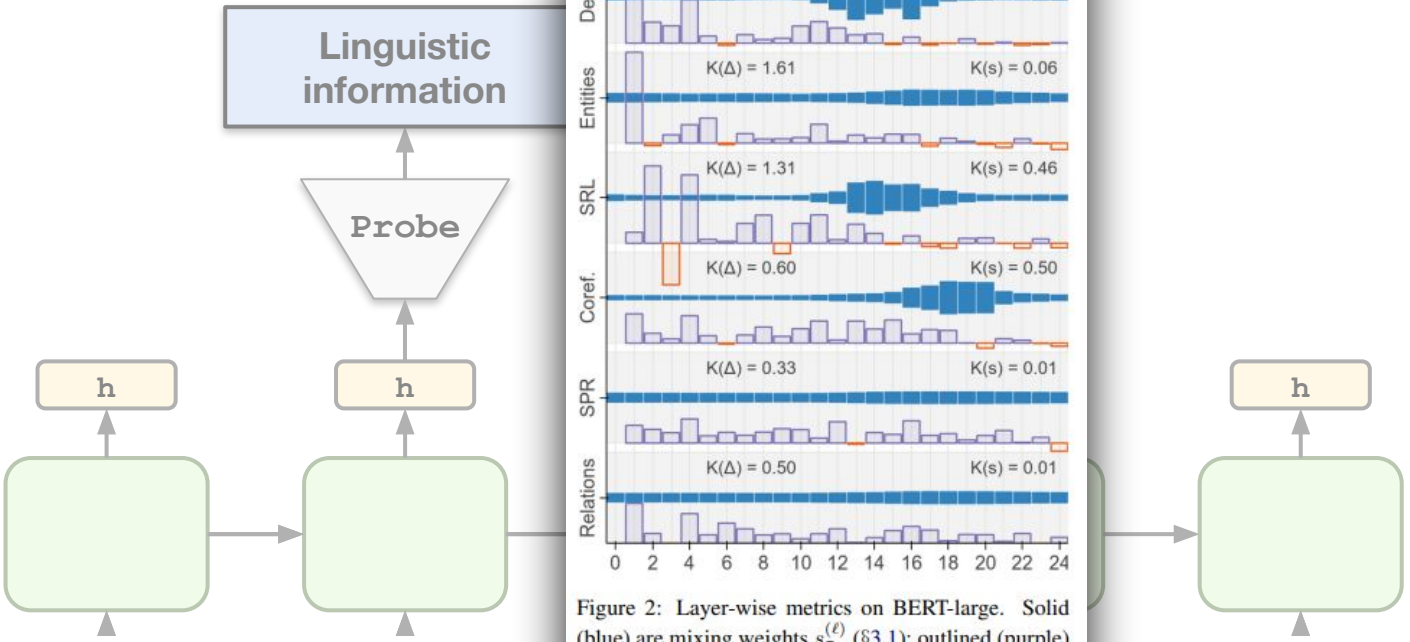


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.

BERT Rediscovers the Classical NLP Pipeline

Tenney et al. (2019)

Limitations of Probing

Probing shows us whether abstract concepts are decodable.

✗ It does not show us whether the model actually *uses* these concepts for its predictions

For this we need a **causal** methodology.

Measure the impact on model performance after **removing** a concept from the representation.

What is mechanistic interpretability?

You're not the only one asking!

Sasha Rush
@srush_nlp

I recently asked pre-PhD researchers what area they were most excited about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am interested how it came about.

Jacob Andreas @jacobandreas · Jan 23

I still don't totally understand the difference between "mechanistic" and "non-mechanistic" interpretability but it seems to be mainly a distinction of the authors' social network?

Andrew Gordon Wilson @andrewgwils · Jan 24

Did they seem to know much about it and the foundations? I've also noticed a major increase in interest in this area, and alignment, but I suspect unfortunately for many it's just trendy buzzwords.

Mark Riedl @mark_riedl · Jan 23

Mechanistic explainability doesn't require human-participant studies for evaluation. Pesky humans always being noisy and requiring IRB protocols and requiring months and months of time.

Mechanistic XAI as a term exists to differentiate from human-centered

Sarah Wiegreffe @sarahwiegreffe · Jan 24

FWIW, I gave a talk at ACL in July on this topic. The framework in the talk doesn't capture everything, but I think it gives some credence as to why the terminology might be useful.

"Two Views of LM Interpretability" (starting at 7:46):

What is mechanistic interpretability?

Mechanistic interpretability: subfield of interpretability that aims to reverse-engineer neural network behavior, mapping low-level mechanisms to higher-level human-interpretable algorithms.

It frequently uses causal interpretability techniques, and studies the sub-layer level (e.g. attention heads, MLPs, or neurons).

This contrasts with work that:

operates at the input-output level (behavioral interpretability)

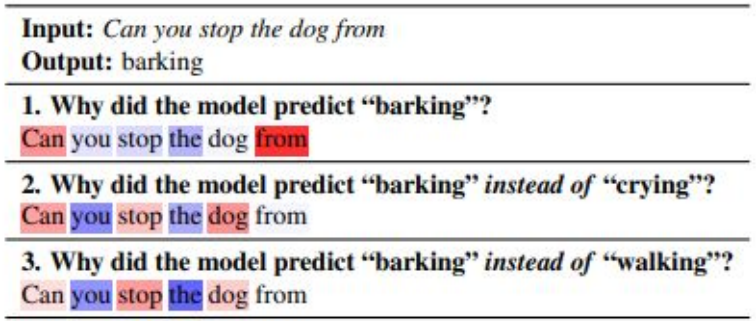
What is mechanistic interpretability?

Mechanistic interpretability: subfield of interpretability that aims to reverse-engineer neural network behavior, mapping low-level mechanisms to higher-level human-interpretable algorithms.

It frequently uses causal interpretability techniques, and studies the sub-layer level (e.g. attention heads, MLPs, or neurons).

This contrasts with work that:

finds important input tokens (input attribution)



Yin and Neubig, 2022

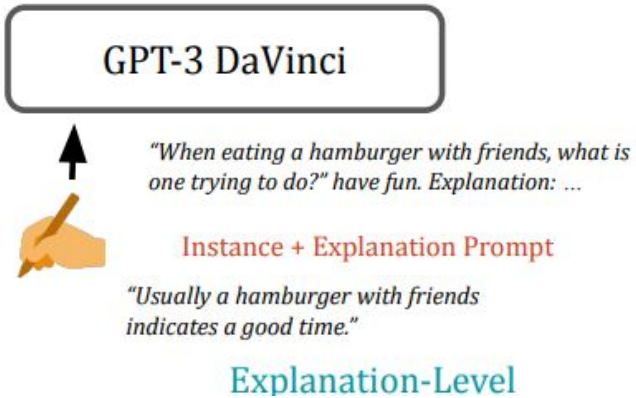
What is mechanistic interpretability?

Mechanistic interpretability: subfield of interpretability that aims to reverse-engineer neural network behavior, mapping low-level mechanisms to higher-level human-interpretable algorithms.

It frequently uses causal interpretability techniques, and studies the sub-layer level (e.g. attention heads, MLPs, or neurons).

This contrasts with work that:

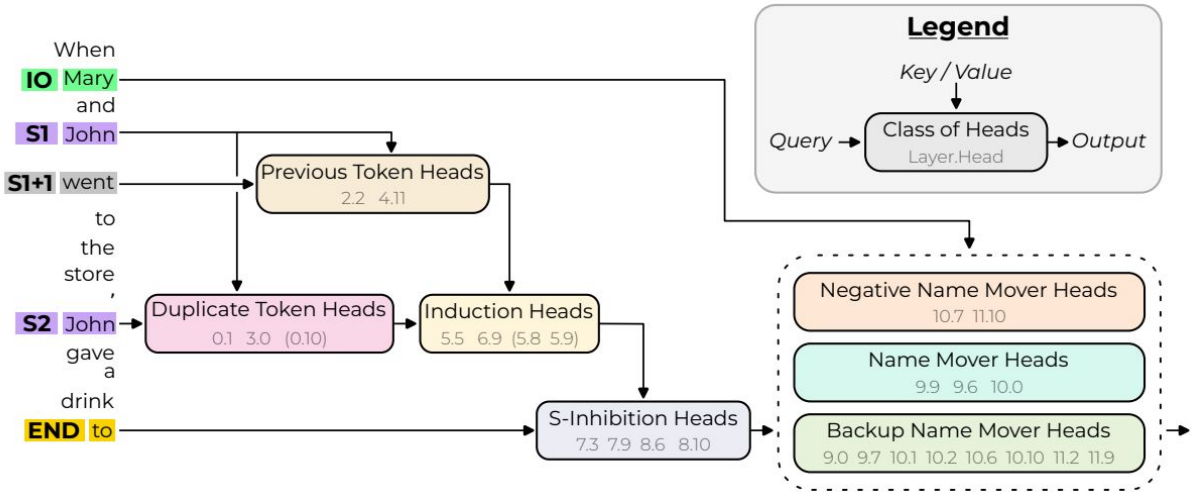
is human/user-centered (HCXAI)



Wiegreffe et al., 2022

What is mechanistic interpretability?

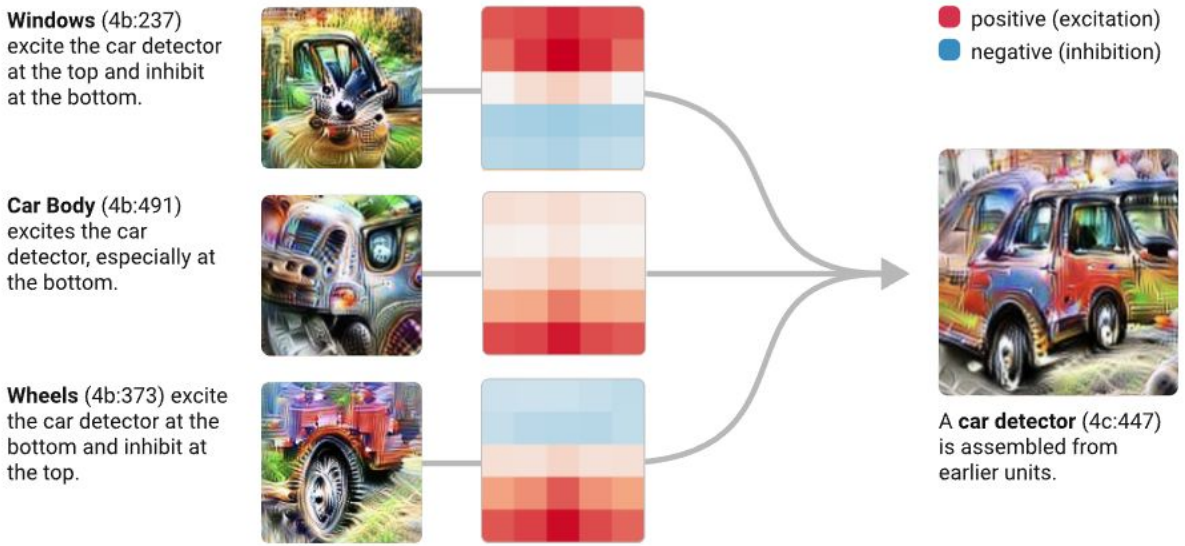
There are still many ways to reverse-engineer model behavior! But today, I'll focus on **circuits**: one particular framework for characterizing model behaviors.



Wang et al., 2023

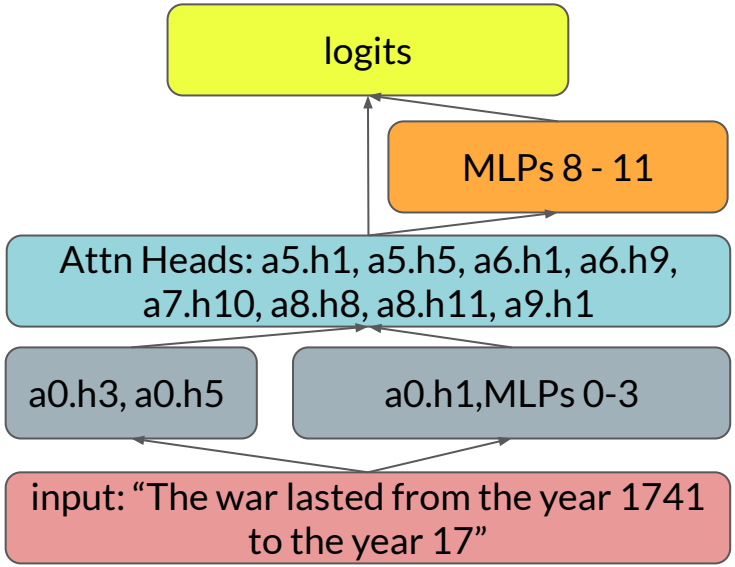
What are circuits?

Olah et al. (2020) defined circuits as “sub-graphs of the network, consisting a set of tightly linked features and the weights between them”



What are *transformer* circuits?

Transformer circuits localize and characterize transformer LM behavior in a (small) set of components of the model.



Upweight years > 41

Identify and upweight 41

Subject enrichment

Hanna et al., 2023

Circuits

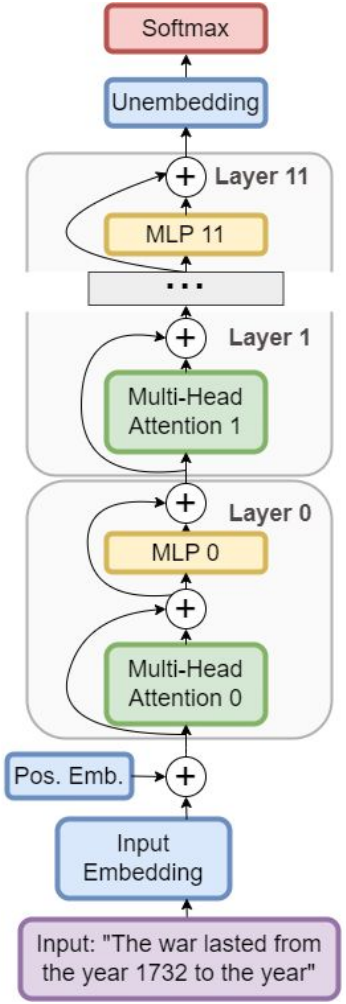
Circuit: minimal computational subgraph of a model that is faithful to model performance on a given task

- **Minimal computational subgraph**: minimal set of model nodes and edges
- **Task**: collection of **inputs** and **expected outputs**, measured by some **loss**.
- **Faithful**: loss remains the same when all non-circuit edges are *ablated*

But what does that mean?

What computational subgraph? The transformer LM architecture

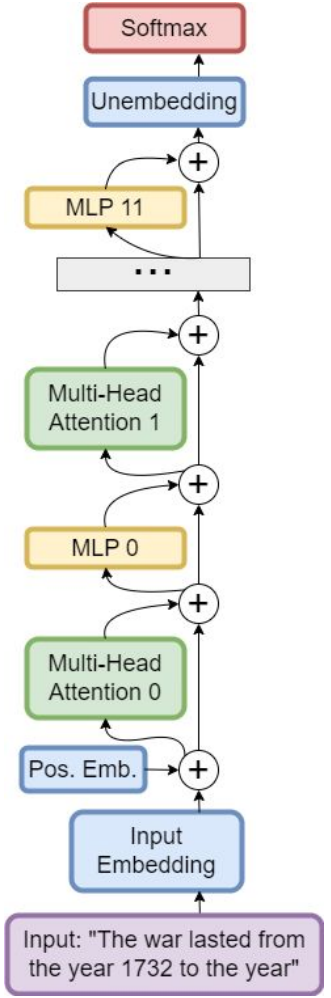
This is a traditional diagram of an autoregressive decoder-only LM.



The Residual Stream View

Centering the residuals reveals:

- Every component reads from and writes to the residual stream!
- Every component's input is the sum of the outputs of the components that came before

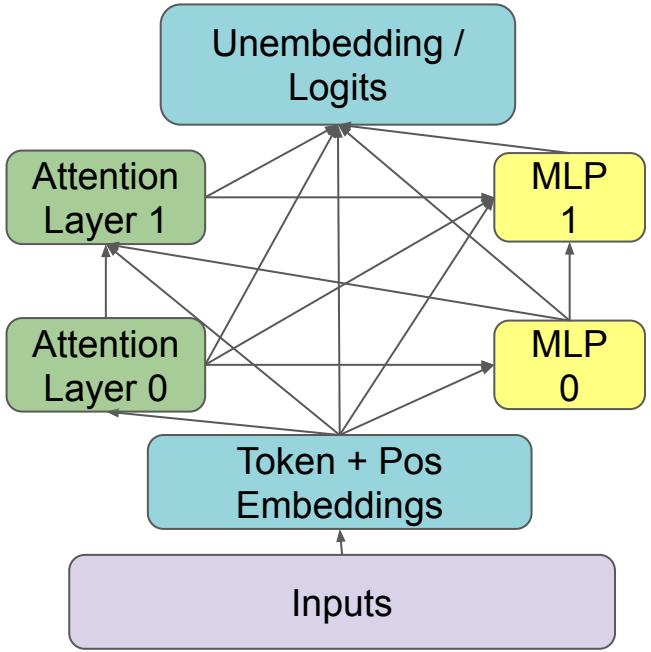


Elhage et al. (2021)

Computational Graph

For our circuit, we want the **minimal subgraph** that is faithful to model behavior. This view lets us specify the specific node-node interactions that count.

Note: we could have chosen other levels of granularity for this graph!



Task: Greater-Than

A **task** consists of:

Inputs: “The war lasted from 1741 to 17”
Expected outputs: a 2-digit number greater than 41
Metric: $\sum_{y>41} p(y) - \sum_{y<=41} p(y)$

Tasks should be solvable by your model, and evaluable in one forward pass.

Average Metric Value: 0.817

For circuit-finding, we also need corrupted inputs.

Corrupted inputs: “The war lasted from 1701 to 17”

Task: Subject-Verb Agreement

A **task** consists of:

Input: “The keys on the cabinet”
Expected output: a verb that agrees with the subject (“keys”)
Metric: $\sum_{y,\text{agree}(y, \text{“keys”})} p(y) - \sum_{y,\text{disagree}(y, \text{“keys”})} p(y)$

Tasks should be solvable by your model, and evaluable in one forward pass.

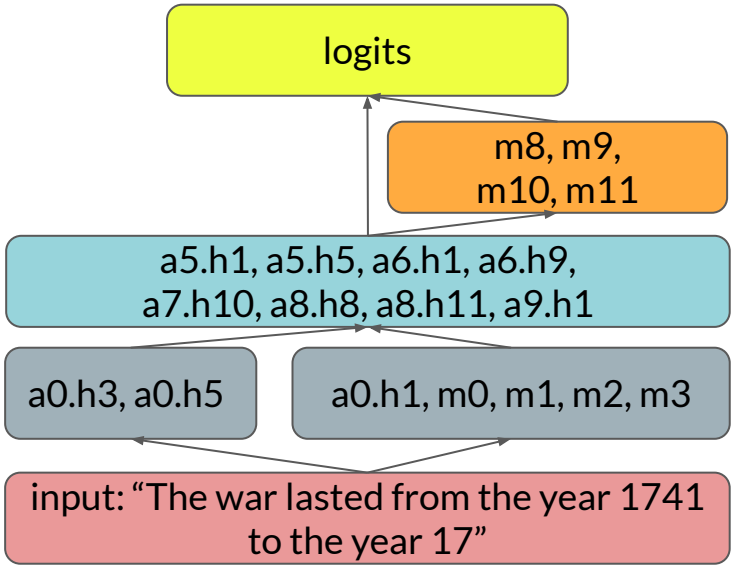
Average Metric Value: 0.351

For circuit-finding, we also need corrupted inputs.

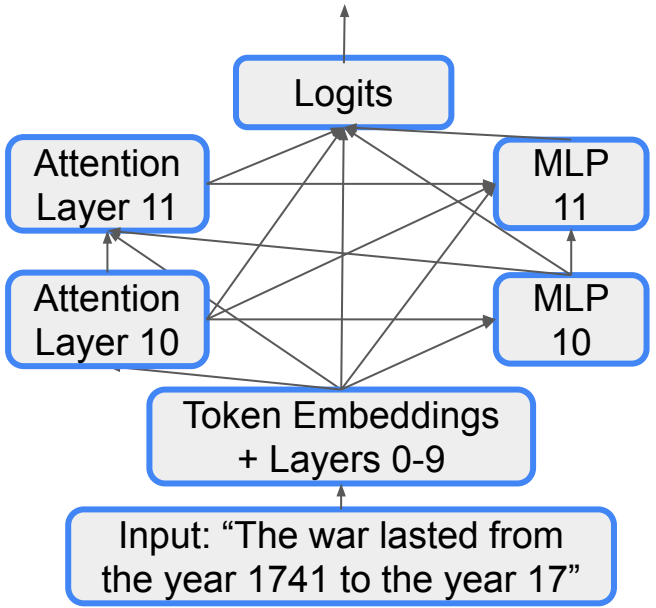
Corrupted Input: “The key on the cabinet”

Faithfulness

If a circuit is faithful to model behavior, we can ablate all nodes outside the circuit, with little to no behavior change!

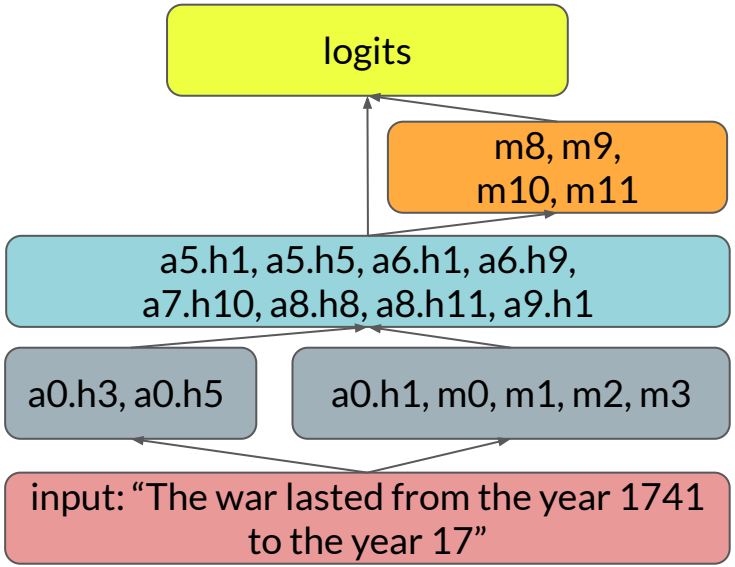


Metric: 0.817

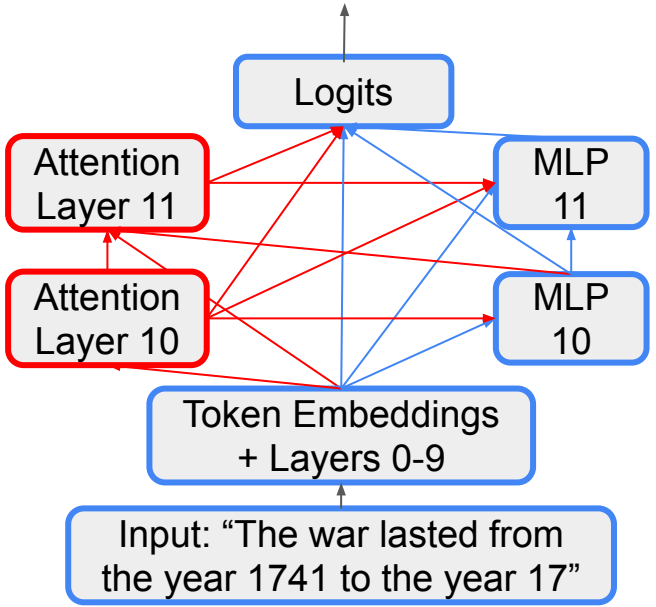


Faithfulness

If a circuit is faithful to model behavior, we can ablate all nodes outside the circuit, with little to no behavior change!



Metric: 0.78



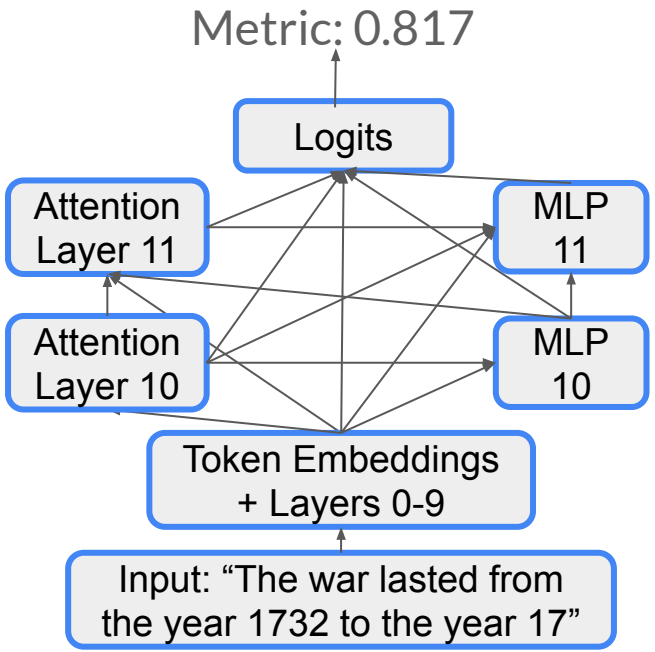
Finding Circuit Structure

Finding important nodes

We want to find nodes / edges that are important for a task.

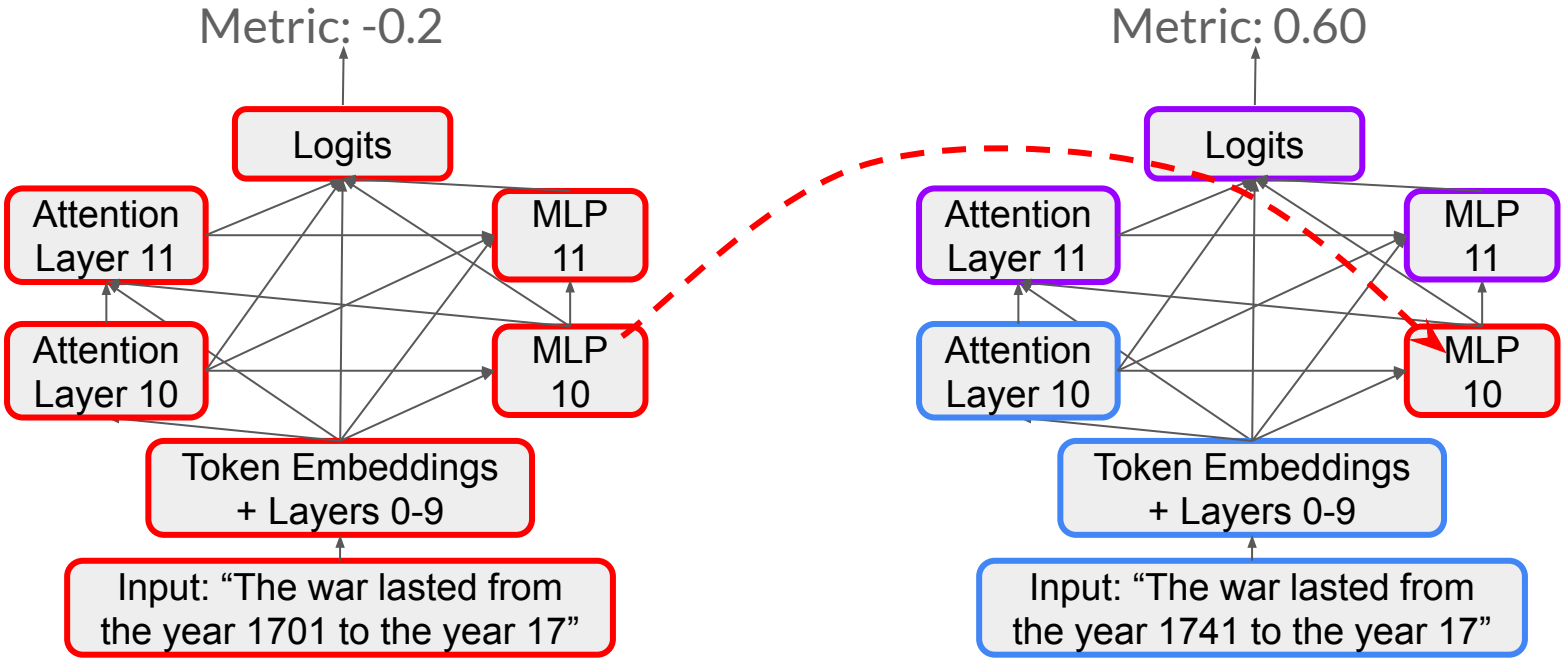
Core Idea: Important nodes / edges can't be ablated without hurting model performance.

But how do we ablate? Don't use zero ablations!



Activation Patching

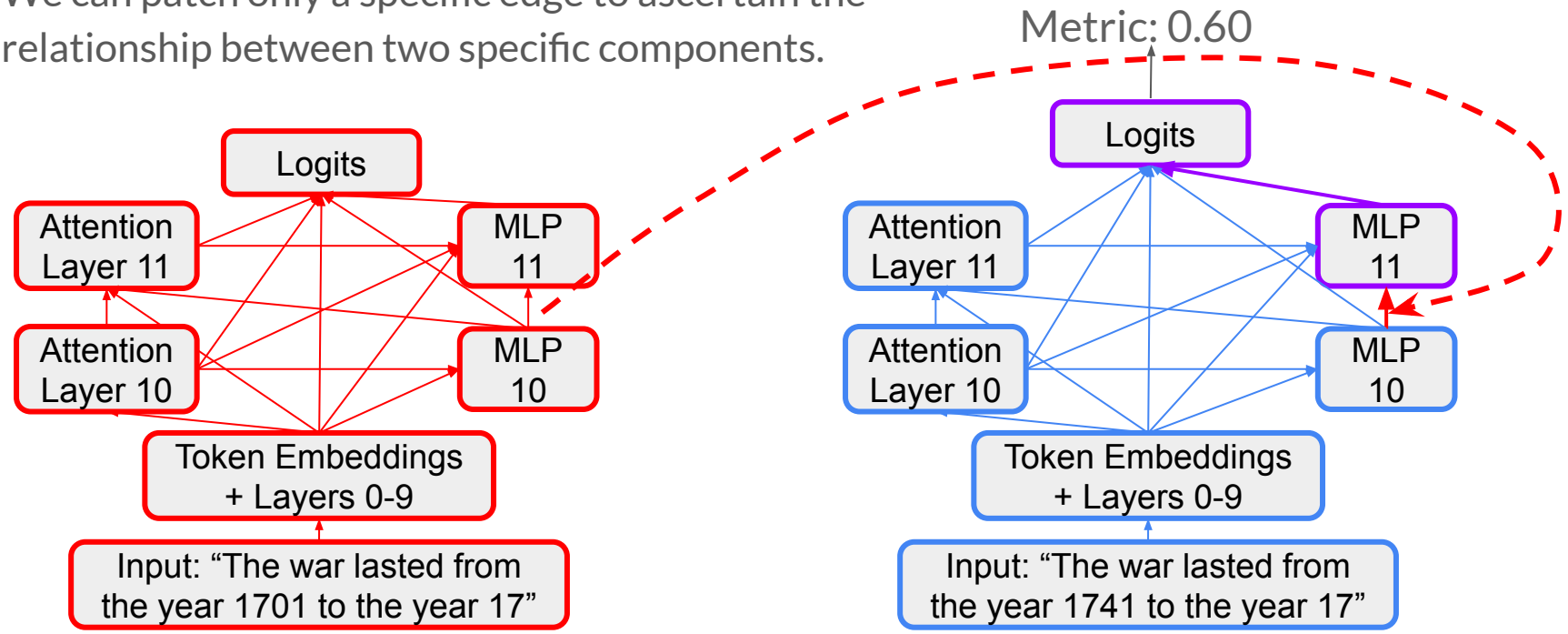
Replace a component's activation on one example, with an activation from another!



Activation patching (Vig et al., 2020; Geiger et al., 2020) predates LM circuits work

Edge Patching

We can patch only a specific edge to ascertain the relationship between two specific components.



Wang et al. (2022), Goldowsky-Dill et al. (2023)

How to perform patching?

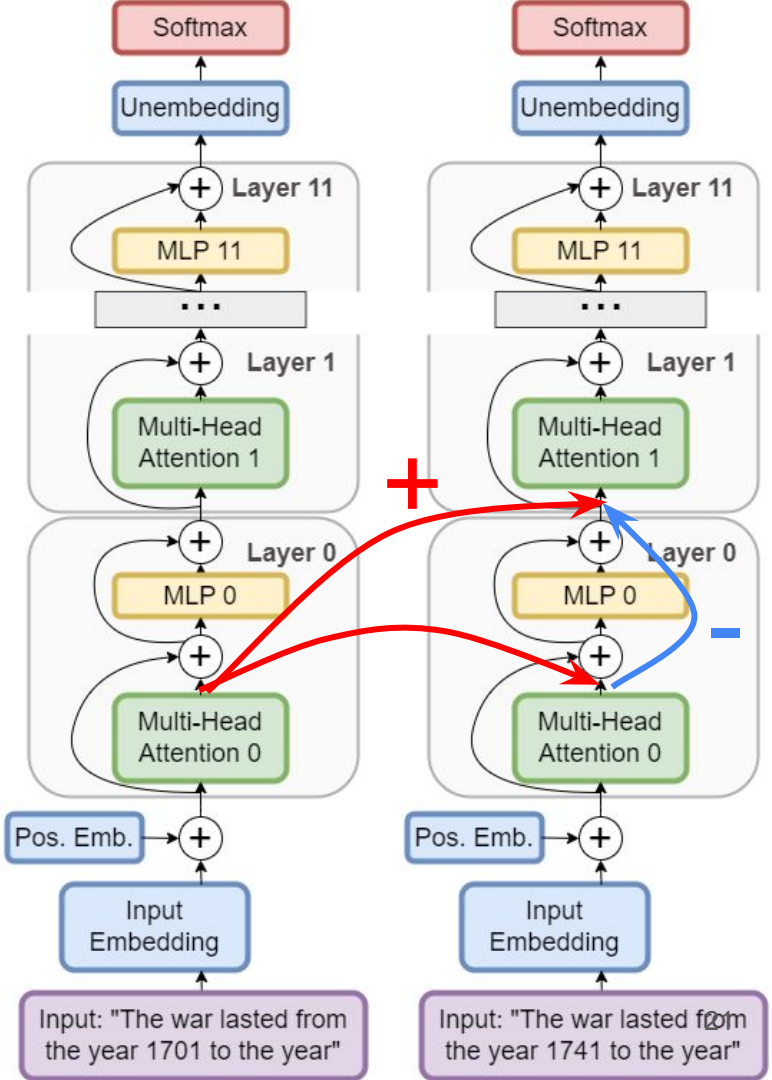
Node-level patching:

Replace the output of the node (e.g. Attn 0) with its output on another input!

Edge-level patching:

Exploit the linearity of the residual stream! Say we're patching the edge Attn0->Attn1.

- Take the input to Attn1
- Subtract the output of Attn0 on normal input
- Add in the output of Attn0 on corrupted input

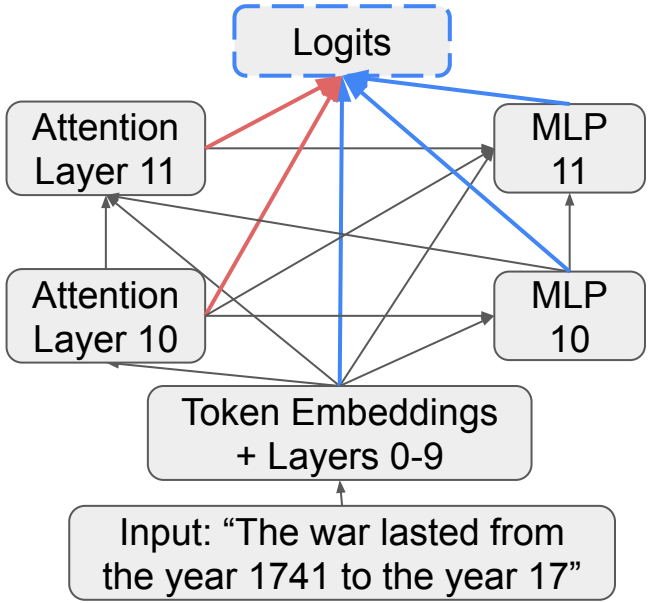


Circuit Finding: Activation Patching

How can we use patching to find an entire circuit?

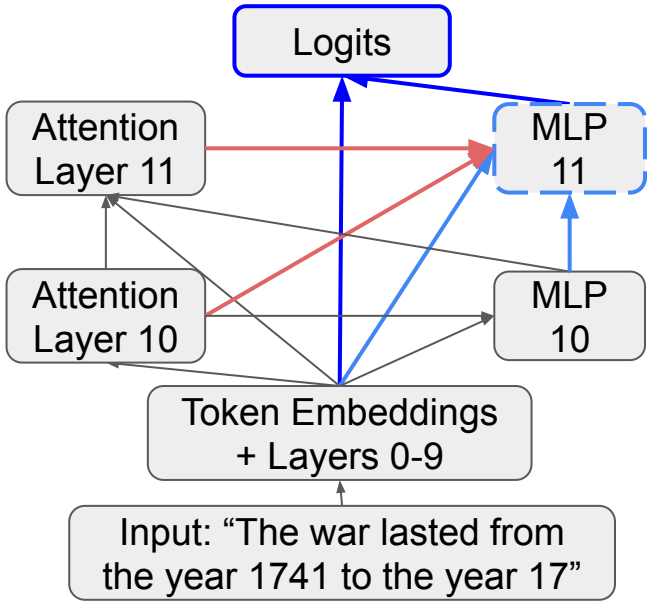
One approach: iteratively patch to find important nodes / edges.

First, find the nodes connected directly to the logits…



Circuit Finding: Activation Patching

Then find the nodes directly connected to those nodes, and then…

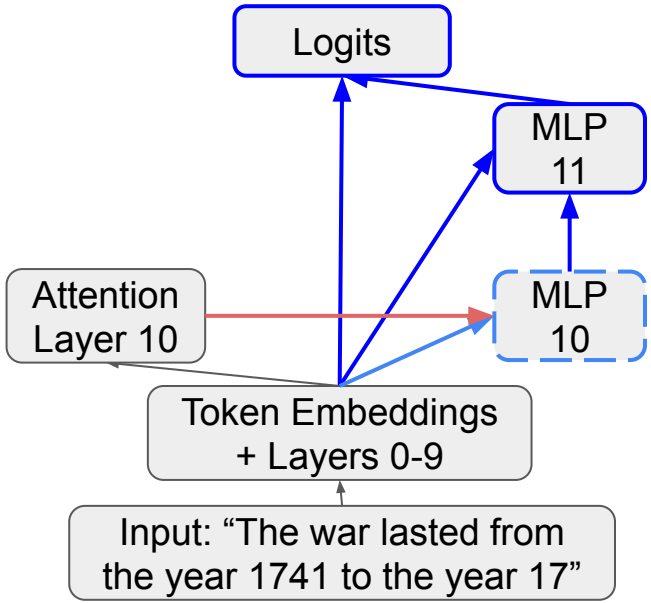


Circuit Finding: Activation Patching

Once we've reached the embeddings, we've found the circuit.

Techniques like automatic circuit discovery (ACDC, Conmy et al. (2023)) use similar approaches.

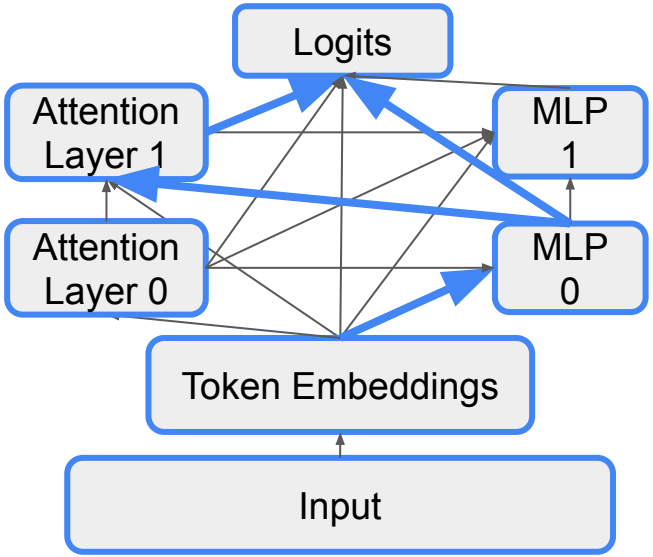
This is very slow! The solution: approximations to activation patching



Proving Circuit Faithfulness

How to prove circuit faithfulness?

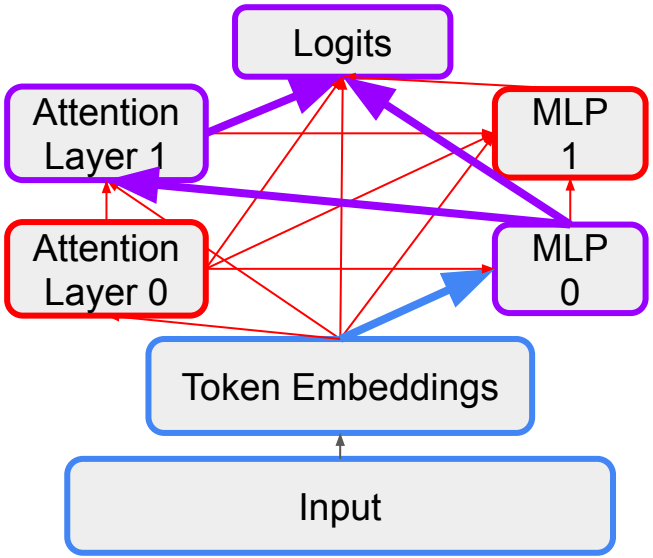
Perform another patching experiment!
Corrupt everything but your circuit.



Proving Circuit Faithfulness

A faithful circuit will have task performance close to that of the whole model! See also:

- **Completeness**: Have we discovered all components, even negative ones?
- **Minimality**: Are all components in the circuit necessary?



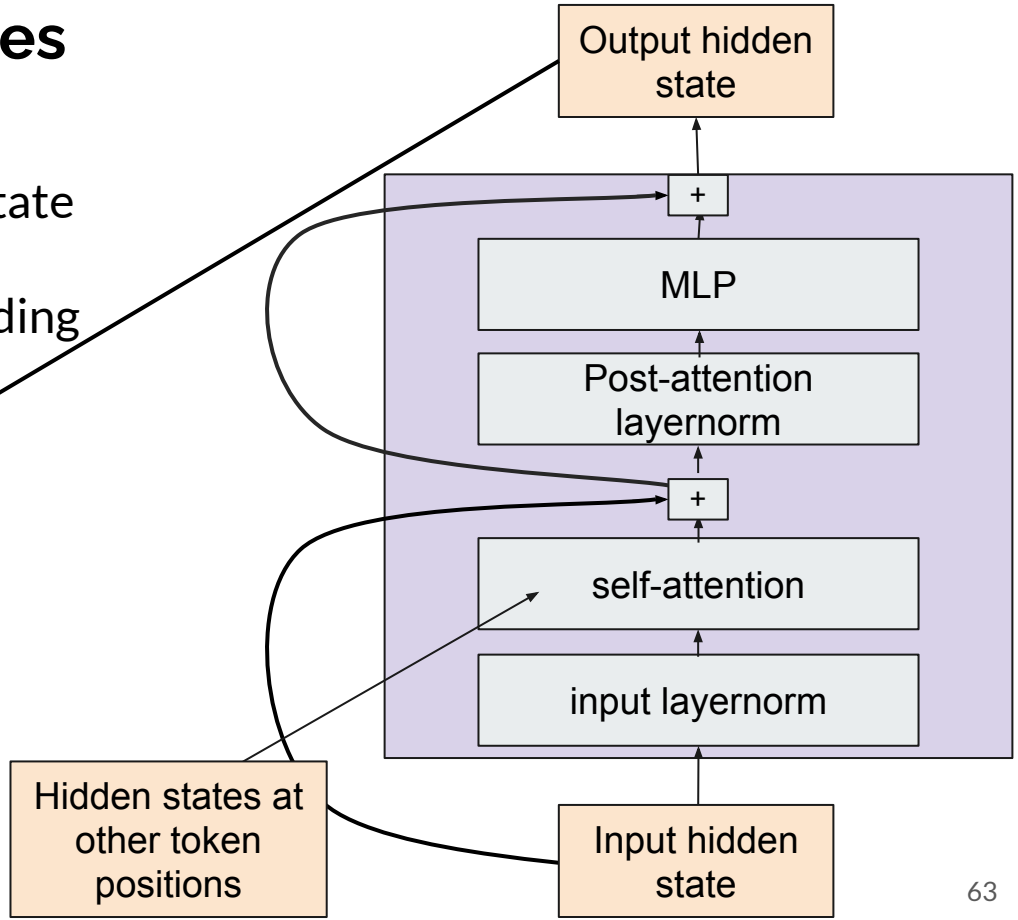
[Nostalgebraist 2020, Geva et al. 2022]

Vocabulary Projection on Transformer Hidden States

- Propose to project each hidden state to the space of probabilities over vocab tokens using the unembedding matrix

$$\mathbf{y} = \text{Softmax}(W_U \cdot \text{LN}(\mathbf{x}_L))$$

Final hidden state - **replace with any d-dimensional hidden state from the network.**



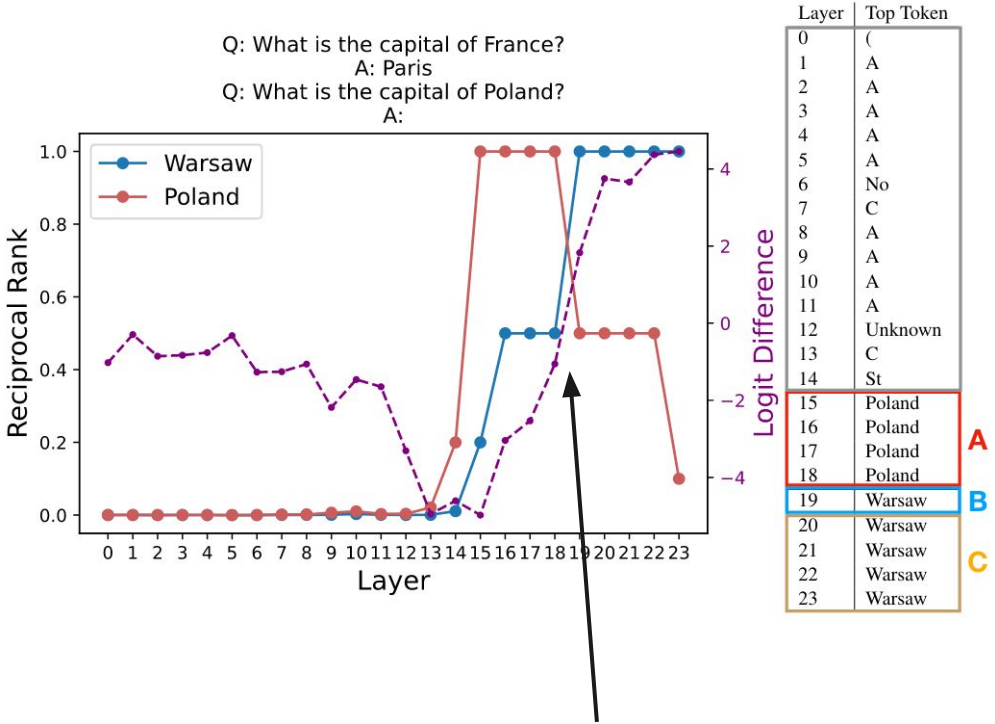
Vocabulary Projection on Transformer Hidden States

		Concept	Sub-update top-scoring tokens
GPT2	$\mathbf{v}^{3}_{1018}$	Measurement semantic	kg, percent, spread, total, yards, pounds, hours
	$\mathbf{v}^{8}_{1900}$	WH-relativizers syntactic	which, whose, Which, whom, where, who, wherein
	$\mathbf{v}^{11}_{2601}$	Food and drinks semantic	drinks, coffee, tea, soda, burgers, bar, sushi
WIKILM	$\mathbf{v}^{1}_{1}$	Pronouns syntactic	Her, She, Their, her, she, They, their, they, His
	$\mathbf{v}^{6}_{3025}$	Adverbs syntactic	largely, rapidly, effectively, previously, normally
	$\mathbf{v}^{13}_{3516}$	Groups of people semantic	policymakers, geneticists, ancestries, Ohioans

Table 1: Example value vectors in GPT2 and WIKILM promoting human-interpretable concepts.

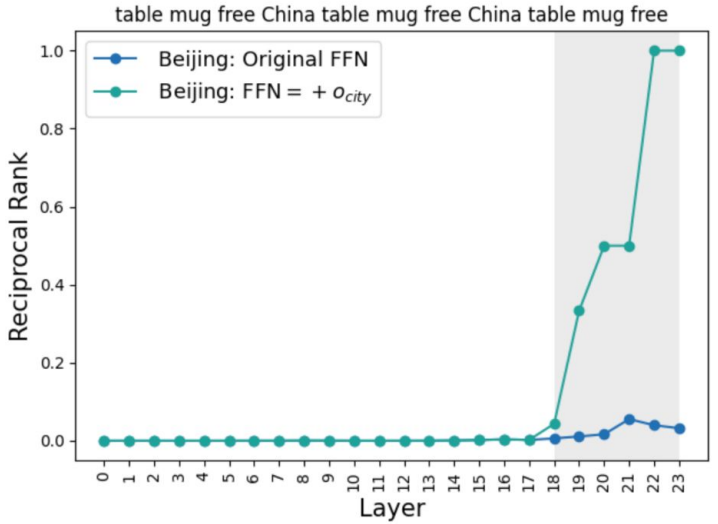
[Merullo et al 2024]

Vocabulary Projection on Transformer Hidden States



Layer	Top Token	
0	(	
1	A	
2	A	
3	A	
4	A	
5	A	
6	No	
7	C	
8	A	
9	A	
10	A	
11	A	
12	Unknown	
13	C	
14	St	
15	Poland	A
16	Poland	A
17	Poland	A
18	Poland	A
19	Warsaw	B
20	Warsaw	C
21	Warsaw	C
22	Warsaw	C
23	Warsaw	C

Validated with causal intervention:



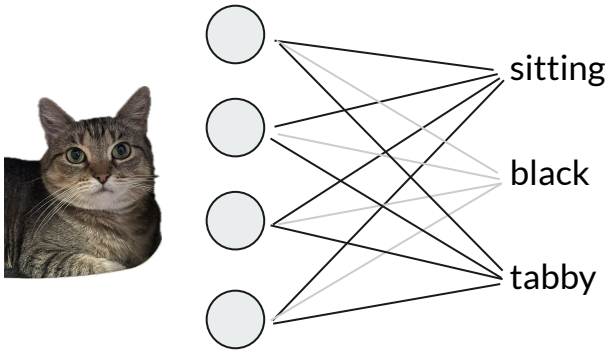
Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:
many neurons respond to mixtures of concepts

[Anthropic, 2023]

Linear Combinations of Neurons as Concepts

Individual neuron-level interpretations are typically not precise:
many neurons respond to mixtures of concepts



Hypothesis

Neurons together (as opposed to individual neurons) respond to concepts

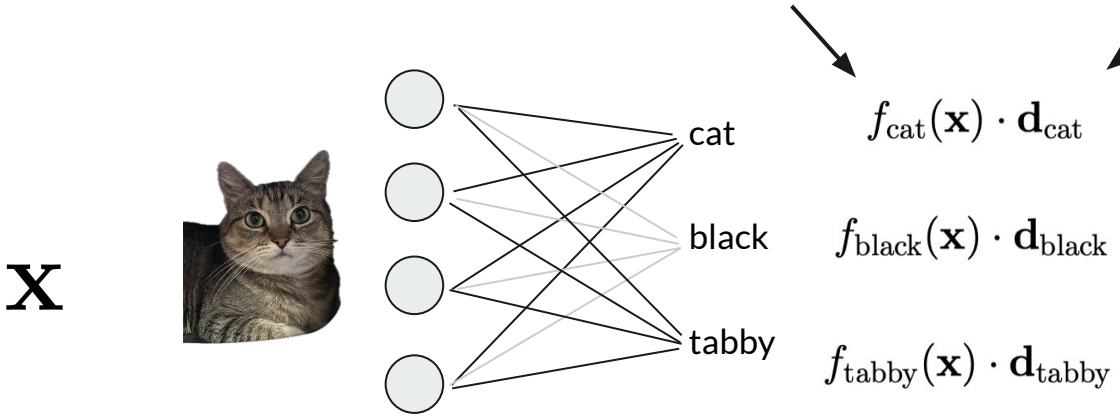
Neuron activations can be decomposed into linear combinations of concept directions (called features)

[Anthropic, 2023]

Decomposing Activations

scalar:
strength of the feature

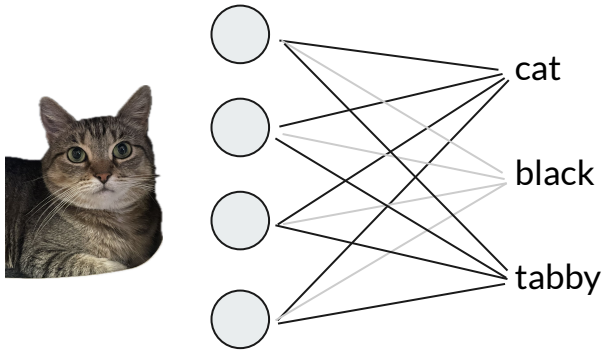
vector:
direction of the feature



$\mathbf{x}$ —decompose→ $f_{\text{cat}}(\mathbf{x}) \cdot \mathbf{d}_{\text{cat}} + f_{\text{black}}(\mathbf{x}) \cdot \mathbf{d}_{\text{black}} + f_{\text{tabby}}(\mathbf{x}) \cdot \mathbf{d}_{\text{tabby}}$

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{X}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x}) > 0$)

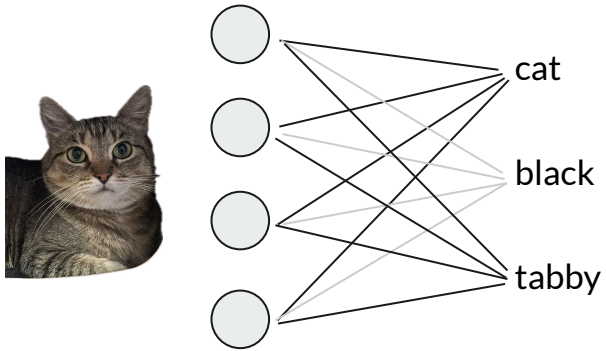
$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\mathbf{d}_c$: decoder parameters

[Anthropic, 2023]

Decomposing Activations with Sparse Autoencoders



Sparsity: for $\mathbf{x}$, we expect only a small number of feature **c** is activated ($f_c(\mathbf{x}) > 0$)

$$\mathbf{x} \approx \mathbf{b} + \sum_c f_c(\mathbf{x})\mathbf{d}_c$$

using **Sparse Autoencoders** to find decompositions

$f_c(\mathbf{x})$: encoder parameters $\quad \mathbf{d}_c$: decoder parameters

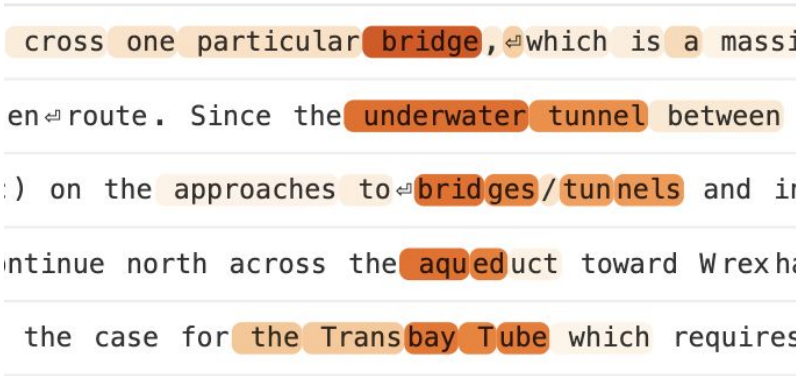
reconstruction loss **sparsity**

Loss: $$\mathbb{E}_x \left[\|x - \hat{x}\|_2^2 + \lambda \sum_c f_c(x) \right]$$

[Anthropic, 2023]

Notes on Learned Features

Example: 1M/3 **Transit infrastructure**



Intervening on feature activations has an influence on behavior

Default output gives reasonable navigation directions

```
Human: What's the best way to get to the grocery
store down the street? Be brief.

Assistant: 1. Walk.
```

with Transit infrastructure F#1M/3
clamped to 5× its max
It confabulates a bridge

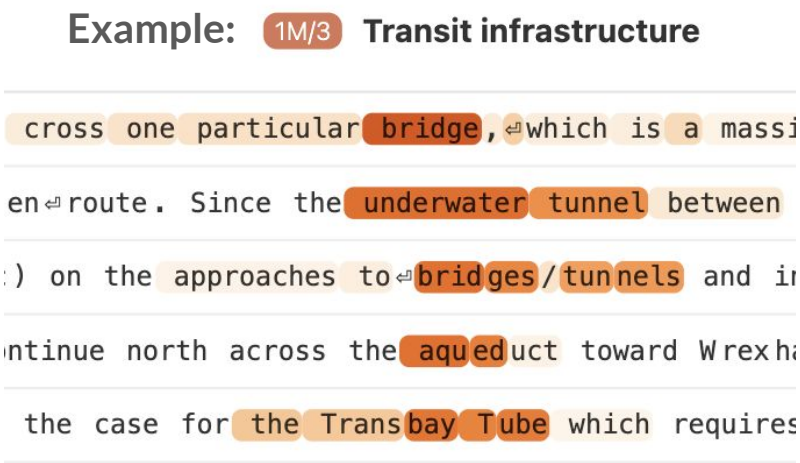
```
Human: What's the best way to get to the grocery
store down the street? Be brief.

Assistant: 1. Walk across the bridge.
```

[Anthropic, 2024]

[Anthropic, 2023]

Notes on Learned Features



Intervening on feature activations has an influence on behavior

Feature activations are more specific than neurons

- “upon manual inspection of a random sample of 50 neurons and features each, the neurons appear significantly less interpretable than the features, typically activating in multiple unrelated contexts..”

[Anthropic, 2024]

[Anthropic, 2023]

Concluding Remarks