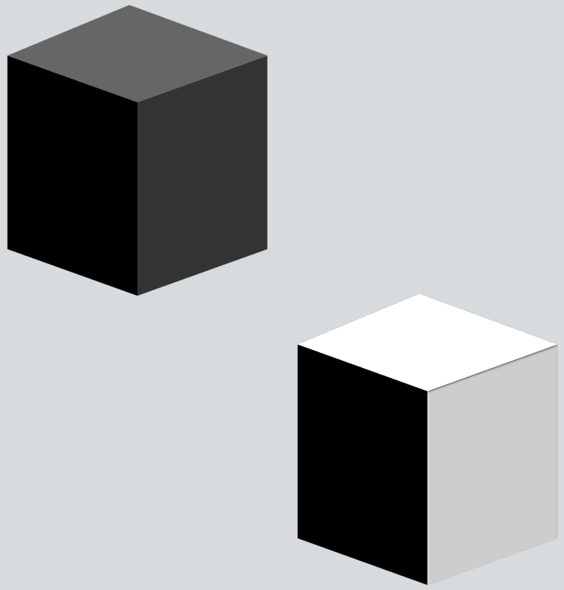
White Box & Black Box

Two Perspectives on Explainable Natural Language Processing

May 16 2024 | TaCoS



Dr. Simon Ostermann

Efficient and Explainable NLP Group

Multilinguality and Language Technology Lab, DFKI Saarbrücken

Outline

- Explainable AI – A Quick Overview
- Black Box Explainable NLP: Dialogue-based Explanations
- White Box Explainable NLP: Feature Textualization

BIG KUDOS to my colleagues Tanja Bäumel and Nils Feldhus for their work and for making available their slides to me!

Explainable Artificial Intelligence

What is Explainable AI/NLP?

What is Explainable AI/NLP?

Technology that makes it possible for humans to understand the reasoning behind the behaviour of an AI system.

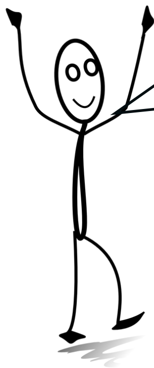
What is Explainable AI/NLP?

Technology that makes it possible for humans to understand the reasoning behind the behaviour of an AI system.

Sometimes, the technology is inherently interpretable, sometimes we need „helpers“. Both can be considered XAI.

An example

A million years ago in 201X ...



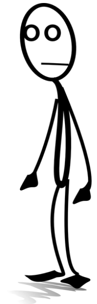
Dear AI, may I please get a loan?

AI

- 0.2 x number of credit cards
- 0.3 x age
+ 1.4 x current bank balance
+ 0.8 x monthly income

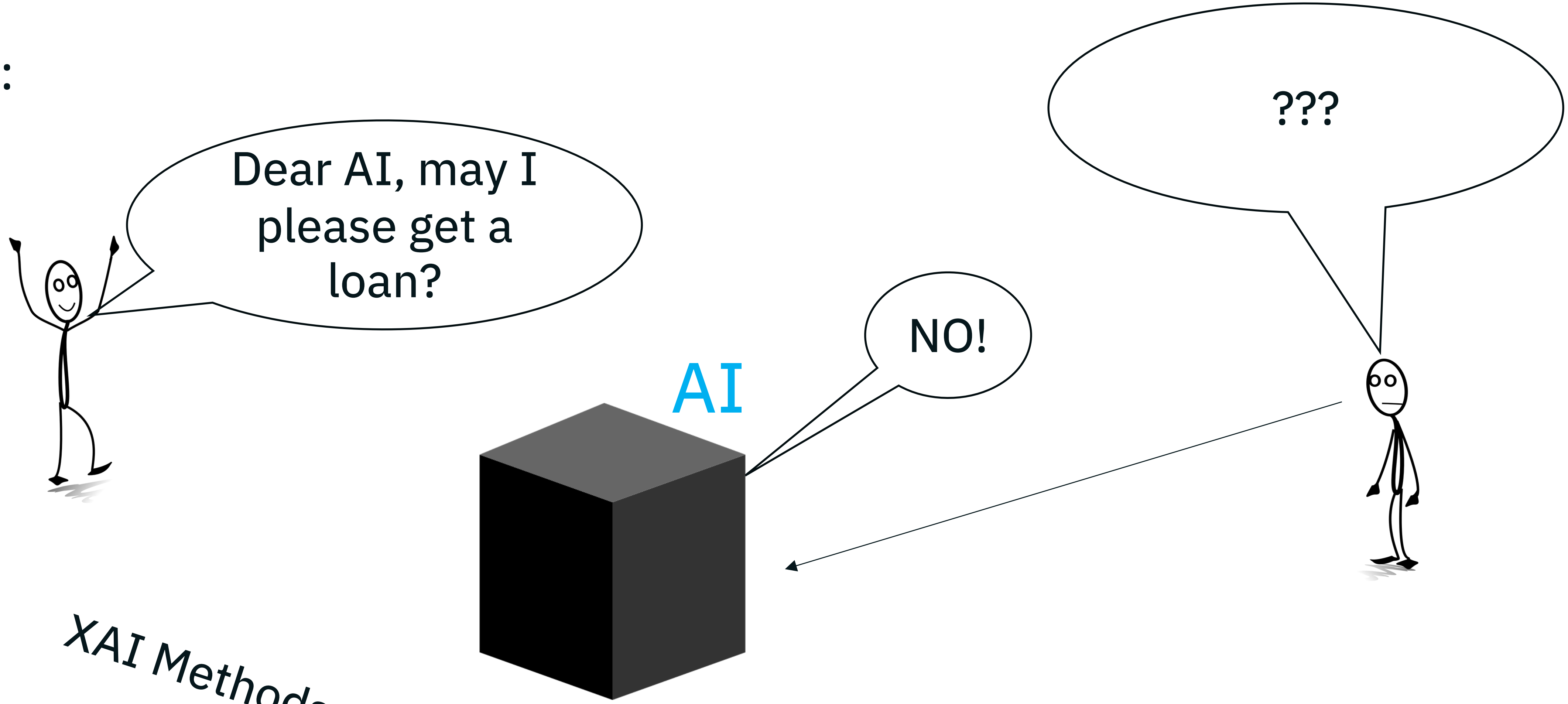
NO!

Probably due to my low bank balance...



An example

Nowadays:



XAI Methods to the help?

Motivation

Why and when should AI be explainable?

Motivation

Advantages of understanding a model:

Detecting bias/
Fairness

Debugging

Safety

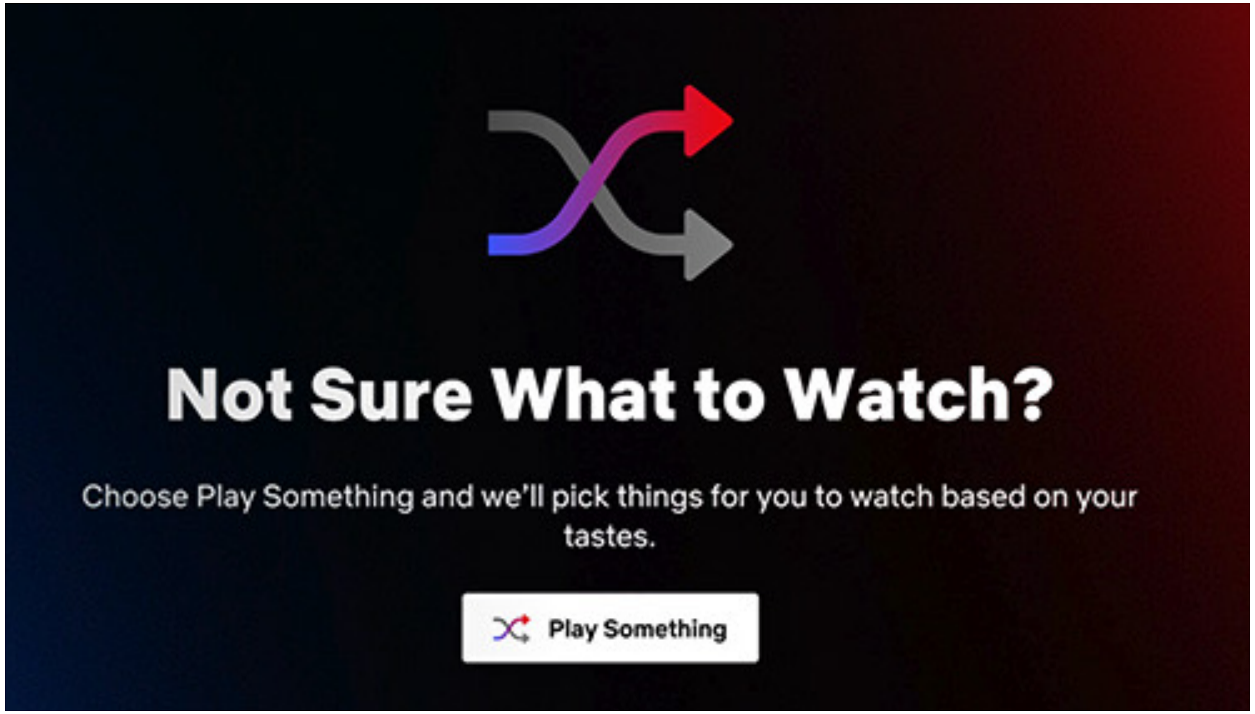
Human
curiosity

Social acceptance

Establish trust

Motivation

Not everything is high stakes!



Wait a sec...

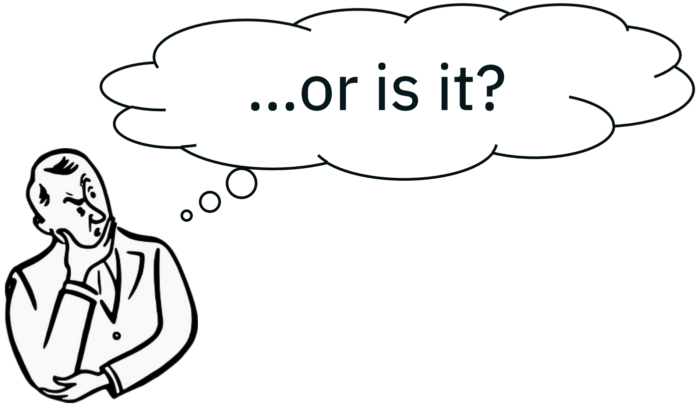
Why don‘t we simply trust high accuracy models?!

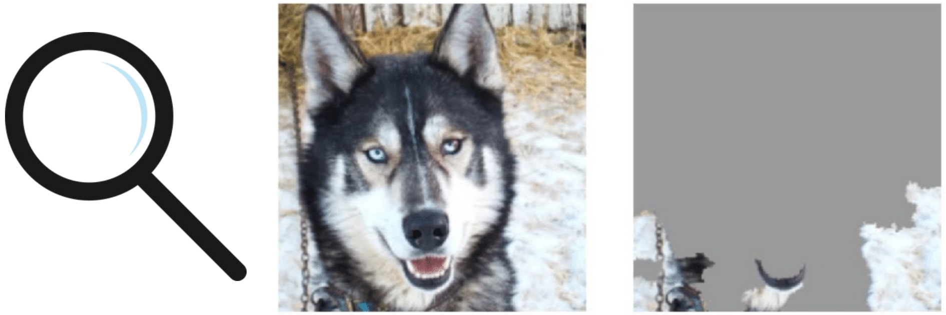
- Real data ≠ test data
- Correct decision for the wrong reasons
- Accuracy not the only criterion (fairness, safety, ...)

The famous husky example

The famous husky example

80% Accuracy
→ pretty decent...





Snow detector,
100% Accuracy

XAI Methods

less information → more information

post-hoc | intrinsic

lower abstraction ↓ higher abstraction

	black-box	dataset	gradient	embeddings	white-box	model specific
local explanation						
input features	SHAP § A.2	LIME § 6.2, Anchors § A.3	Gradient § 6.1, IG § A.1			Attention
adversarial examples	SEA$^{\mathcal{M}}$ § B.1		HotFlip § 7.1			
influential examples		Influence Functions$^{\mathcal{H}}$ § 8.1 TracIn$^{\mathcal{C}}$ § 8.3		Representer Pointers† § 8.2		Prototype Networks
counter-factuals	Polyjuice$^{\mathcal{M},\mathcal{D}}$ § C.1	MiCE$^{\mathcal{M}}$ § 9.1				
natural language	CAGE$^{\mathcal{M},\mathcal{D}}$ § 10.1					GEF$^{\mathcal{D}}$, NILE$^{\mathcal{D}}$
class explanation						
concepts					NIE$^{\mathcal{D}}$ § 11.1	
global explanation						
vocabulary				Project § 12.1, Rotate § 12.2		
ensemble	SP-LIME § 13.1					
linguistic information	Behavioral Probes$^{\mathcal{D}}$ § 14.1			Structural Probes$^{\mathcal{D}}$ § 14.2	Structural Probes$^{\mathcal{D}}$ § 14.2	Auxiliary Task$^{\mathcal{D}}$
rules	SEAR$^{\mathcal{M}}$ § 15.1		Compositional Explanations of Neurons† § D.1			

Table from Madsen et al. (2022): "Post-hoc Interpretability for Neural NLP: A Survey"

A classical view

Intrinsically interpretable AI ⟷ Black Box XAI

- Classical ML models were interpretable: Regression, Feature-based, etc.
- Modern models are black boxes often

… or are they? We have access to all parameters! (sometimes)

Blackbox vs Whitebox XAI

White Box XAI	Black Box XAI
Interpret model components and insides of the model	Interpret model behaviour or representations generated
Access to the Model Parameters necessary	Access to the Model Parameters not always necessary
Target Group: Research, AI-Developers	Target Group: End users, AI users

Blackbox vs Whitebox XAI

White Box XAI ⟷ Black Box XAI

Let’s dive into two examples!

Black Box XAI

InterroLang

Exploring NLP Models and Datasets through Dialogue-based Explanations

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, Sebastian Möller

Dialogue-based explanations?

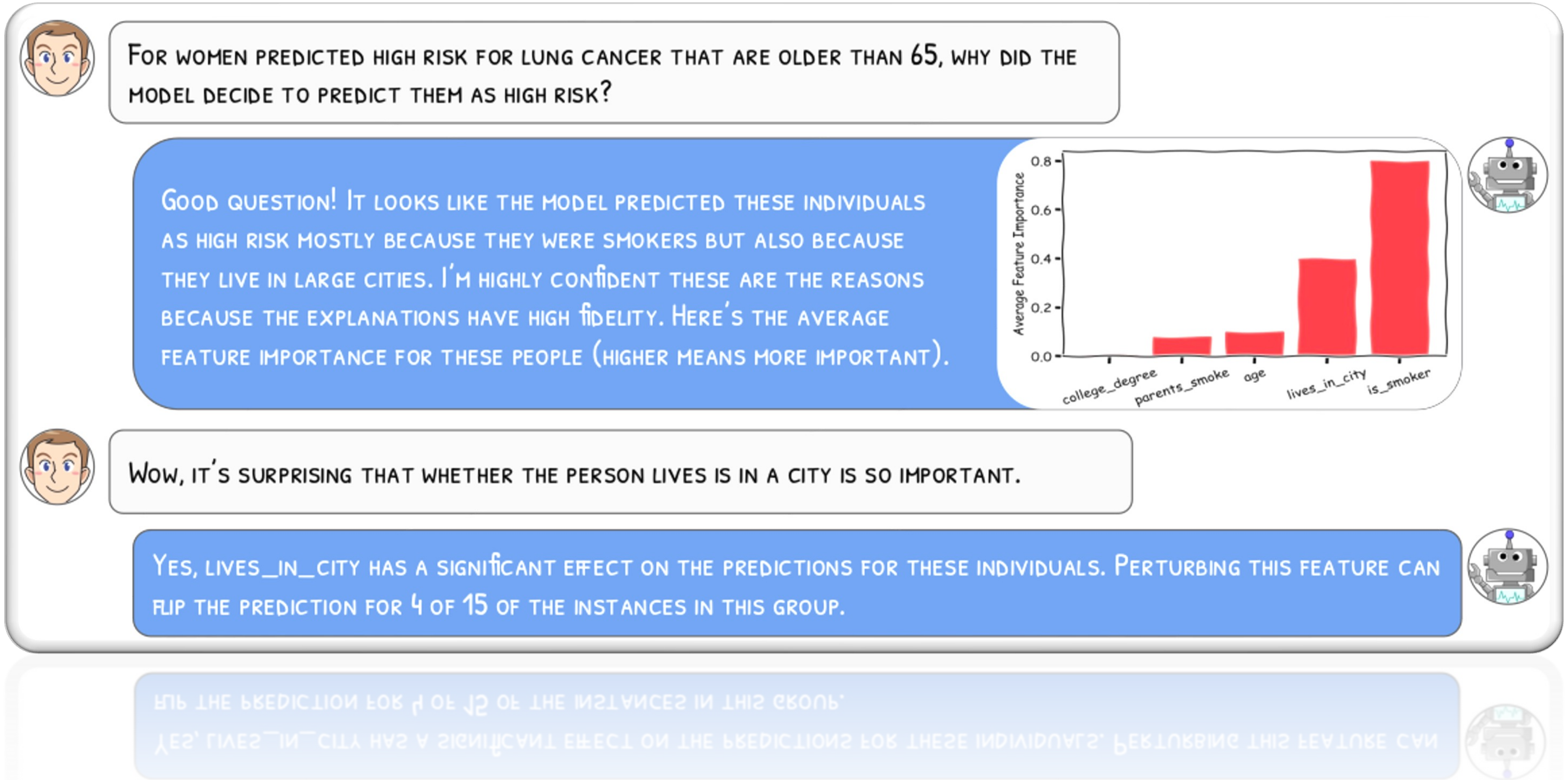
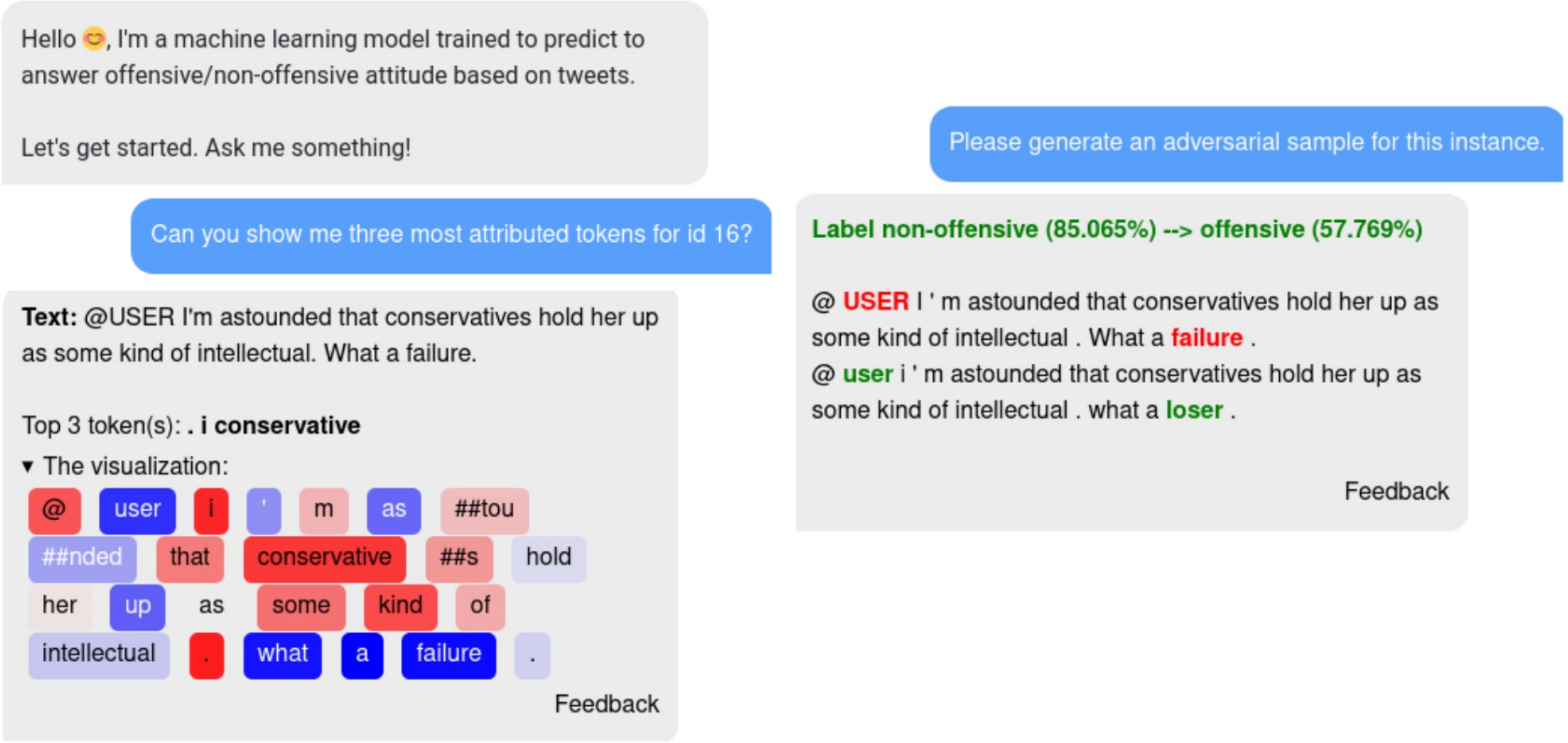


Figure from Lakkaraju et al. (2022): "Rethinking Explainability as a Dialogue: A Practitioner's Perspective"

Dialogue-based explanations!

- Interactive conversational interface providing multi-turn dialogues and context
- One-off explanations not sufficient, potentially ambiguous
- Ease of use; More accessible to laypeople
- Support various explanations in one single system

Interrolang – an Example



Another Example: Rationale generation

Under the Hood

TalkToModel
Interactive Dialogues with ML Models

HUGGING FACE

\+

NLP Model Token Attributions	Natural Language Counterfactuals
Rationale Generation with LLMs	Semantic Similarity

=

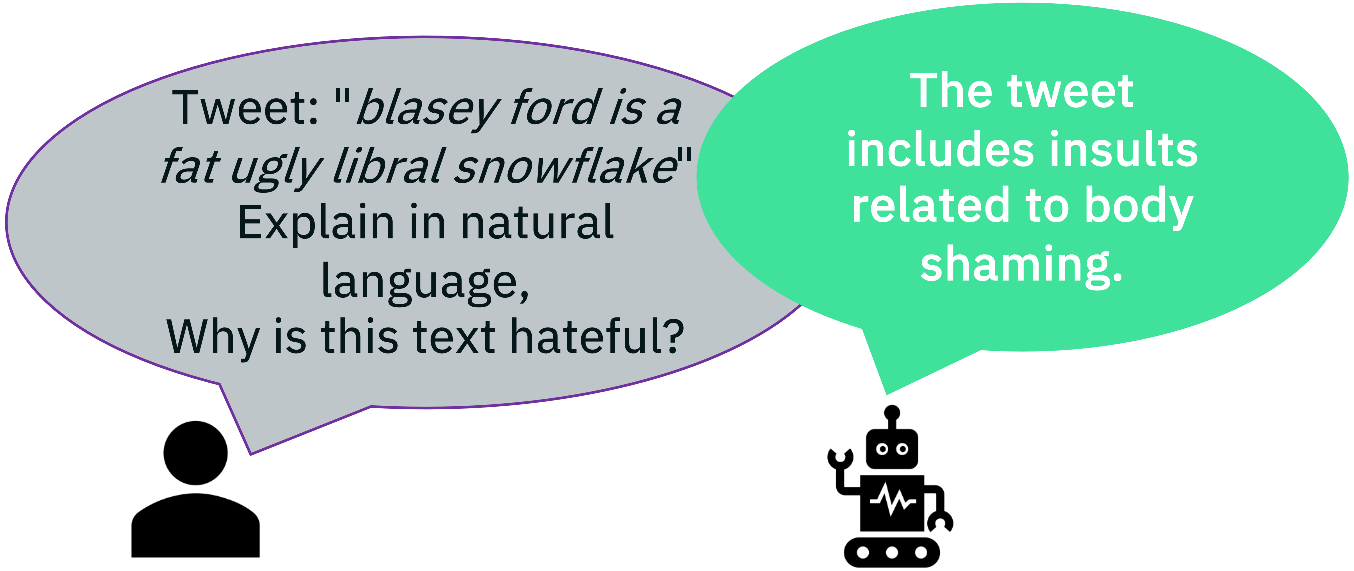
!?!

Task: Dialogue Act Classification

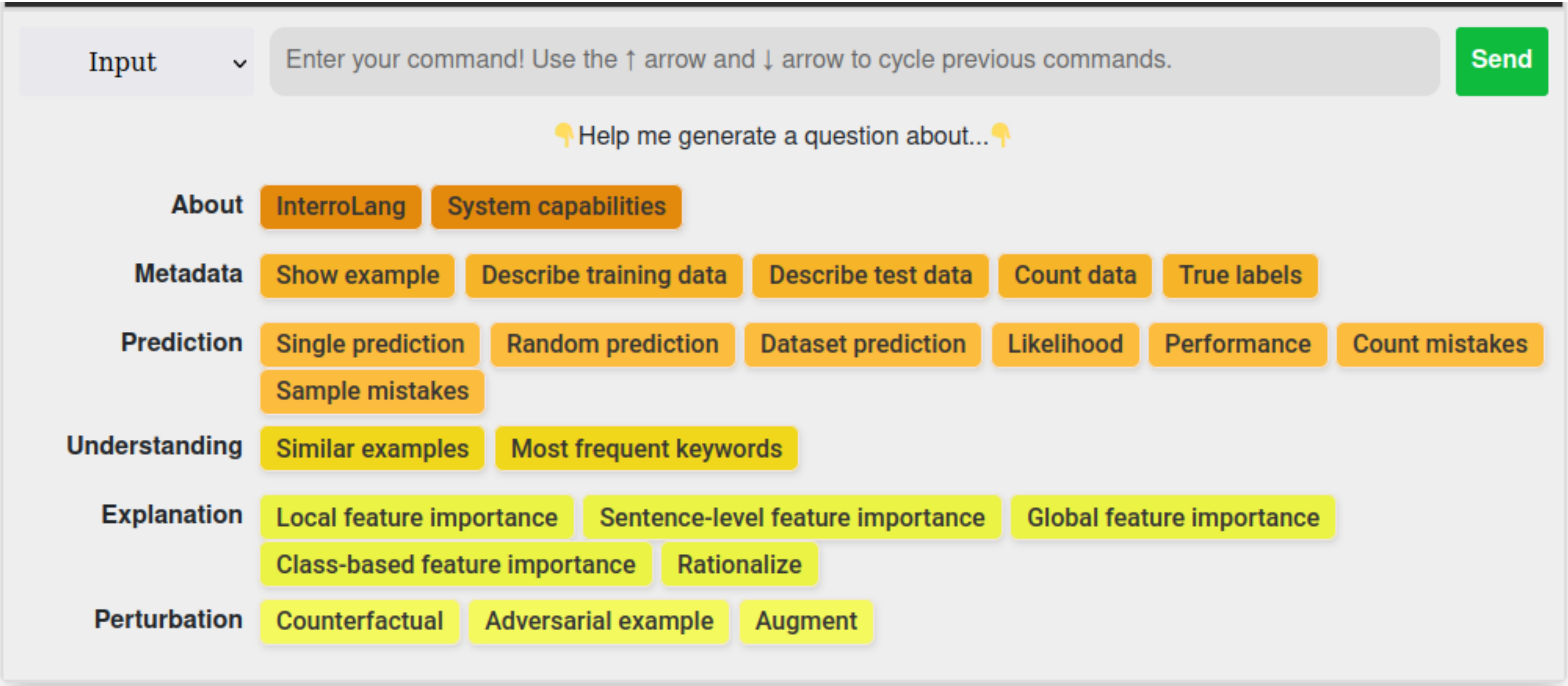
…

Task: Question Answering

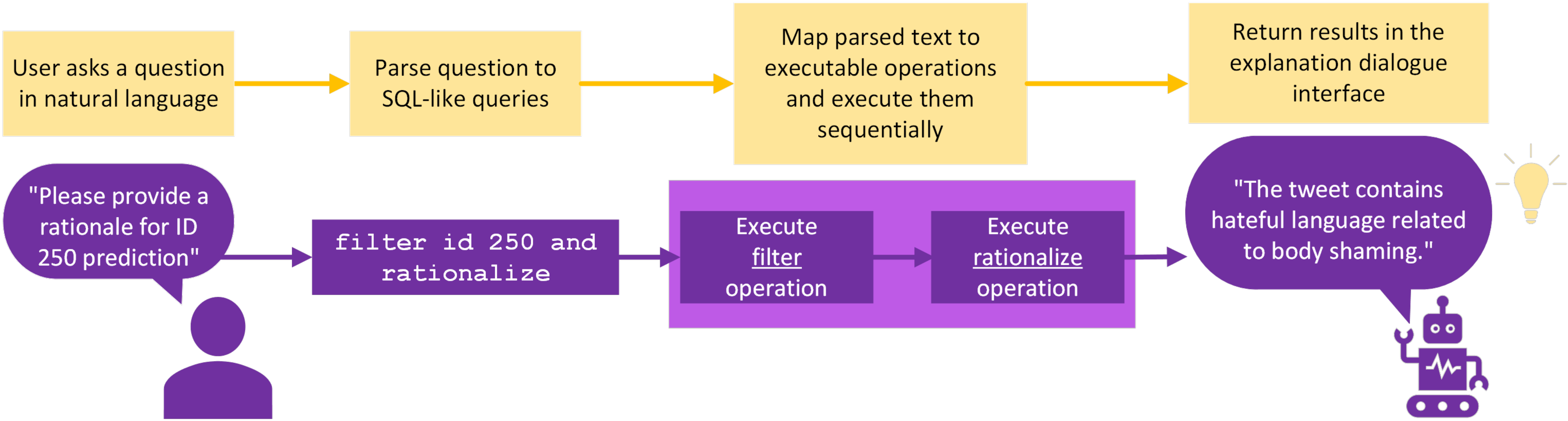
Task: Hate Speech Detection



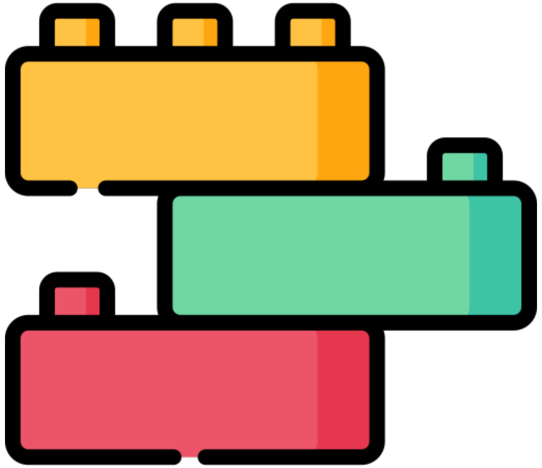
Operations


Input
Enter your command! Use the ↑ arrow and ↓ arrow to cycle previous commands.
Send
Help me generate a question about...
About
InterroLang
System capabilities
Metadata
Show example
Describe training data
Describe test data
Count data
True labels
Prediction
Single prediction
Random prediction
Dataset prediction
Likelihood
Performance
Count mistakes
Sample mistakes
Understanding
Similar examples
Most frequent keywords
Explanation
Local feature importance
Sentence-level feature importance
Global feature importance
Class-based feature importance
Rationalize
Perturbation
Counterfactual
Adversarial example
Augment

Intent Recognition



Building Blocks



Operation	Tool / Model
Intent recognition / Parsing	GPT-Neo (2.7B) FLAN-T5-base (250M) BERT + Adapter (110M)
Feature Attribution / Saliency Method	Captum Integrated Gradients
Counterfactuals	Polyjuice (GPT-2)
Adversarial Examples	OpenAttack
Data Augmentation	NLPAug
Rationalization	Dolly v2 (3B)
Similar Examples	SBERT

Human Evaluation: Simulatability

Simulatability = "Forward prediction"

- User is exposed to: Input + Explanation
- User has to predict the expected model outcome
- Simulation accuracy: How often user prediction == Actual model outcome

Explanation types	**Sim** (all)	**Sim** ($t = 1$)	**Help** Ratio	**#Turns** Avg.
Local feature importance	91.43	93.10	**82.86**	3.85
Sent. feature importance	90.00	94.44	60.00	3.84
Free-text rationale	**94.74**	**100.00**	68.42	**3.70**
Counterfactual	85.00	80.00	25.00	4.14
Adversarial example	84.00	85.71	56.00	4.00
Similar examples	88.46	87.50	61.54	4.00

Table 5: Task B of the user study: Simulatability. Simulation accuracy (in %), simulation accuracy for explanations deemed helpful (in %), helpfulness ratio (in %), average number of turns needed to make a decision.

Human Evaluation: Subjective Ratings

	Operations	Corr.	Help.	Sat.
Metadata	Show example	52.94	44.44	42.19
	Describe data	89.66	87.27	87.72
	Count data	56.41	44.44	45.83
	True labels	58.82	64.71	72.22
	Model cards	56.25	43.75	45.06
Prediction	Random prediction	57.59	60.71	65.52
	Single/Dataset prediction	53.42	53.52	54.17
	Likelihood	62.86	67.50	63.41
	Performance	72.50	65.79	76.19
	Mistakes	81.25	68.75	77.09

NLU	Similar examples	53.57	45.61	62.50
	Keywords	60.34	54.00	60.00
Expl.	Feature importance	55.88	42.25	50.00
	Global feature importance	50.00	50.00	31.32
	Free-text rationale	62.07	62.50	65.45
Pertb.	Counterfactual	40.00	27.03	21.62
	Adversarial example	61.90	40.00	37.50
	Augmentation	62.50	52.17	60.00

Subjective ratings (% positive) on Correctness, Helpfulness and Satisfaction for single turns, macro-averaged.

Takeaways

- Human evaluators preferred global explanations and analyses
    1. Metadata (Model cards / Datasheets)
    2. Common mistakes made by the model
    3. Performance metrics (Accuracy, F1, etc.)
- Simulatability shows multi-turn explanations are necessary. Most useful explanation types:
    1. Feature attribution
    2. Free-text rationales

White Box XAI

Investigating the Encoding of Words in BERT's Neurons using Feature Textualization

Tanja Bäumel, Soniya Vijayakumar, Josef van Genabith, Günter Neumann, Simon Ostermann

Feature Visualization

Goal: Find words in an LM. Interpret the meaning of a single neuron!

WHY?!

Identify biases, prune the model, localize domains...

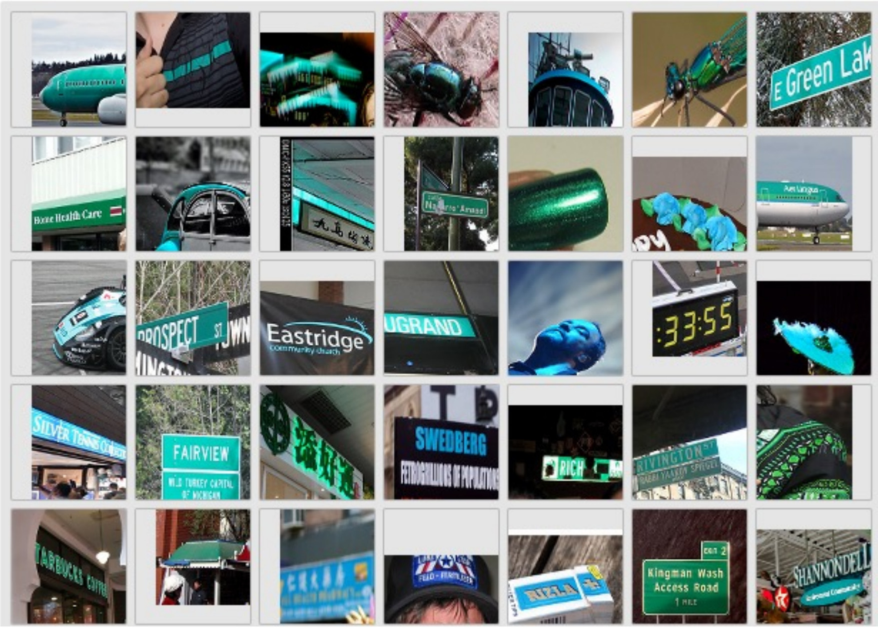
=> Mechanistic XAI

Feature Visualization

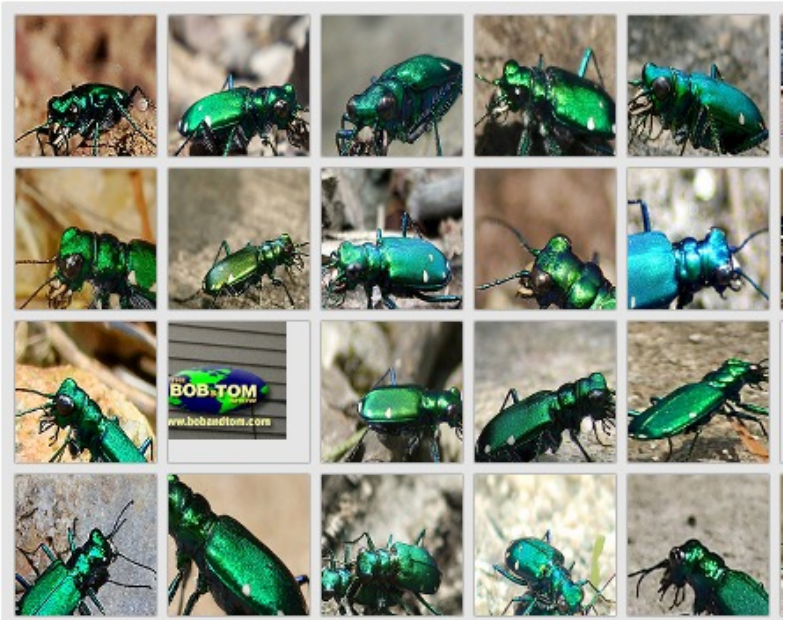
Assumption: The input that maximally excites a specific part of a Neural Network, gives insight into what that part of the NN is sensitive to.

What does unit 16 in Neuron 12 of layer 5 encode?

DATASET: YAHOO FLICKR CREATIVE COMMONS



DATASET: IMAGENET



- Look at Neuron Activations in data sets
- Might differ between data sets!

Feature Visualization

Assumption: The input that maximally excites a specific part of a Neural Network, gives insight into what that part of the NN is sensitive to.

What does unit 16 in Neuron 12 of layer 5 encode?

Feature Visualization

Use **Activation Maximization** to synthesize an optimized input image to maximize activations of a given neuron/component.

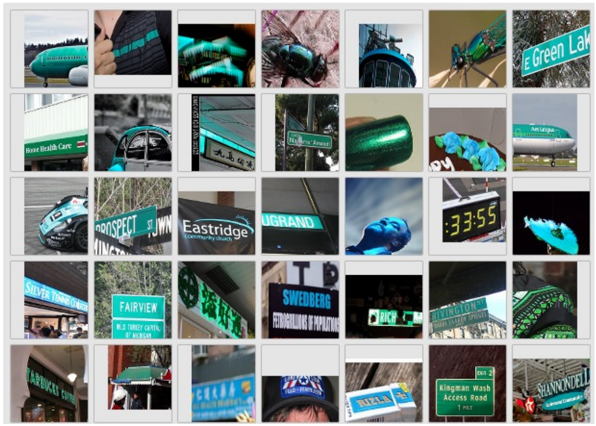
"Learn an input" with the activation size as objective

Previous work: Attempts on finding word representations in BERT

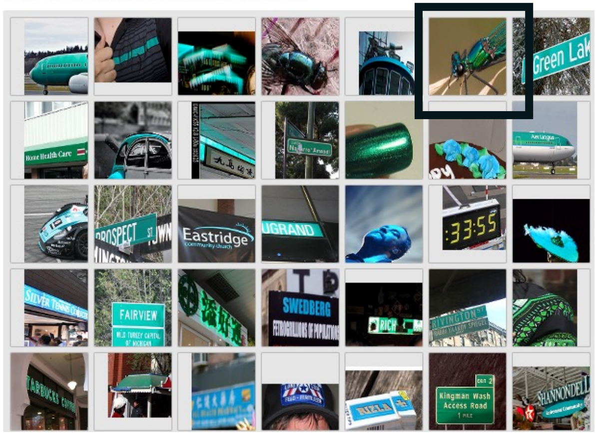
- Simplest case: Feed vocabulary terms to BERT, observe activation patterns

- Try to learn the ideal one-hot representation for a neuron

DATASET: YAHOO FLICKR CREATIVE COMMONS



DATASET: YAHOO FLICKR CREATIVE COMMONS



Problems with previous work

Try to force interpretations towards words.

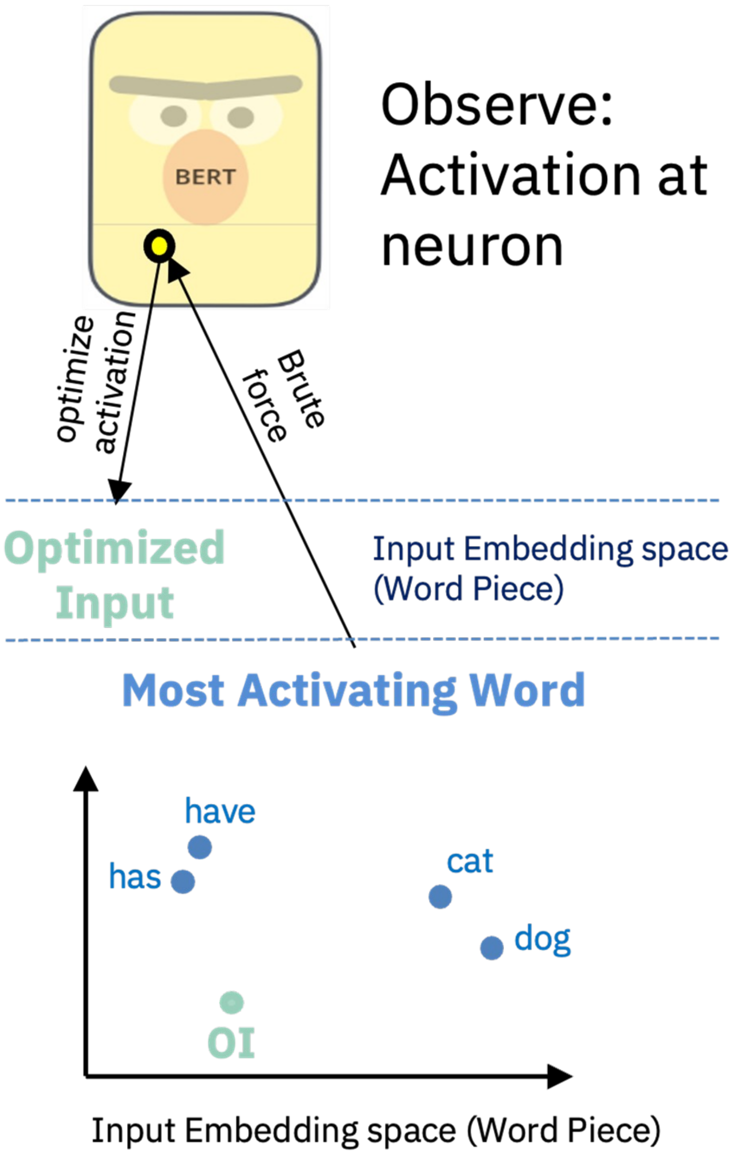
But what if neurons do not encode clear-cut linguistic concepts, such as words?

Language is not continous!

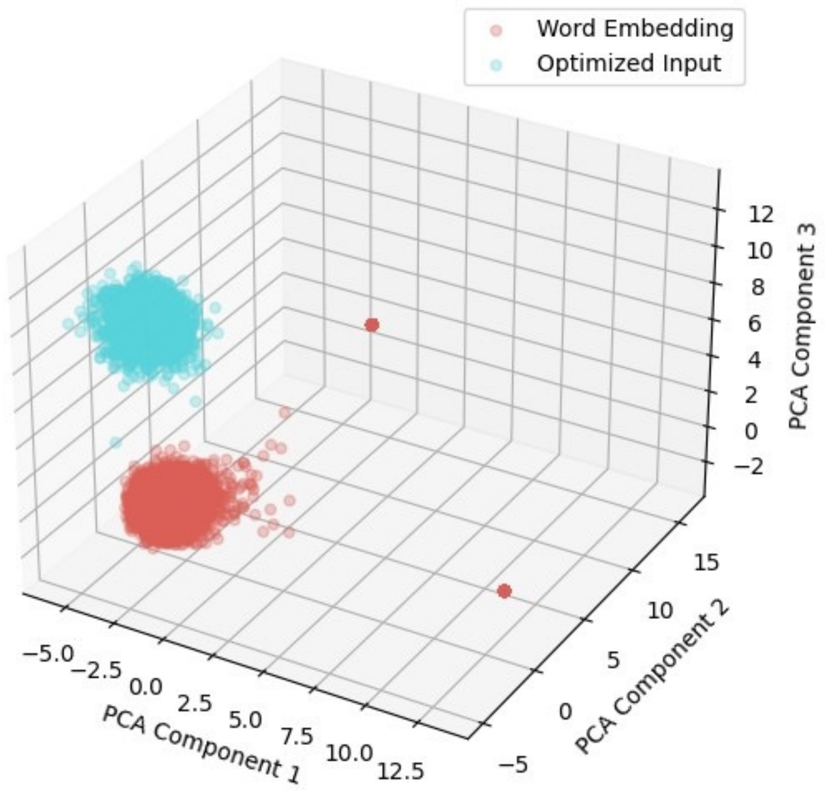
How can we interpret information in between linguistic units?

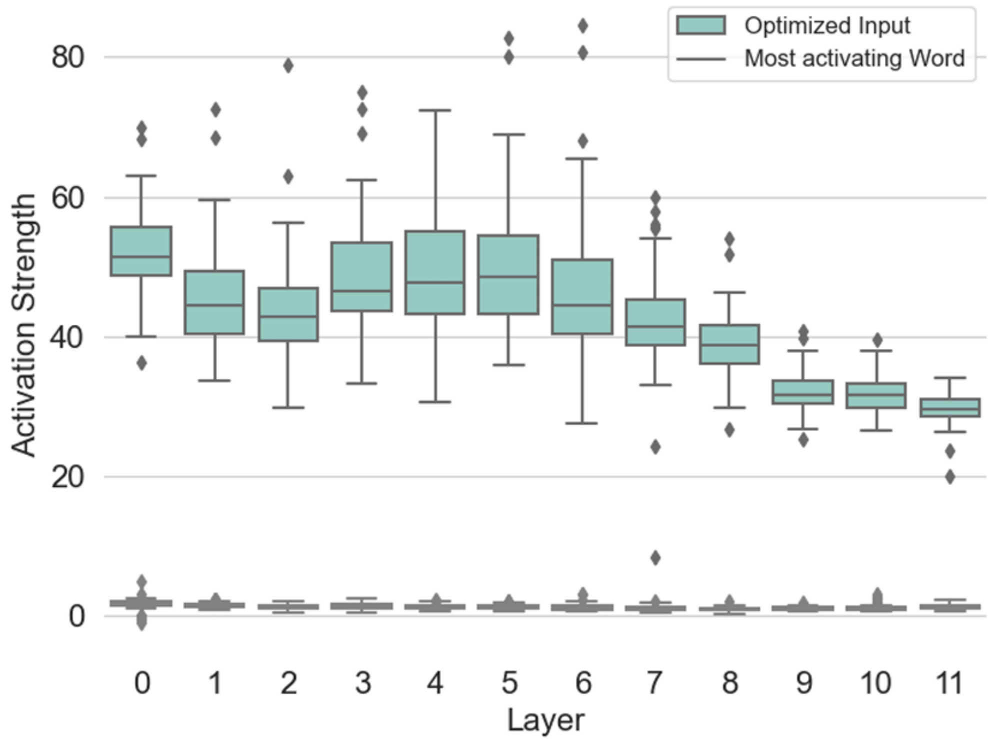
Feature Textualization

- Feature Textualization: Obtain optimized inputs for random neurons **in the embedding space**
- Evaluate Symbolizability by comparing them to actual words with continuous measures
- If a neuron encodes a symbolizable unit, then its optimized input should be similar to a word
  ➔ Similar Vectors
  ➔ Similar Activation Potential



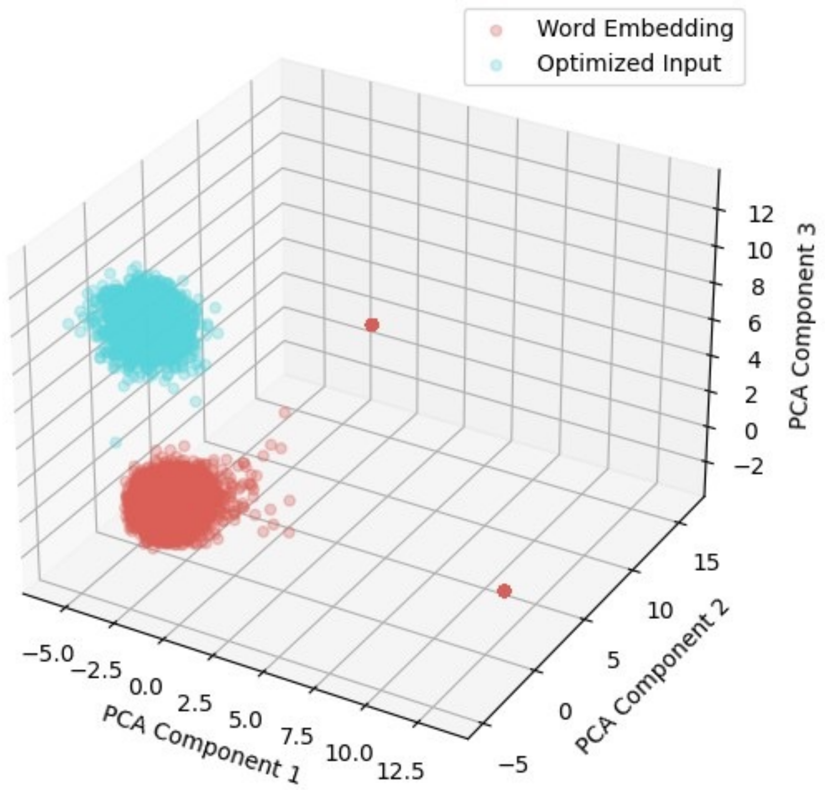
Optimal Inputs for Single Neurons

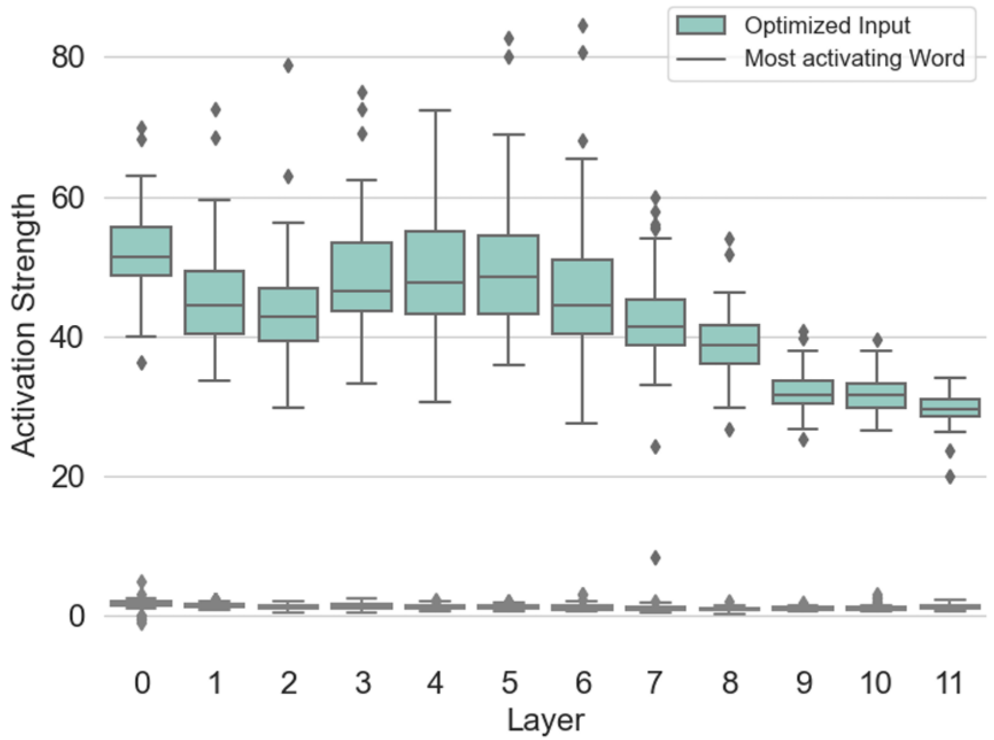




Vector positions and activation potentials are **very** different between optimized inputs and actual words.

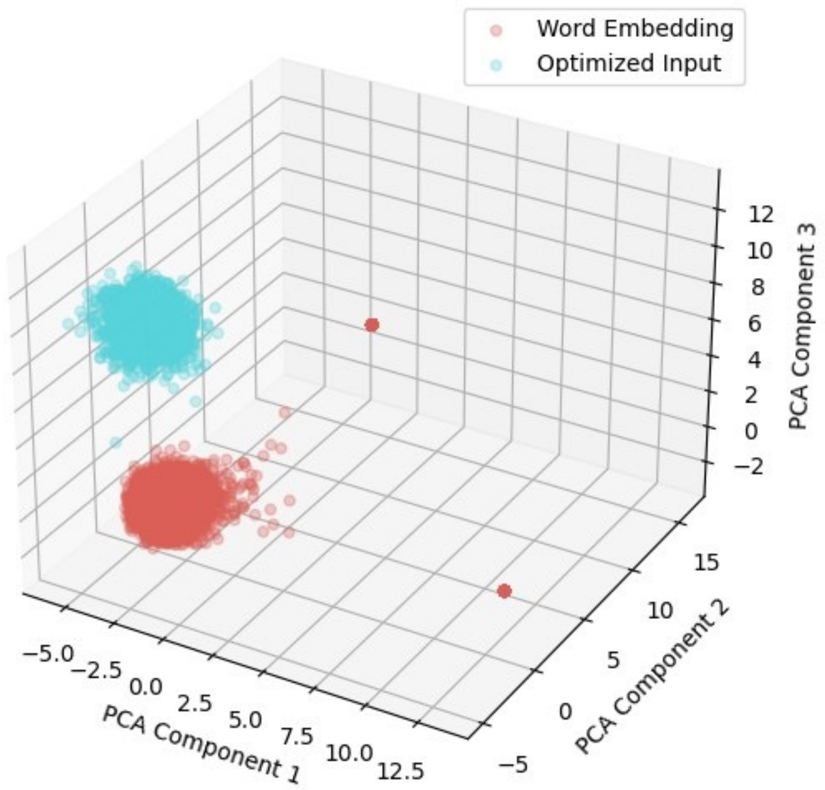
Optimal Inputs for Single Neurons

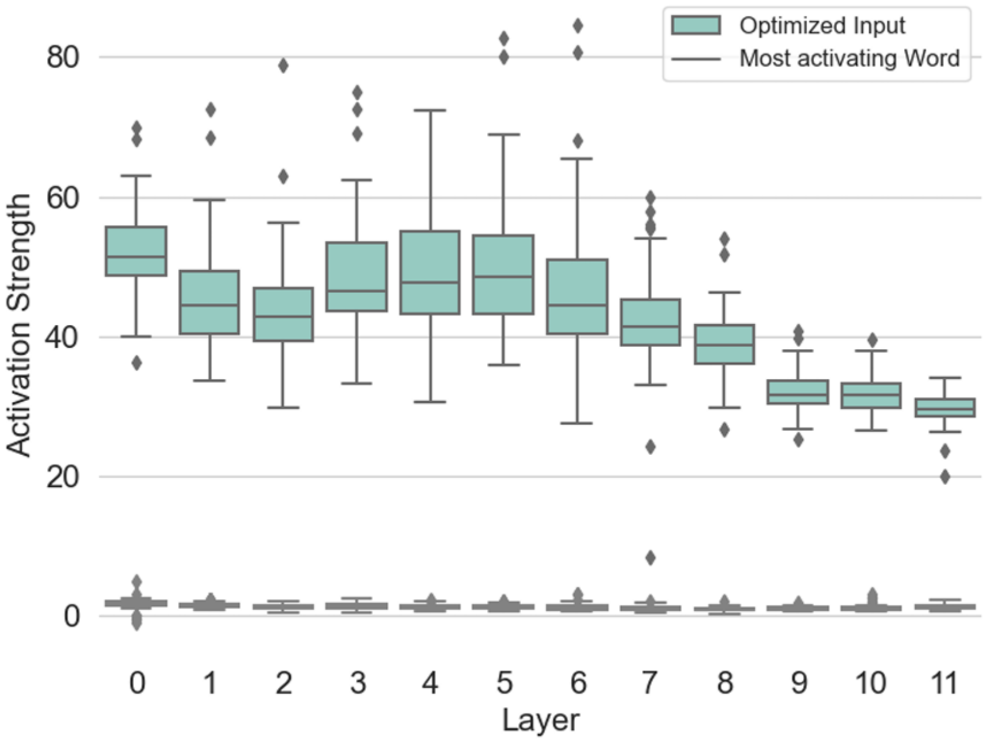




Apparently single neurons don‘t encode words.

Optimal Inputs for Single Neurons

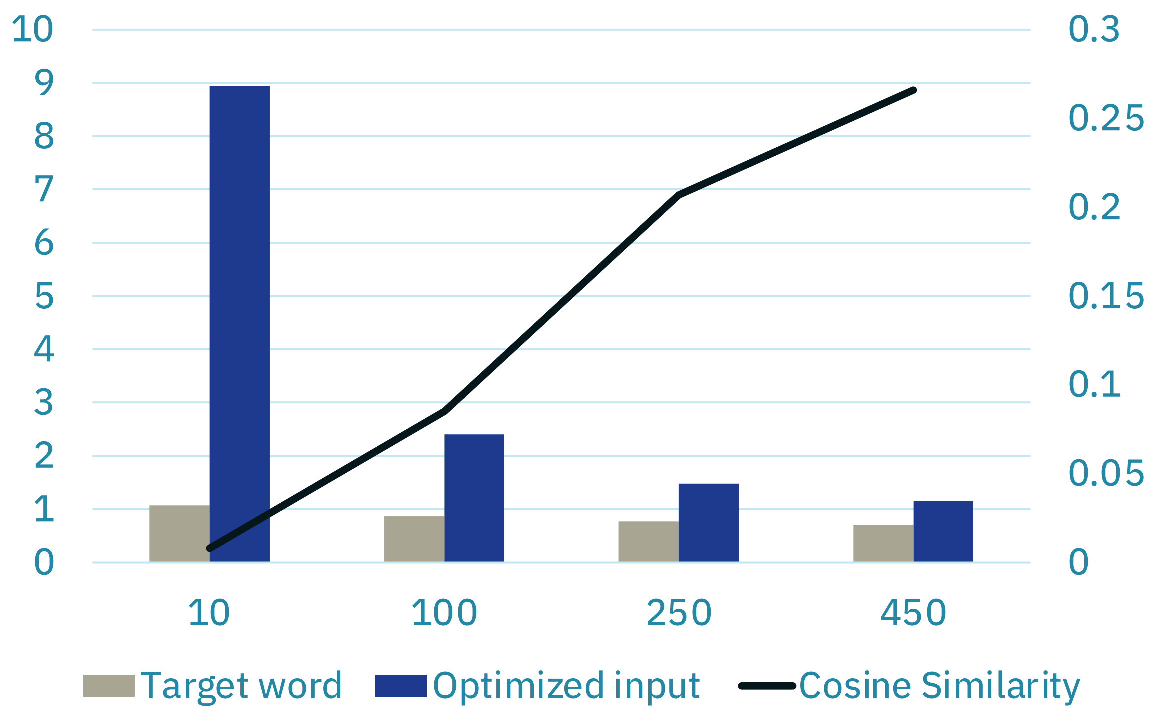




So where are they?!

Optimal Inputs for Multiple Neurons

- We can optimize the activations of multiple neurons at once
- During training, just average over their absolute activations
- But which neurons to pick?
  - Proof-of-Concept experiments!
  - Pick the top n activated neurons for random words
  - Optimize them together
  - Do we end up close to the original word?



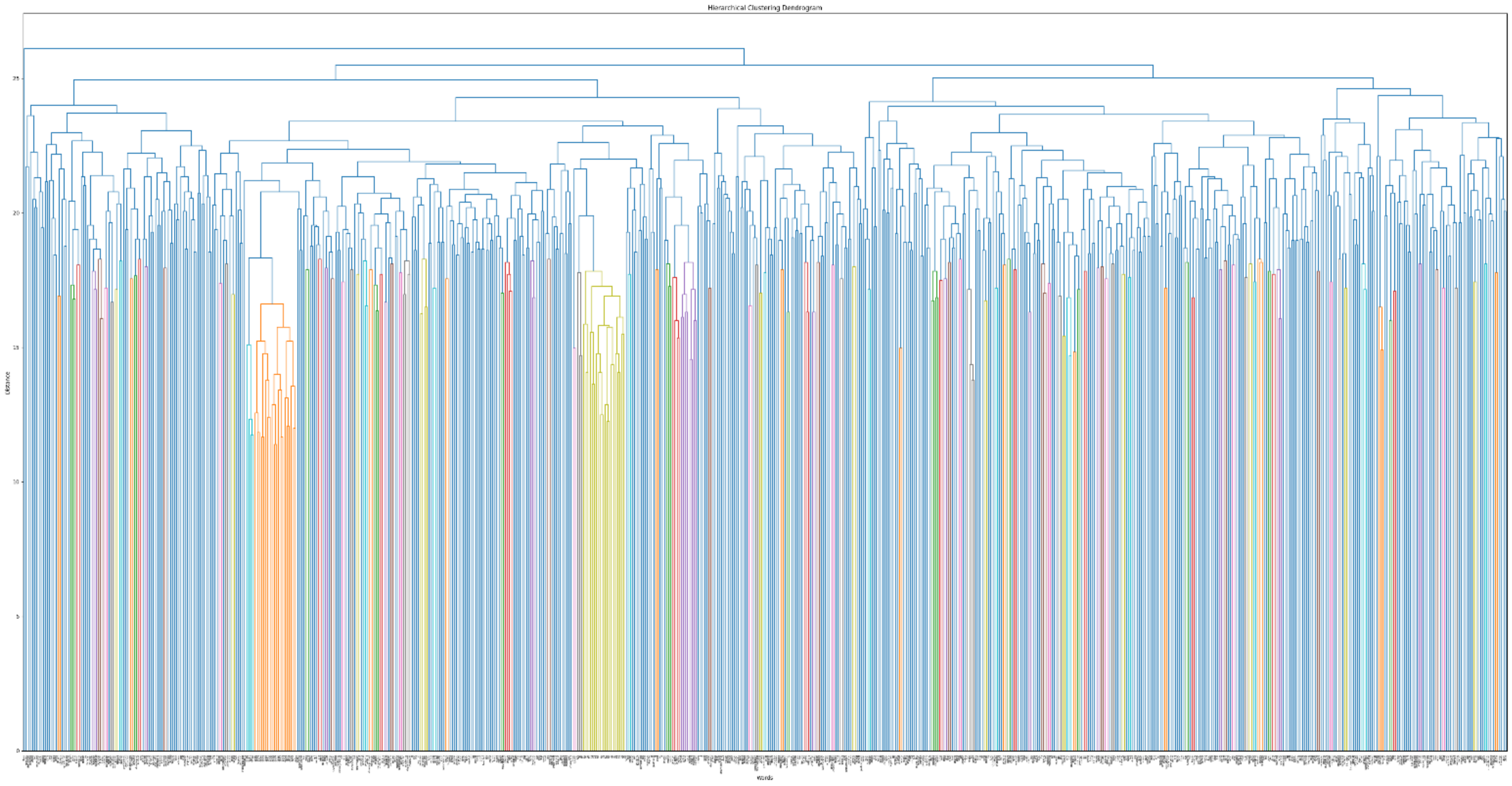
Interesting Observations with Multiple Neuron activations

The top 500 activated neurons are basically semantic vectors.

Largest overlap in activated neurons:

romanian	**english**	**butler**	**get**	**1**
albanian	arabic	gilbert	gets	2
croatian	french	barnes	got	3
indonesian	japanese	hughes	getting	4
thai	spanish	sullivan	gotten	5
iranian	latin	bennett	catch	7
argentine	irish	murphy	analyze	9
armenian	italian	wallace	respond	11
bulgarian	hindi	phillips	deliver	8
hindi	thai	edwards	boil	14
byzantine	filipino	montgomery	drown	13

Interesting Observations with Multiple Neuron activations



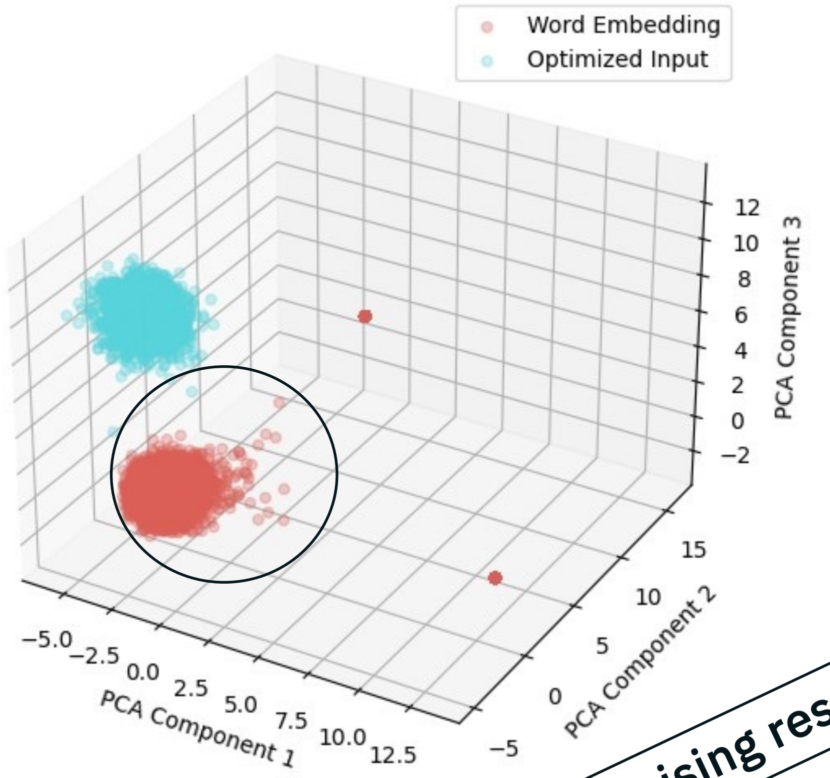
Feature Textualization - Some next steps

On the technical side:

- Vanilla gradient ascent. Maximize a single unit activation w.r.t. the input
- Often results in finding local/global minima that are far from the embedding space
- Next steps: Try to counteract this by using priors based on the embedding space

Example: **Membership prior**. Test if the optimized input falls into a particular part of space.

Intuition: compute an objective that is 0, if the optimal input is in a hypothetic cone around the embeddings (i.e. diff to the center < cone radius), and large if it's far away



Just initial promising results so far

Feature Textualization - Some next steps

On the conceptual side: Right now this is a more theoretical kind of work.

- Make this more usable to researchers
- Connect it to other efforts around mechanistic XAI

Takeaways

- Single neurons do not encode words
  - Optimized inputs are far away
  - They lead to much higher activations
- Apparently, more than 400 neurons are needed to get close to words
  - There are structures to be found in BERT, when looking at sets of neurons needed to encode words
  - Much more work needs to be done to determine „good combinations“ of neurons
- There is still a gap to feature visualization in computer vision, need for priors!

Summary

- Black Box XAI:
  - Useful for end users
  - Doesn‘t look into the model but rather tries to interpret using data operations
- White Box:
  - More useful for researchers
  - Try to find meaning in network components, but hard to understand for non-AI researchers
- Dialogue-Based explanations and feature textualization as two examples

Questions?

- Black Box XAI:
  - Useful for end users
  - Doesn‘t look into the model but rather tries to interpret using data operations
- White Box:
  - More useful for researchers
  - Try to find meaning in network components, but hard to understand for non-AI researchers
- Dialogue-Based explanations and feature textualization as two examples