Interpretability in NLP

CS 780/880 Natural Language Processing Lecture 25

Samuel Carton, University of New Hampshire

Last lecture

LLM evaluation—nontrivial and probably not adequate currently

Special data collection tricks to create hard examples for LLMs

Historical movement from low-level linguistic capabilities (i.e. inference, sentiment) to high-level capabilities (world knowledge, solving logic puzzles)

Movement from classification to generation

This class so far

Goal so far: Make models more better

- Probabilistic models (N-gram language models, Naïve Bayes)
- Neural models
  - Word vectors
  - RNNs
  - Transformers
  - Really big transformers (and prompt engineering)

But what happens when a human being sits down to try to use a model?

Human-model collaboration

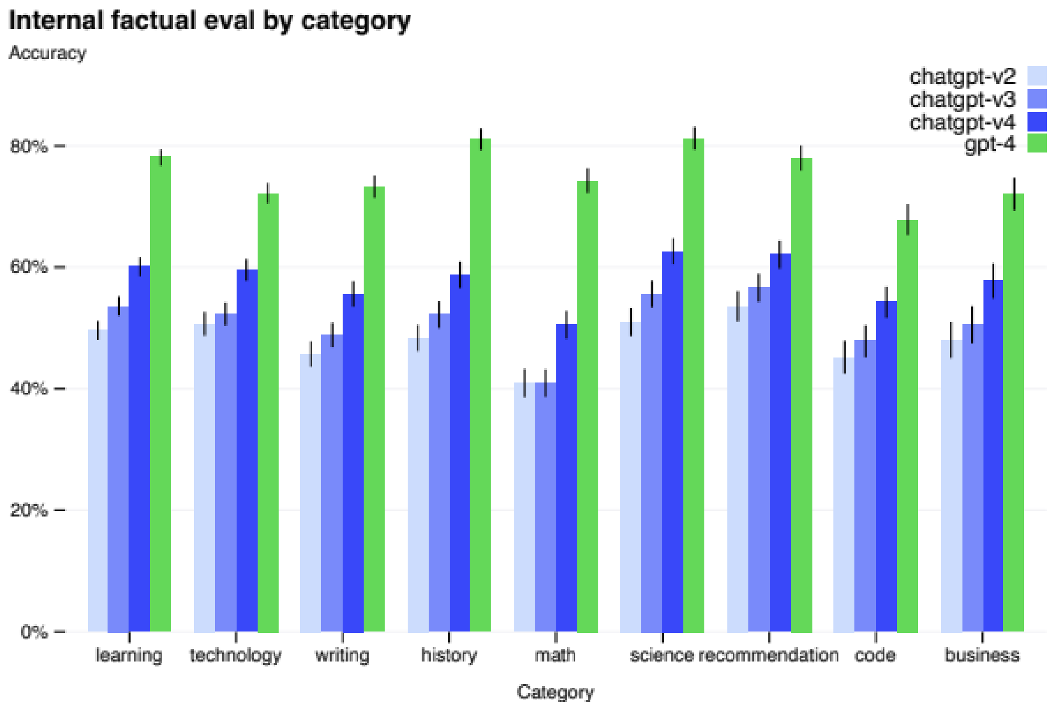
Scenario:

- You are a physician
- You have a model that is trained (or prompted) to look at a patient medical record and predict what diseases they have
  - E.g. "This patient is likely to have breast cancer."
- You want to use this model as an advice-giving assistant.
  - **Artificial**: no doctor would actually do this

Key problem: models never perfectly accurate!

- E.g. GPT-4 hallucinating

Leads to a couple follow-on problems



https://arxiv.org/pdf/2303.08774.pdf

Verification

Basic idea: When the model makes a prediction, how can we verify that it is correct?

Two basic options for medical scenario:

1. Review the medical record yourself
   - If you do this, what was the point in having the model at all?
2. Blindly trust the model
   - If you do this, what is the point in having **you**?

Other options out there, but you have to get creative (more on this in a minute)

Accountability

Basic idea: In high-stakes fields like medicine and law, we have systems of accountability for dealing with errors and harm:

- Malpractice laws/insurance
- Firing
- Sensitivity training
- Appeals, impeachment, ethics suits

But what happens when someone makes a mistake while relying on a model?

- If you "fire" the model, what do you replace it with?
- If you fire the person, what guarantee is there that the next person will be any better?

No way to constructively deal with errors!

Other scenarios

Many examples of human-model collaboration in NLP:

- Reddit moderator using a toxicity classifier to help moderate posts
- Investor using a model to suggest stocks to buy/sell
- Teacher using model to help grade essays
- Student using a model to help write essays
- Materials scientist using a model to suggest new hypotheses to investigate
- Hiring manager using a model to filter resume
- Programmer using a model to help code

Accountability and verification more/less important for some of these

- Not that important for stock trading as long as earnings outweigh losses
- Important for medicine, coding, essay grading, etc.

What’s missing?

What’s missing here is something that can **make people better at working with models**

- Make it easier for humans to verify model predictions
    - While still deriving some benefit from those predictions
- Make it easier to hold humans accountable for decisions made with model advice
    - In a way that actually makes systems better

One (potential) (partial) solution: **interpretability**

Model interpretability

Basic idea: make predictions and then explain those predictions in ways that are useful

Examples:

- "This Reddit post is toxic because of these X, Y and Z words which are personal attacks"
- "This tweet is misinformation because it contains a specific rumor which is known to be false"
- "This product review is fake because it contains unnecessary details"
- "This essay is poorly-written because it uses overly verbose diction"

I focus on NLP, but important everywhere

Feature attribution

Explanations mostly take the form of **feature attribution** or **explanatory examples**

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

Feature attribution

Feature attribution: which features (words) of the input were most responsible for the model’s output

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

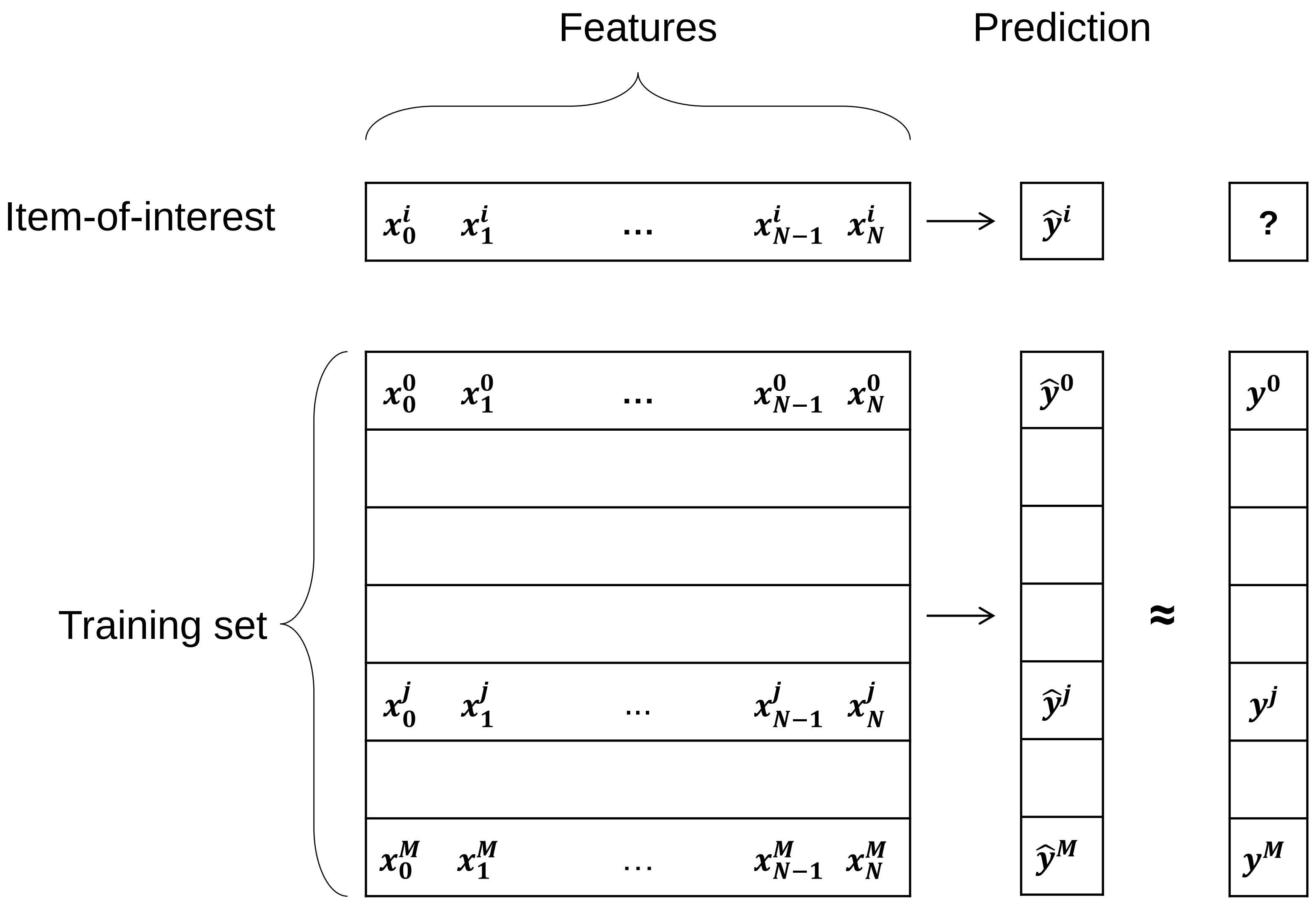
Explanatory examples

Explanatory examples: what other texts can we retrieve from the data (or generate) which explain the prediction for this text

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

Explanatory examples

Text	Prediction	True label
Go to hell, you miserable piece of crap!	Toxic	Toxic
Why are you so stupid? Just shut up.	Toxic	Toxic
You are wildly incorrect.	Nontoxic	Nontoxic



Features

Prediction

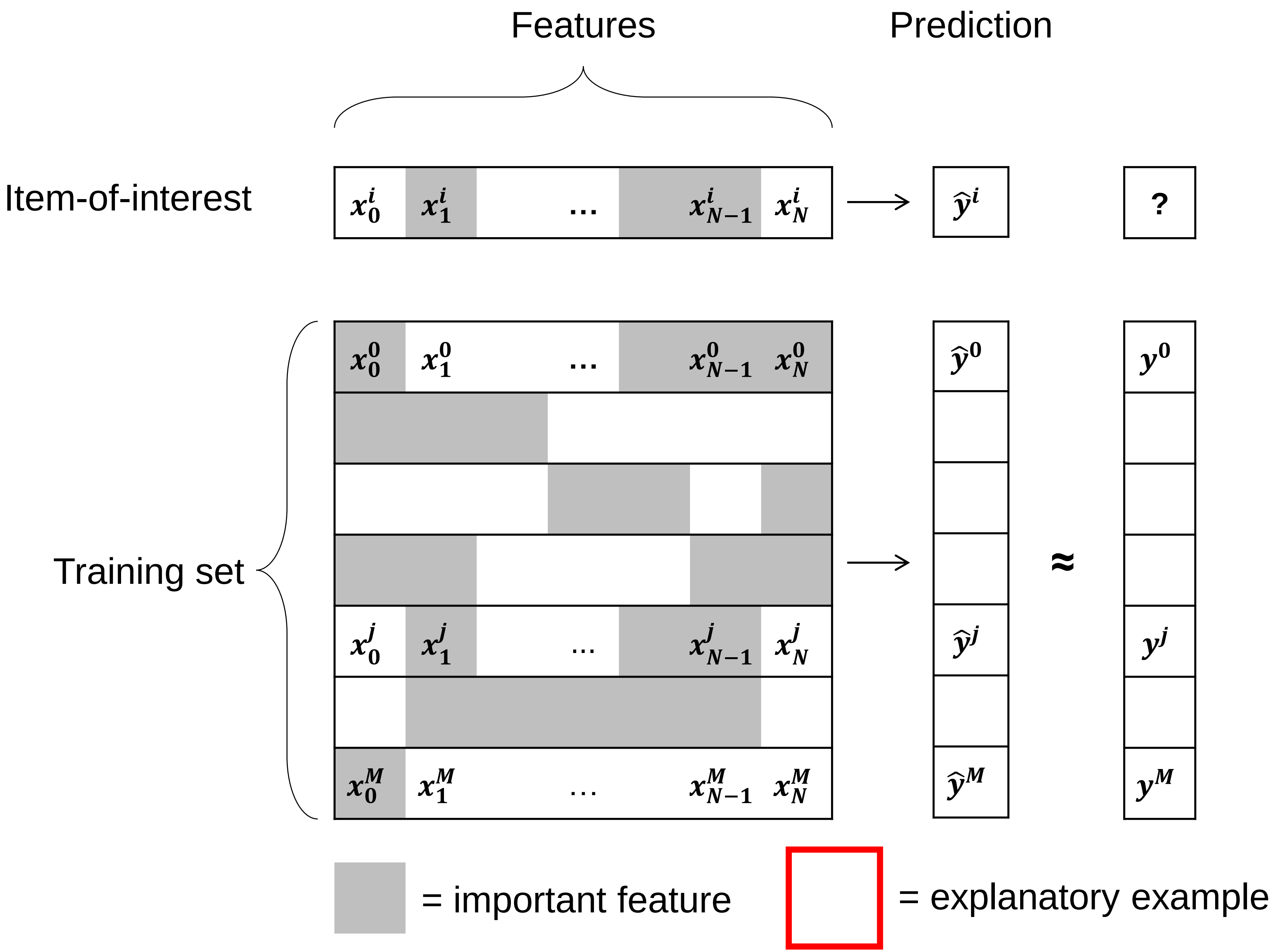
Item-of-interest

Training set

Item-of-interest	x_0^i	x_1^i	…	x_{N-1}^i	x_N^i	→	$\hat{y}^i$	?
Training set	x_0^0	x_1^0	…	x_{N-1}^0	x_N^0		$\hat{y}^0$	y^0
						→		≈
	x_0^j	x_1^j	…	x_{N-1}^j	x_N^j		$\hat{y}^j$	y^j
	x_0^M	x_1^M	…	x_{N-1}^M	x_N^M		$\hat{y}^M$	y^M

▨ = important feature

□ = explanatory example



Features

Prediction

Item-of-interest

x_0^i x_1^i … x_{N-1}^i x_N^i → $\hat{y}^i$?

Training set

x_0^0 x_1^0 … x_{N-1}^0 x_N^0 $\hat{y}^0$ y^0

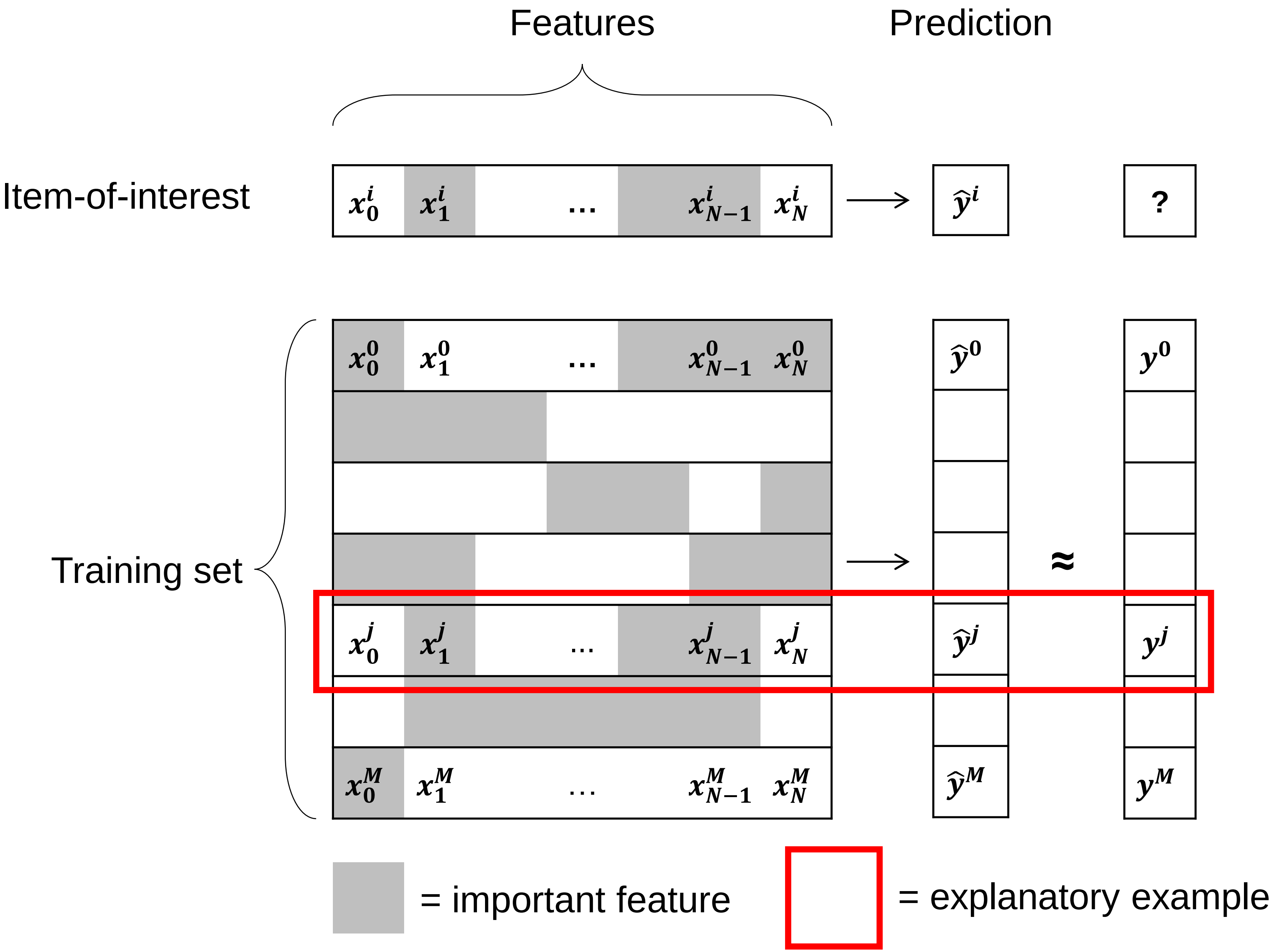
→ ≈

x_0^j x_1^j … x_{N-1}^j x_N^j $\hat{y}^j$ y^j

x_0^M x_1^M … x_{N-1}^M x_N^M $\hat{y}^M$ y^M

▨ = important feature

□ = explanatory example



Features

Prediction

Item-of-interest

x_0^i x_1^i … x_{N-1}^i x_N^i → $\hat{y}^i$?

Training set

x_0^0 x_1^0 … x_{N-1}^0 x_N^0 $\hat{y}^0$ y^0

→ ≈

x_0^j x_1^j … x_{N-1}^j x_N^j $\hat{y}^j$ y^j

x_0^M x_1^M … x_{N-1}^M x_N^M $\hat{y}^M$ y^M

▨ = important feature

🟥 = explanatory example

Stakeholder view of interpretability

Who are the stakeholders in NLP (and ML generally)?

Model engineers: why is my model failing?

- "What kinds of toxicity is the model failing to recognize, so we can label more data?"

Decision makers: can I trust the model?

- "Is this particular comment *really* toxic, and I (the moderator) should remove it?"

Decision subjects: why was this decision made about me?

- "Hey, why was my comment removed?"
- GDPR "right to explanation"

Scientists: what is the model discovering about the data?

- What kinds of toxicity are people using on the internet these days?

Desiderata for interpretability

Three major desiderata: plausibility, fidelity

Plausibility: does the explanation make sense to human?

- Pretty hard to define “make sense”

Fidelity: does the explanation actually explain the model?

- I.e. whatever the explanation is implying about the behavior of the model, is that implication actually true?

Usefulness: does the explanation help a human use it effectively?

- Very difficult to evaluate, and may not really be related to the other two

Local versus global interpretability

Local interpretability: why is the model making this decision for this particular input?

- “Why is this particular comment toxic?”

Global interpretability: why does the model generally make the decisions it makes?

- “When you see the word ‘hell’, does that usually indicate toxicity?”
- Problem: contemporary models (and reality) are context-sensitive
  - “Go to hell” → toxic
  - “Today was fun as hell” → nontoxic

Most focus is on local

Feature attribution

A.k.a “feature importance”

Which features (words) of the input were most responsible for the model’s output?

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

But how to do it?

Three basic ways:

1. Perturbation
2. Gradient analysis
3. Attention

Perturbation

Basic idea: Characterize how the model responds when you fiddle with the input

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

Text	Prediction	True label
You are [redacted] completely wrong. [redacted]	Nontoxic	?

Leave-one-out

Basic idea: Remove words one-by-one and see how the model responds

Text	Prediction	True label
You are [MASK] and completely wrong. [MASK] [MASK] [MASK].	Nontoxic	?

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

✓

Leave-one-out

Basic idea: Remove words one-by-one and see how the model responds

Text	Prediction	True label
You are [MASK] and completely wrong. [MASK] [MASK] [MASK].	Nontoxic	?

↓

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

✓

But collinearity…

Text	Prediction	True label
You are [MASK] and completely wrong. Go to hell.	Toxic	?

↓

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

✗

LIME & SHAP

Basic idea: Learn a linear model that describes how each word affects the model output when other words might also be missing. Then use coefficients of that model as scores.

Text	Prediction	True label
[MASK] are stupid [MASK] completely [MASK]. Go to [MASK].	Nontoxic	?
[MASK] are [MASK] and completely wrong. [MASK] to [MASK].	Nontoxic	?
You [MASK] [MASK] and [MASK] wrong. Go [MASK] hell.	Toxic	?
[MASK] are stupid [MASK] completely [MASK]. Go to hell.	Toxic	?
You [MASK] stupid and [MASK] [MASK]. [MASK] to [MASK].	Toxic	?
You are stupid [MASK] [MASK] [MASK]. Go [MASK] hell.	Toxic	?
[MASK] [MASK] [MASK] and completely [MASK]. Go to [MASK].	Nontoxic	?

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

Perturbation: LIME & SHAP

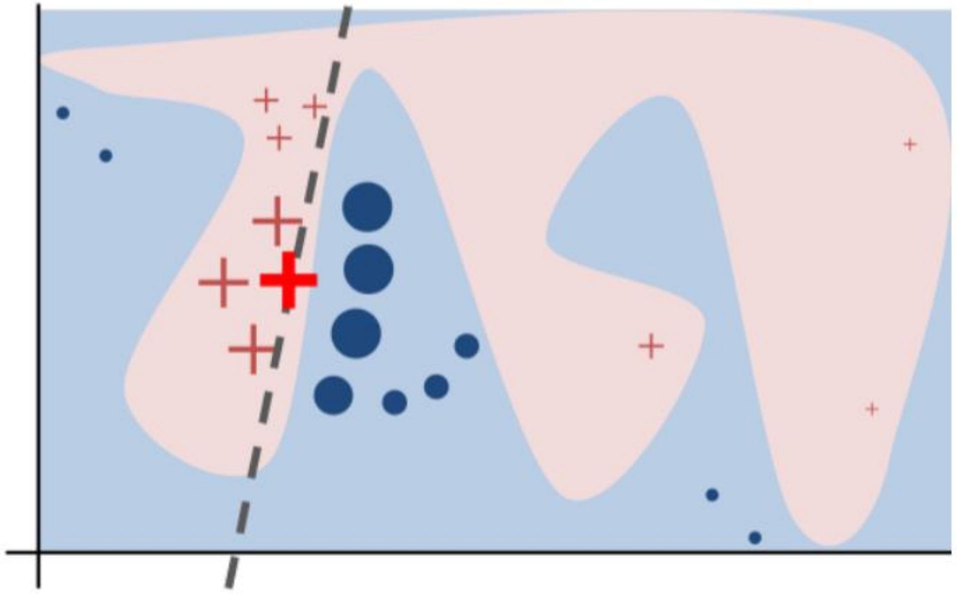
Both very common methods that show up all over the place

LIME

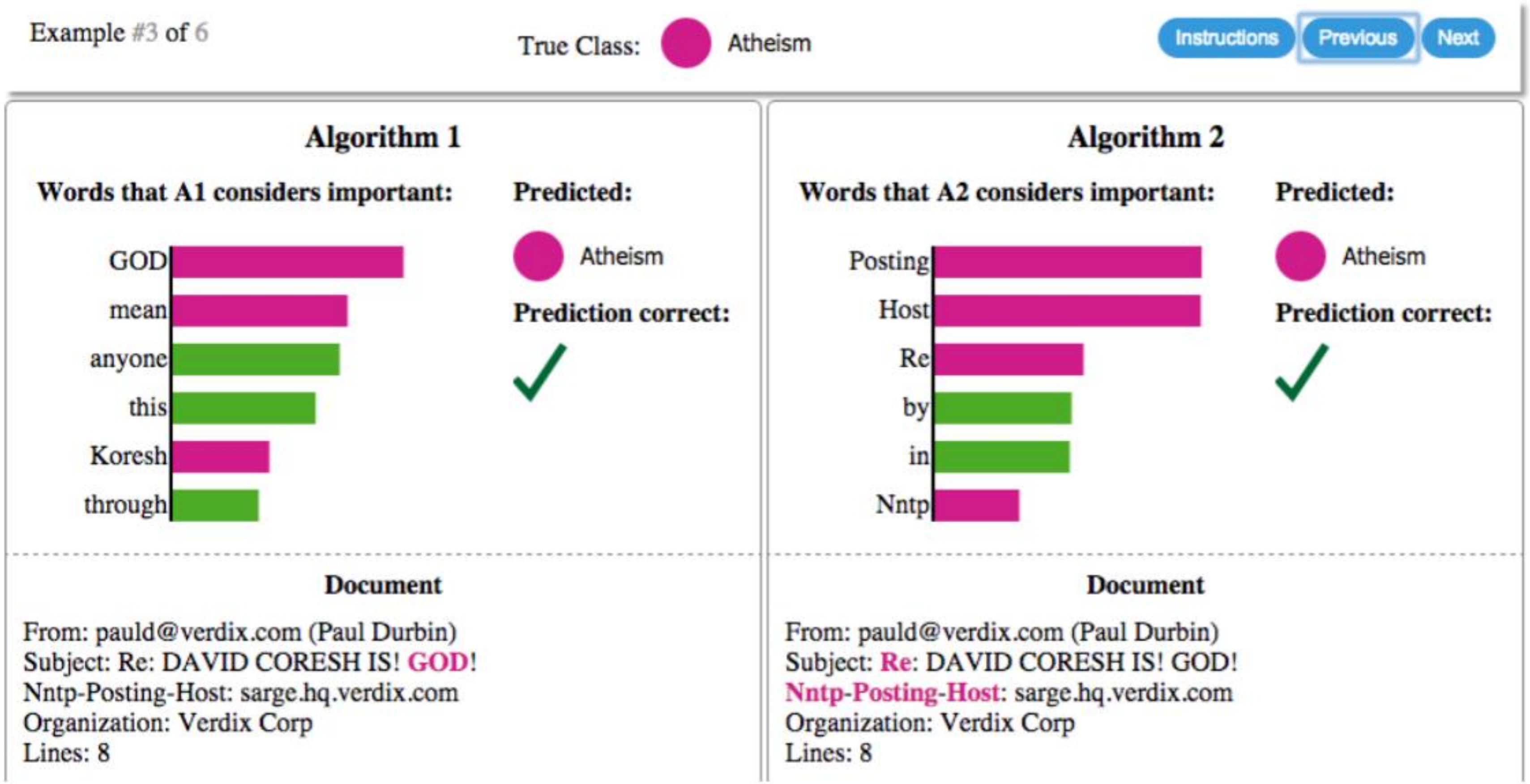
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

SHAP

- Basically just a variant of LIME that obeys certain axiomatic properties
- Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*:10.



LIME & SHAP



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted:

● Atheism

Prediction correct:

✓

Algorithm 2

Words that A2 considers important:

Predicted:

● Atheism

Prediction correct:

✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

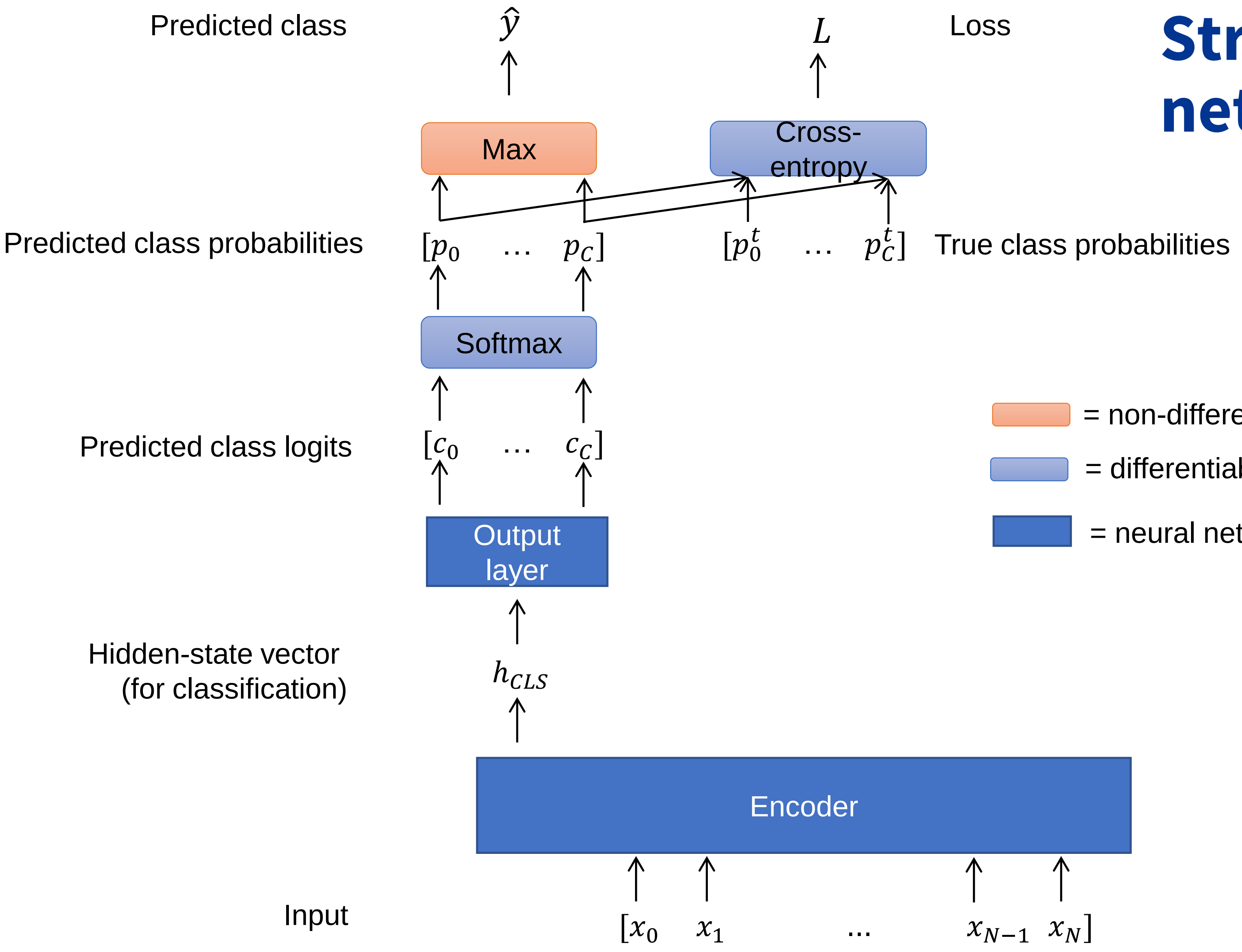
Gradient analysis

Basic idea: use basic mathematical/structural properties of the model to infer feature importances without explicitly perturbing input

Take advantage of the fact that entire neural net is a differentiable function

- We can calculate derivative of any output with respect to any input (or parameters)

Structure of a neural net



Predicted class $\hat{y}$

L Loss

Max

Cross-entropy

Predicted class probabilities $[p_0 \quad \dots \quad p_C]$

$[p_0^t \quad \dots \quad p_C^t]$ True class probabilities

Softmax

Predicted class logits $[c_0 \quad \dots \quad c_C]$

Output layer

Hidden-state vector (for classification) h_{CLS}

Encoder

Input $[x_0 \quad x_1 \quad \dots \quad x_{N-1} \quad x_N]$

🟧 = non-differentiable operation

▭ = differentiable operation

🟦 = neural net layer with parameters

Gradient saliency

Basic idea: Calculate partial derivative of model output with respect to each individual input:

$\frac{\partial \hat{y}}{\partial x_i}$ or $\frac{\partial L}{\partial x_i}$ for all $x_i \in \boldsymbol{x}$

If derivative is high, x_i is having a big impact on model output at that point, and thus is important.

Easy to calculate

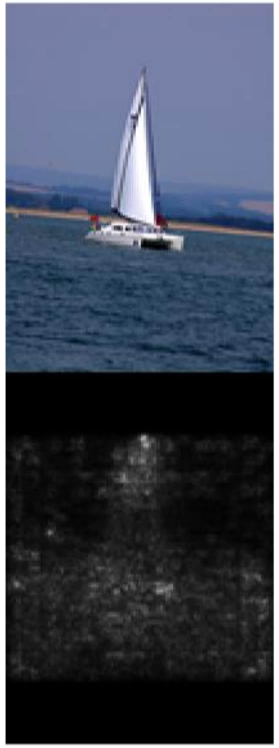
- torch.autograd.grad() for Pytorch

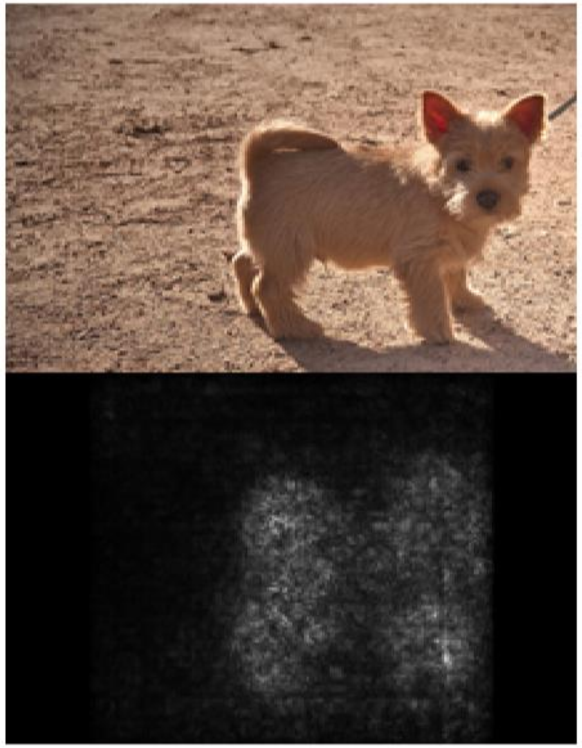
Also has problems with collinearity, and isn't bounded

Gradient saliency

Done more commonly for images

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps.





Attention

Basic idea: Extract internal attention values from model, infer from them which tokens were important to the model

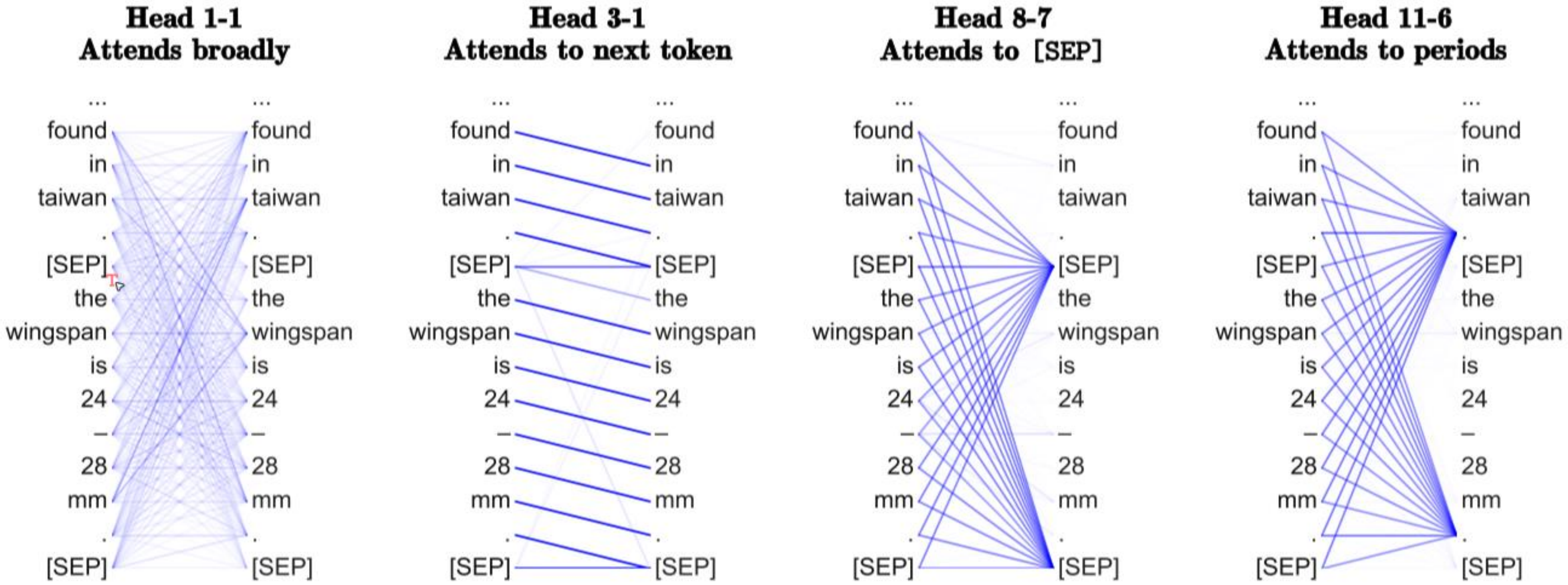
Only works if the model has an attention mechanism

- …but BERT does!

BERT attention

BERT attention

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention.



Rationale model

Basic idea: expand model to two encoders; one to decide attention, one to make prediction

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*

Rationale model

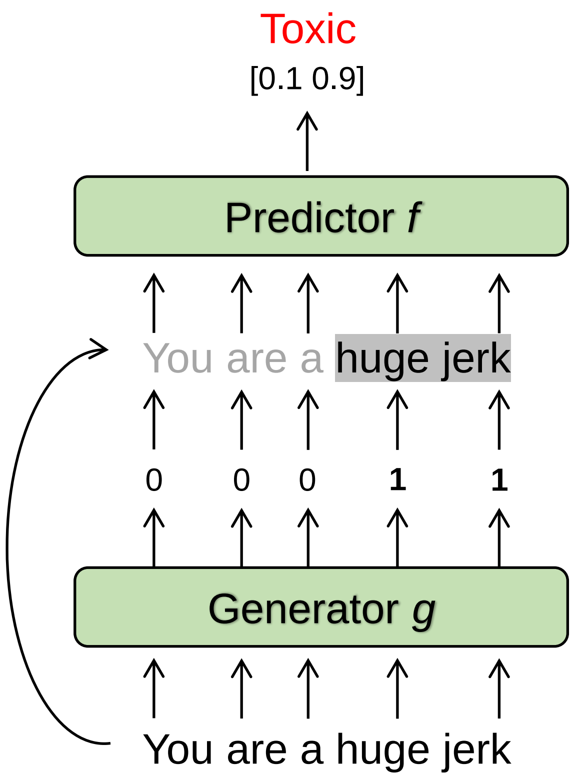
Predictor: $cost_f(z, x, y) = [f(x, z) - y]^2$

Generator: $cost_g(z, x, y) =$

$[f(x, z) - y]^2$ Accuracy

$+\lambda_1 ||z||$ Sparsity

$+\lambda_1 \lambda_2 \sum_t |z_t - z_{t-1}|$ Cohesion



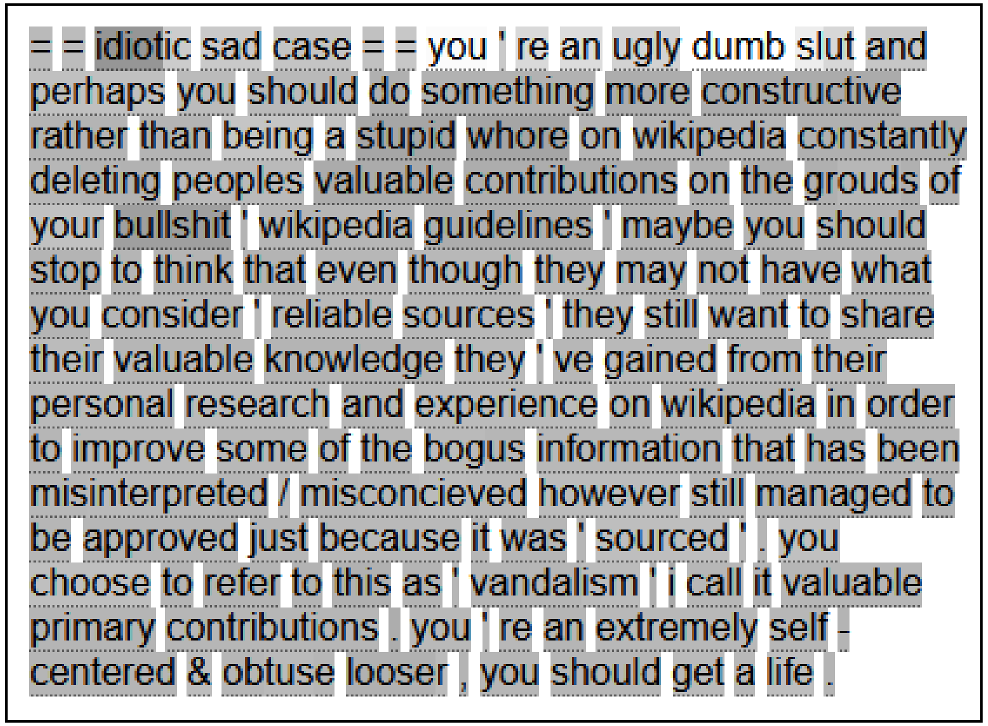
Train layers together, optimizing f for cost_f and g for cost_g

Problems with feature attribution

Different feature attribution methods can do a nice job of calculating exactly what effect each word is having on the model’s output.

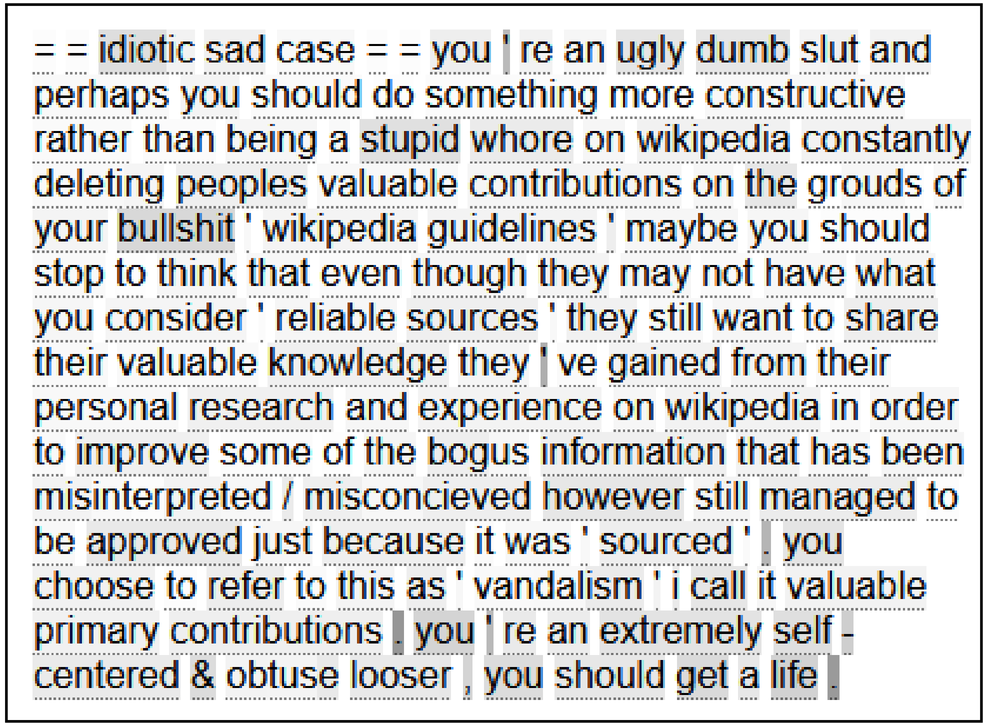
But how useful is this exactly?

Gradient saliency:



= = idiotic sad case = = you ' re an ugly dumb slut and perhaps you should do something more constructive rather than being a stupid whore on wikipedia constantly deleting peoples valuable contributions on the grouds of your bullshit ' wikipedia guidelines ' maybe you should stop to think that even though they may not have what you consider ' reliable sources ' they still want to share their valuable knowledge they ' ve gained from their personal research and experience on wikipedia in order to improve some of the bogus information that has been misinterpreted / misconcieved however still managed to be approved just because it was ' sourced ' . you choose to refer to this as ' vandalism ' i call it valuable primary contributions . you ' re an extremely self - centered & obtuse looser , you should get a life .

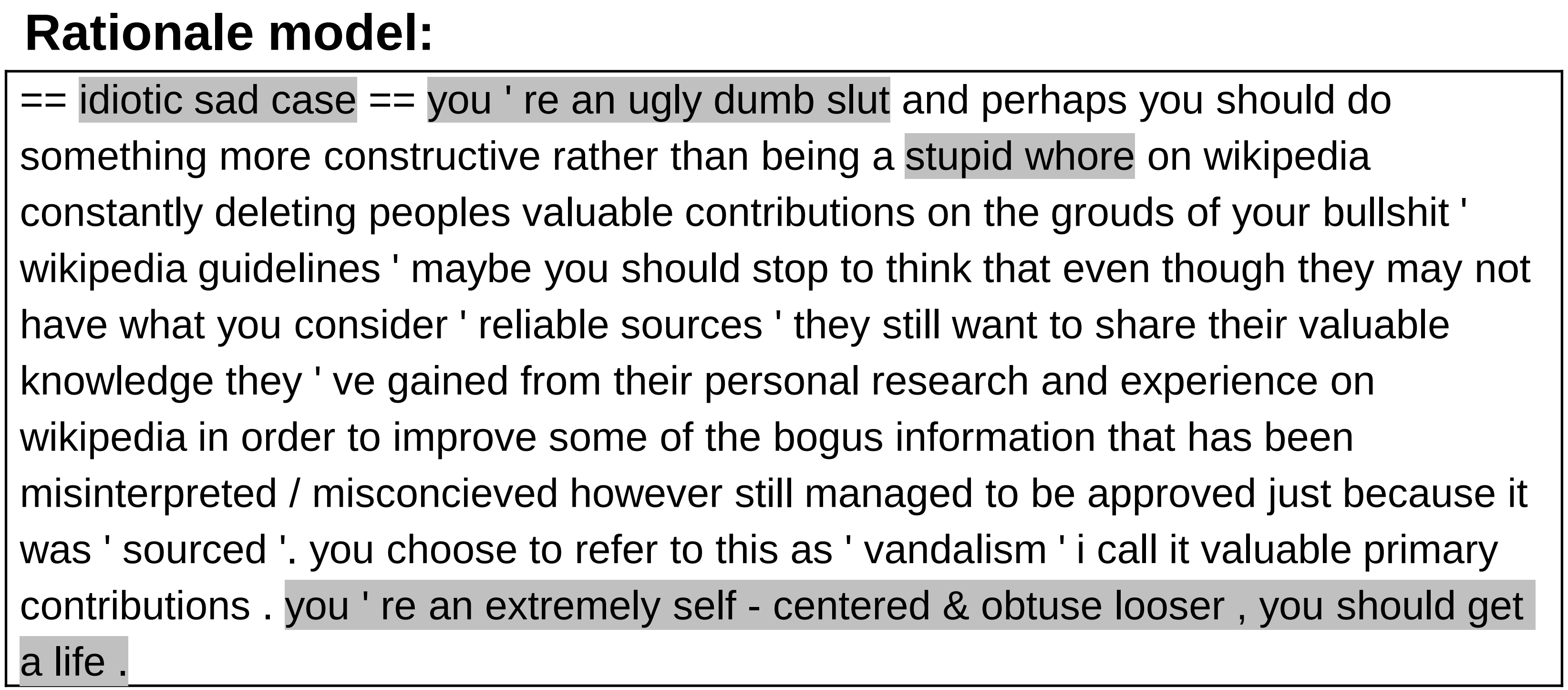
Summed BERT attention:



= = idiotic sad case = = you ' re an ugly dumb slut and perhaps you should do something more constructive rather than being a stupid whore on wikipedia constantly deleting peoples valuable contributions on the grouds of your bullshit ' wikipedia guidelines ' maybe you should stop to think that even though they may not have what you consider ' reliable sources ' they still want to share their valuable knowledge they ' ve gained from their personal research and experience on wikipedia in order to improve some of the bogus information that has been misinterpreted / misconcieved however still managed to be approved just because it was ' sourced ' . you choose to refer to this as ' vandalism ' i call it valuable primary contributions . you ' re an extremely self - centered & obtuse looser , you should get a life .

Problems with feature attribution

Or even this?



Rationale model:

== idiotic sad case == you ' re an ugly dumb slut and perhaps you should do something more constructive rather than being a stupid whore on wikipedia constantly deleting peoples valuable contributions on the grouds of your bullshit ' wikipedia guidelines ' maybe you should stop to think that even though they may not have what you consider ' reliable sources ' they still want to share their valuable knowledge they ' ve gained from their personal research and experience on wikipedia in order to improve some of the bogus information that has been misinterpreted / misconcieved however still managed to be approved just because it was ' sourced '. you choose to refer to this as ' vandalism ' i call it valuable primary contributions . you ' re an extremely self - centered & obtuse looser , you should get a life .

Problems with feature attribution

Feature attribution can be very **faithful** to model behavior

But it is kind of **implausible** (because that's just not how humans explain things)

And it has been found to be **not very useful**

- Carton et al. (2020), Cai et al. (2019), Lage et al. (2019), Poursabzi-Sangdeh et al. (2019), Lai et al. (2019), Jacobs et al. (2021), etc.

Pretty clear at this point that just pointing out what the model is looking at is not a very effective solution to the problem of human-model collaboration.

So what next?

Case-based explanations

Basic idea: Retrieve labeled examples from the training set whose true labels explain why the model made the prediction it made.

Text	Prediction	True label
You are stupid and completely wrong. Go to hell.	Toxic	?

Explanatory examples

Text	Prediction	True label
Go to hell, you miserable piece of crap!	Toxic	Toxic
Why are you so stupid? Just shut up.	Toxic	Toxic
You are wildly incorrect.	Nontoxic	Nontoxic

Text similarity

Basic idea: find similar texts from the training set to the item-of interest

Similarity metrics:

- Word overlap
- TF-IDF similarity
- Word vector centroid cosine distance
- Doc2vec

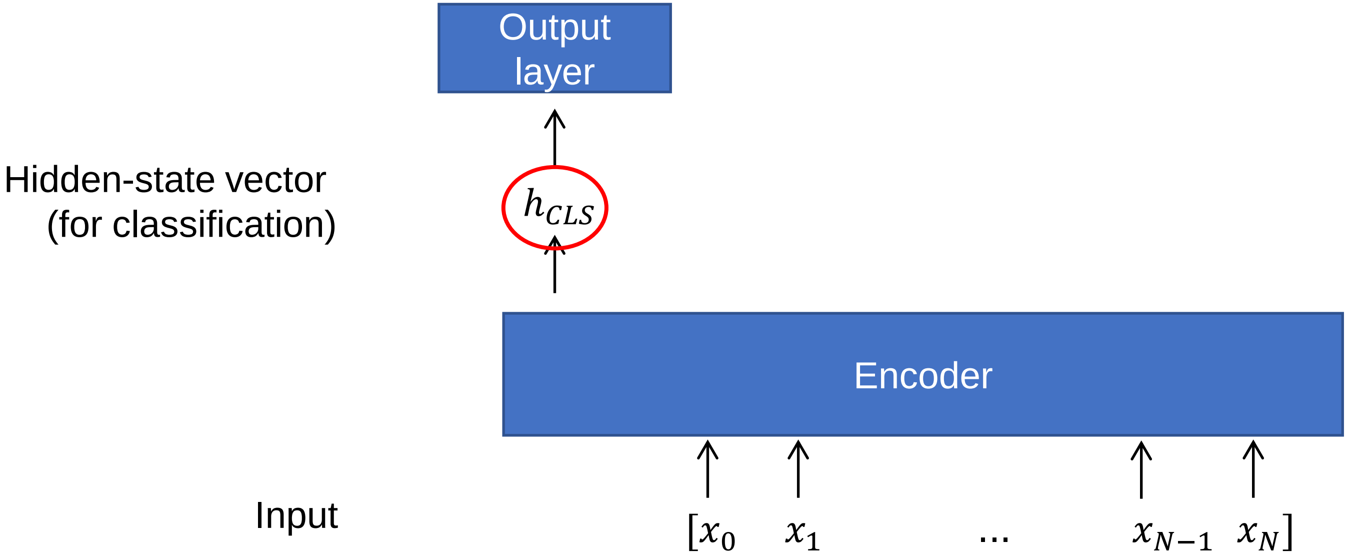
Problem: these two texts might be generally similar, but are they similar **in terms of how the model looks at them?**

- Fidelity

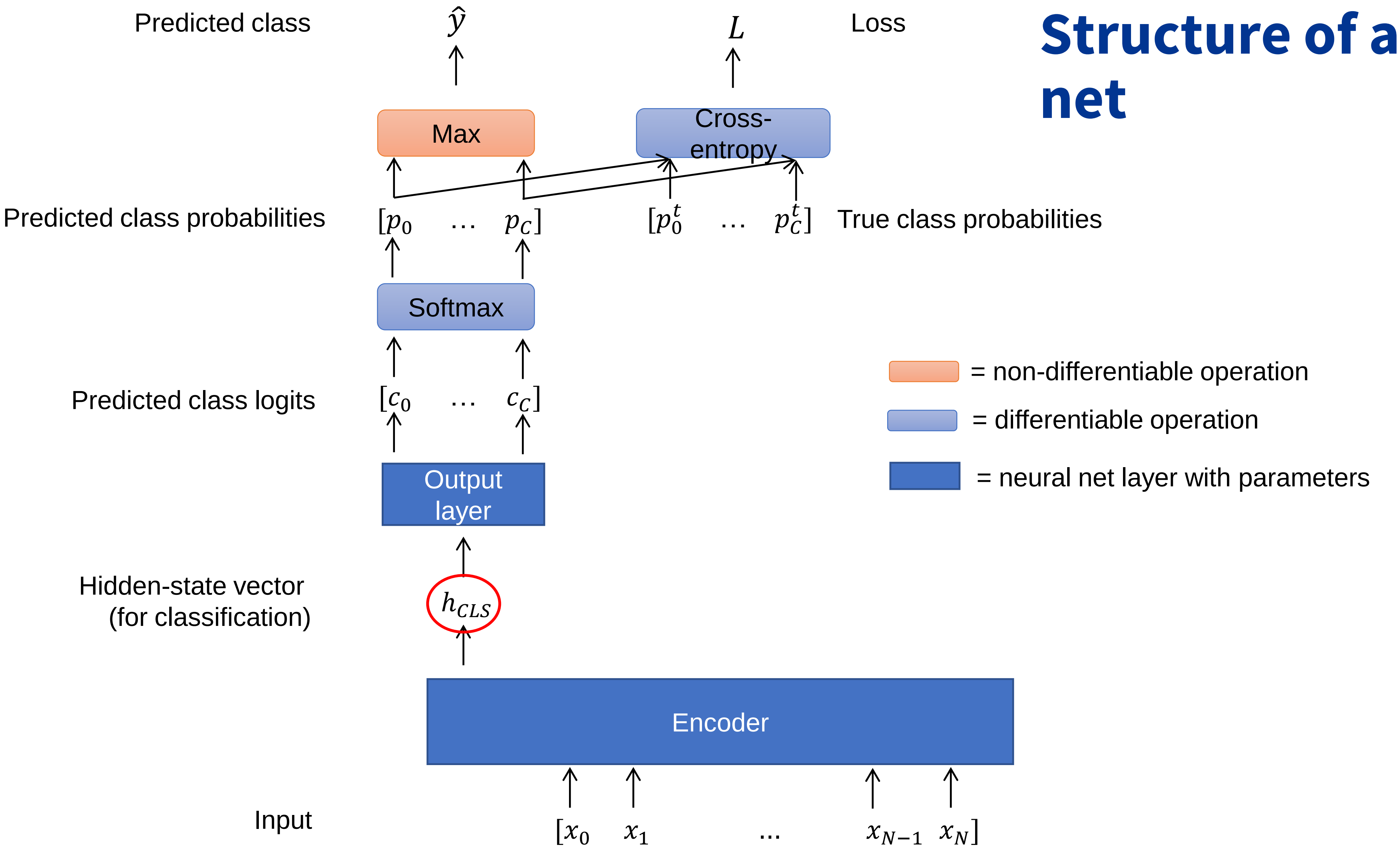
Deep k-Nearest Neighbors

Basic idea: retrieve examples that are close in terms of the final pre-classification hidden-state vector

Nicolas Papernot and Patrick McDaniel. 2018. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv:1803.04765 [cs, stat]*


Output layer
Hidden-state vector (for classification)
h_{CLS}
Encoder
Input
$[x_0 \quad x_1 \quad \ldots \quad x_{N-1} \quad x_N]$

Structure of a neural net



Predicted class $\hat{y}$

L Loss

Max

Cross-entropy

Predicted class probabilities $[p_0 \quad \dots \quad p_C]$

$[p_0^t \quad \dots \quad p_C^t]$ True class probabilities

Softmax

Predicted class logits $[c_0 \quad \dots \quad c_C]$

Output layer

Hidden-state vector (for classification) h_{CLS}

Encoder

Input $[x_0 \quad x_1 \quad \dots \quad x_{N-1} \quad x_N]$

🟧 = non-differentiable operation

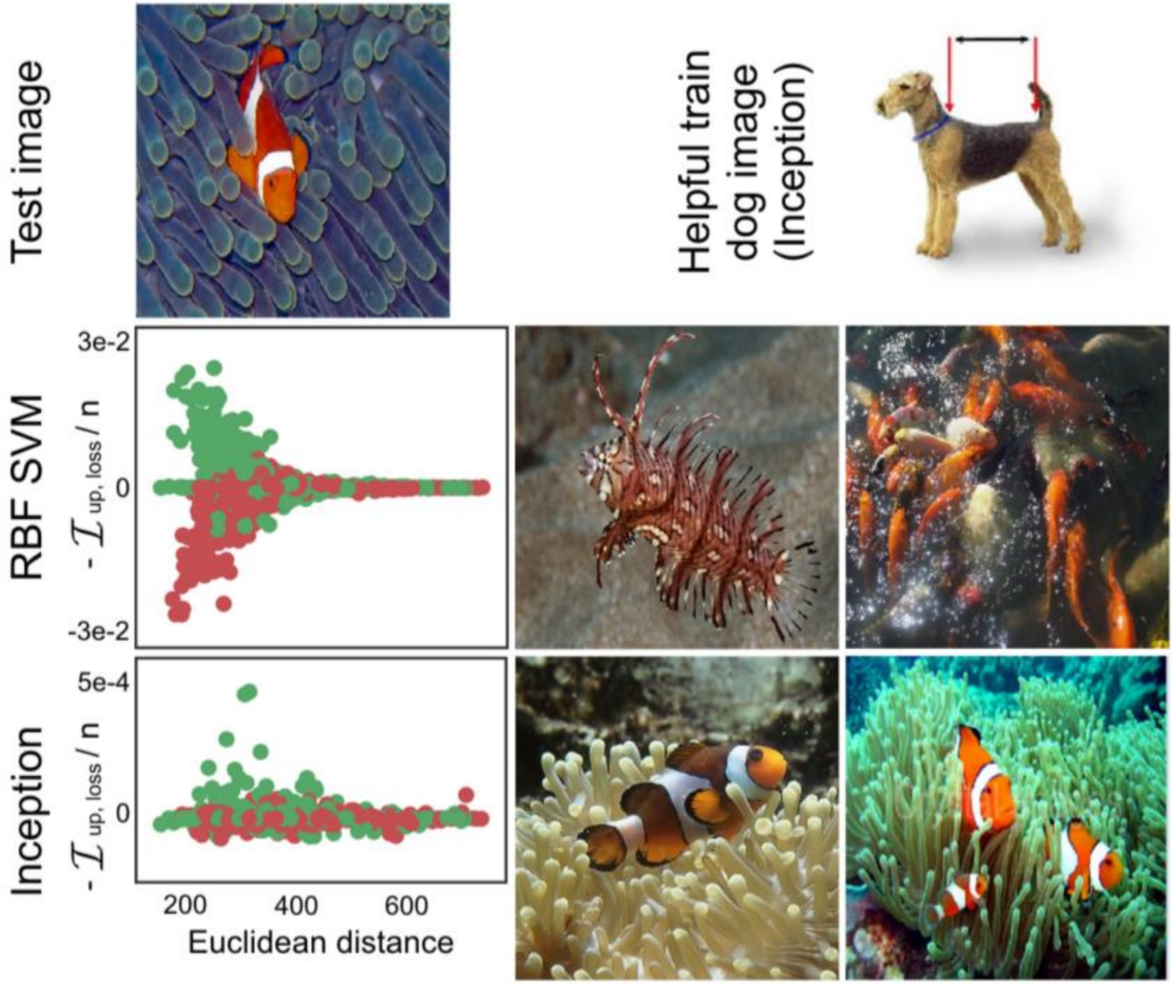
🟦 = differentiable operation

■ = neural net layer with parameters

Influence functions

Basic idea: use Hessian (second derivative) to assess how much each item in the training data would have affected the item-of-interest if it **hadn’t** been used in training

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. *arXiv:1703.04730*



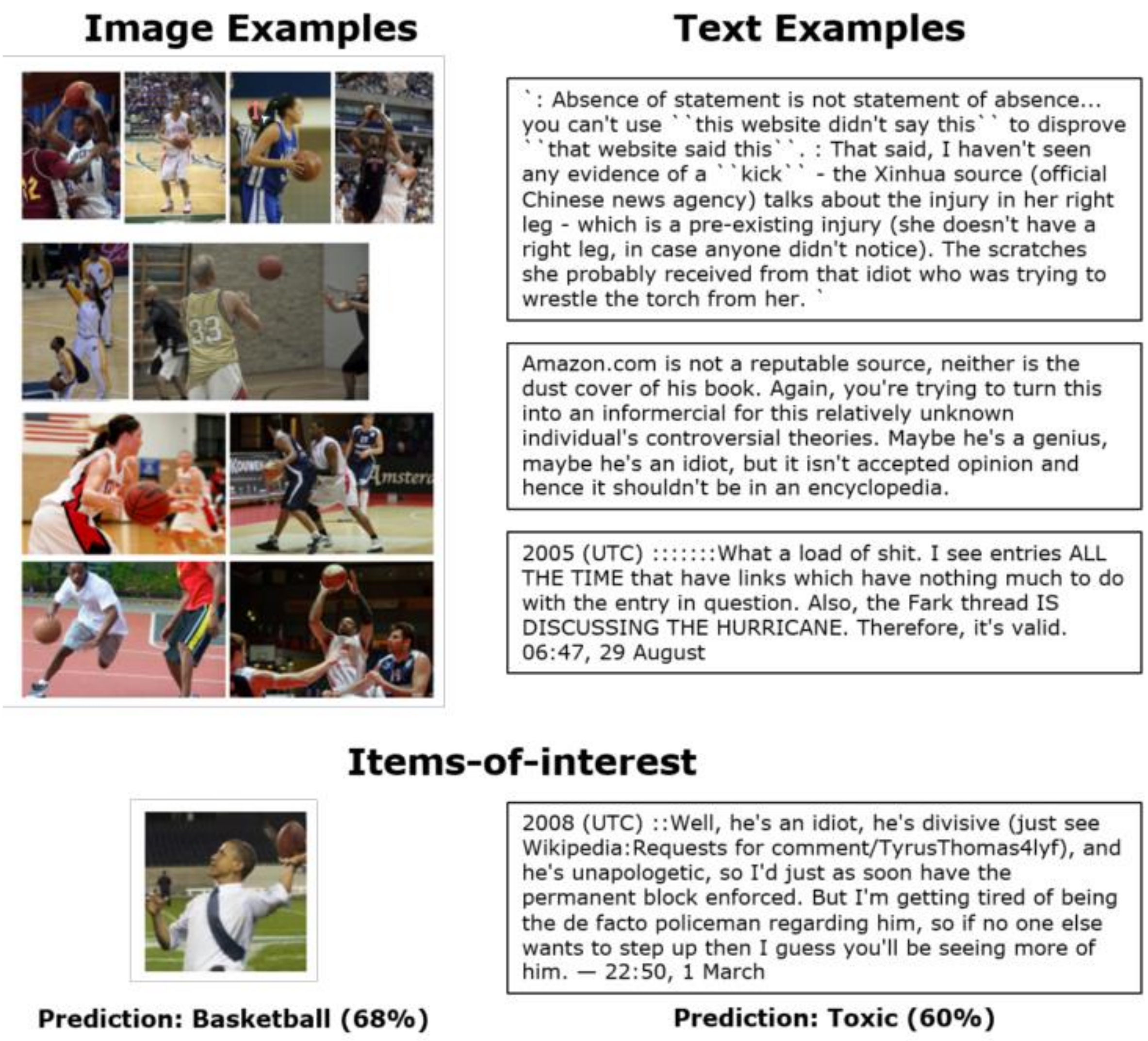
Effectiveness of explanatory examples

Much less well-studied.

Big problem: **information overload**

- Making humans read a bunch of extra stuff might conceivably make them more accurate…
- But at the cost of time and effort!

Naturally works better for images than text



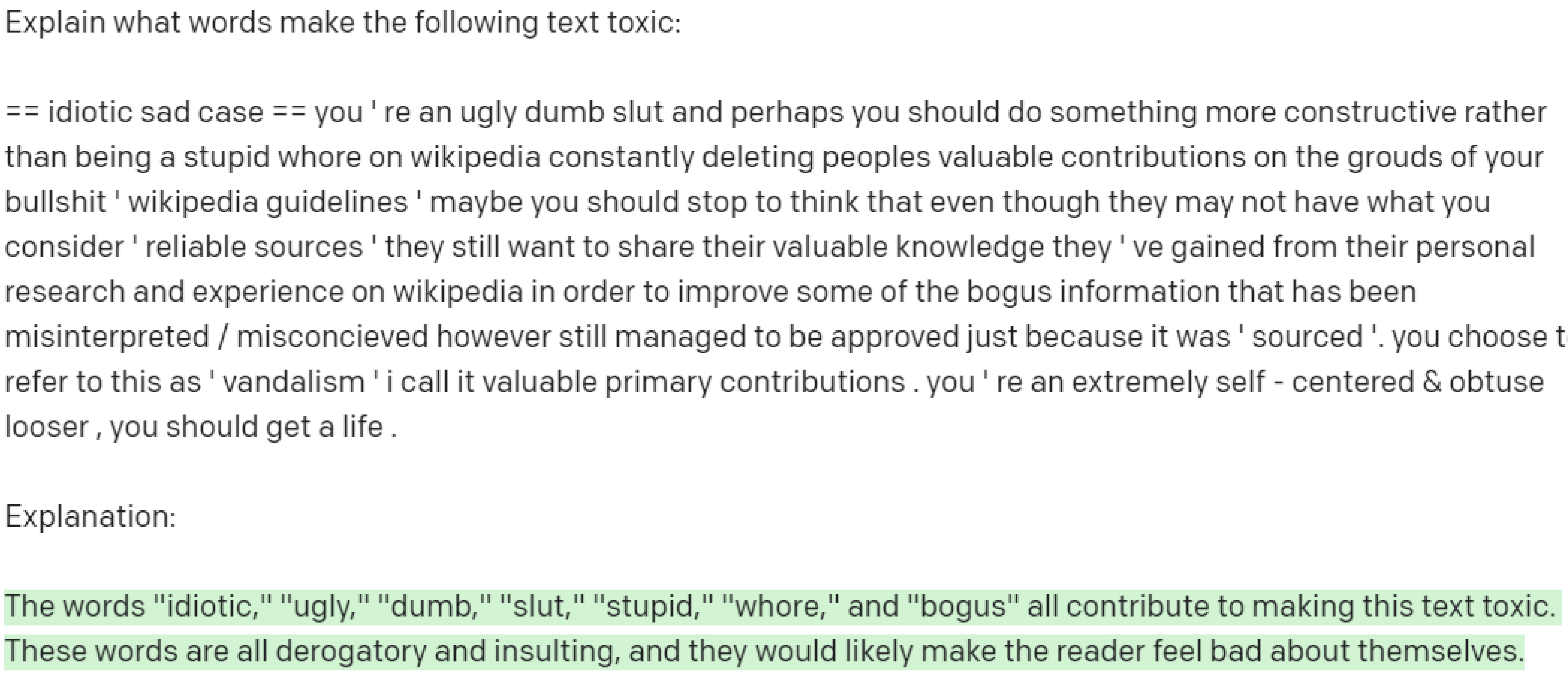
`: Absence of statement is not statement of absence... you can't use ``this website didn't say this`` to disprove ``that website said this``. : That said, I haven't seen any evidence of a ``kick`` - the Xinhua source (official Chinese news agency) talks about the injury in her right leg - which is a pre-existing injury (she doesn't have a right leg, in case anyone didn't notice). The scratches she probably received from that idiot who was trying to wrestle the torch from her. `

Amazon.com is not a reputable source, neither is the dust cover of his book. Again, you're trying to turn this into an informercial for this relatively unknown individual's controversial theories. Maybe he's a genius, maybe he's an idiot, but it isn't accepted opinion and hence it shouldn't be in an encyclopedia.

2005 (UTC) :::::::What a load of shit. I see entries ALL THE TIME that have links which have nothing much to do with the entry in question. Also, the Fark thread IS DISCUSSING THE HURRICANE. Therefore, it's valid. 06:47, 29 August

2008 (UTC) ::Well, he's an idiot, he's divisive (just see Wikipedia:Requests for comment/TyrusThomas4lyf), and he's unapologetic, so I'd just as soon have the permanent block enforced. But I'm getting tired of being the de facto policeman regarding him, so if no one else wants to step up then I guess you'll be seeing more of him. — 22:50, 1 March

Natural language explanations



Explain what words make the following text toxic:

== idiotic sad case == you ' re an ugly dumb slut and perhaps you should do something more constructive rather than being a stupid whore on wikipedia constantly deleting peoples valuable contributions on the grouds of your bullshit ' wikipedia guidelines ' maybe you should stop to think that even though they may not have what you consider ' reliable sources ' they still want to share their valuable knowledge they ' ve gained from their personal research and experience on wikipedia in order to improve some of the bogus information that has been misinterpreted / misconcieved however still managed to be approved just because it was ' sourced '. you choose to refer to this as ' vandalism ' i call it valuable primary contributions . you ' re an extremely self - centered & obtuse looser , you should get a life .

Explanation:

The words "idiotic," "ugly," "dumb," "slut," "stupid," "whore," and "bogus" all contribute to making this text toxic. These words are all derogatory and insulting, and they would likely make the reader feel bad about themselves.

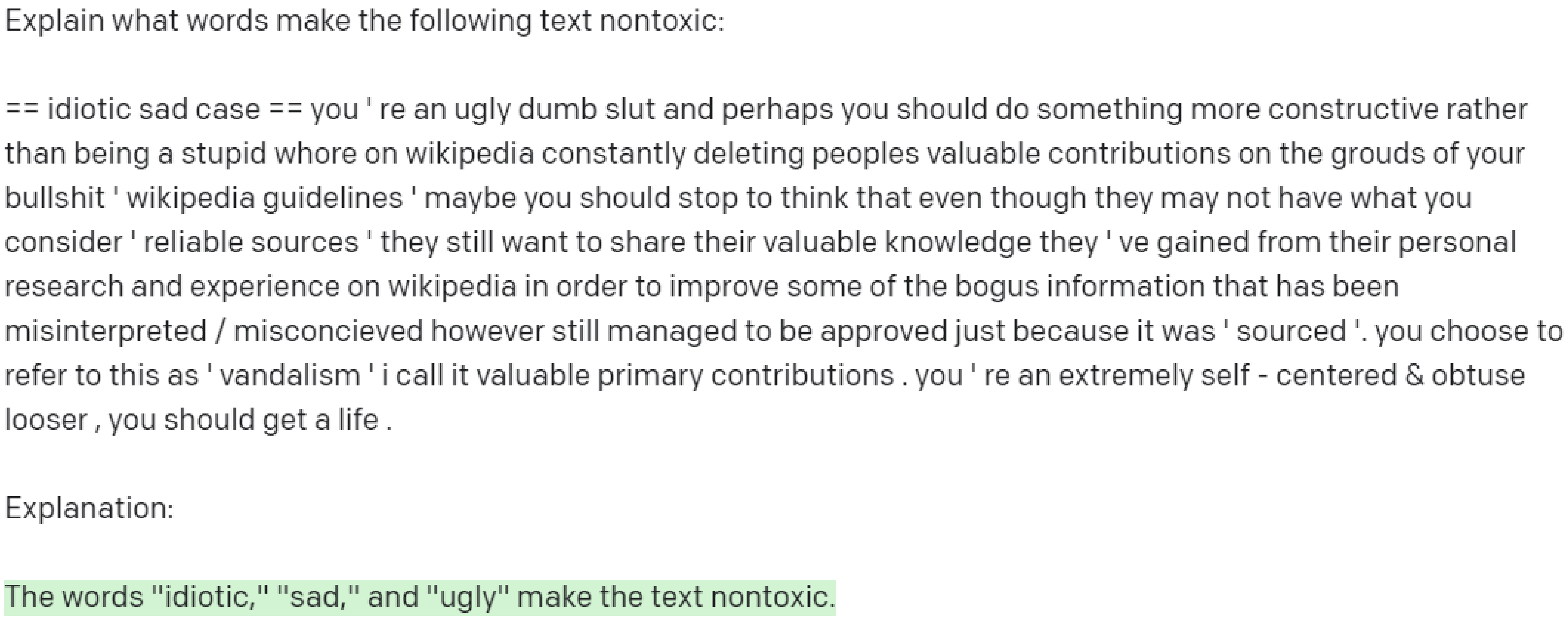
OpenAI playground, davincii-3

Natural language explanations

Now we're getting somewhere.

- Plausibility
  - This is how people explain things, so heck yeah
- Usefulness
  - Are human explanations useful?
    - I'll get back to you on this
- Faithfulness
  - Uh oh.
  - Problem is that a LLM explanation is **just more text** output from the model, and **doesn't necessarily have anything to do with the model's underlying logic**
  - So it might look humanlike, but make no sense

Natural language explanations



Explain what words make the following text nontoxic:

== idiotic sad case == you ' re an ugly dumb slut and perhaps you should do something more constructive rather than being a stupid whore on wikipedia constantly deleting peoples valuable contributions on the grouds of your bullshit ' wikipedia guidelines ' maybe you should stop to think that even though they may not have what you consider ' reliable sources ' they still want to share their valuable knowledge they ' ve gained from their personal research and experience on wikipedia in order to improve some of the bogus information that has been misinterpreted / misconcieved however still managed to be approved just because it was ' sourced '. you choose to refer to this as ' vandalism ' i call it valuable primary contributions . you ' re an extremely self - centered & obtuse looser , you should get a life .

Explanation:

The words "idiotic," "sad," and "ugly" make the text nontoxic.

OpenAI playground, davincii-3

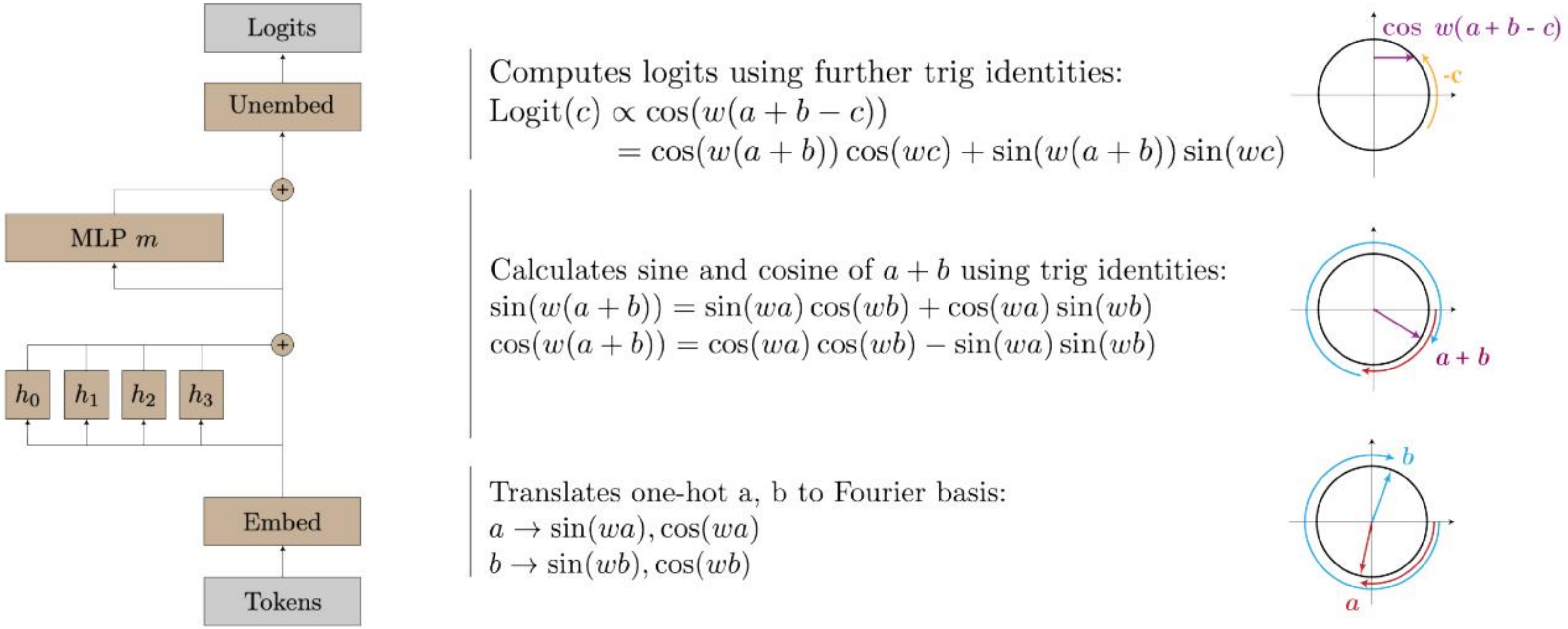
Mechanistic interpretability

Recent trend in ML/NLP

Basic idea: identify **circuits** that emerge within neural net parameters which implement recognizable algorithms to produce output

Currently seems to rely a lot on manual mapping, but automated methods being worked on

- Conmy, Arthur, et al. "Towards automated circuit discovery for mechanistic interpretability." Advances in Neural Information Processing Systems 36 (2023): 16318-16352.



Computes logits using further trig identities:
$$\mathrm{Logit}(c) \propto \cos(w(a+b-c))$$
$$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$$

Calculates sine and cosine of $a+b$ using trig identities:
$$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$$
$$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$$

Translates one-hot a, b to Fourier basis:
$$a \to \sin(wa), \cos(wa)$$
$$b \to \sin(wb), \cos(wb)$$

Nanda, Neel, et al. "Progress measures for grokking via mechanistic interpretability." *arXiv preprint arXiv:2301.05217* (2023).

Prompt engineering

Other forms of interpretability emerge from **prompt engineering**

Examples:

Chain of thought (CoT) prompting

- Tie model output to a series of intermediate steps that can be checked individually

Retrieval-augmented generation (RAG)

- Tie model output to a series of retrieved documents or snippets that can be checked individually

You can think of CoT and RAG as the generative equivalent of the Lei et al. (2016) rationale model: forcing the model output to be **grounded in some intermediate representation which then serves as an explanation.**

Usefulness

Million-dollar question: **are explanations useful?**

Do they make it _____ for humans trying to use them for stuff?

- Safer
- More ethical
- Easier to catch model mistakes
- Easier to learn new knowledge from

So far… not really. Positive results tend to be pretty conditional

- E.g. Vasconcelos, Helena, et al. "Explanations can reduce overreliance on ai systems during decision-making." Proceedings of the ACM on Human-Computer Interaction 7.CSCW1 (2023): 1-38.

Broader impact

Important for everyone to think about it because:

- Every company wants to use AI to automate things (recruiting, writing, documentation, coding, etc)
- But LMs have severe limitations (hallucination, logical mistakes, boringness)
  - Also ethical/legal issues (e.g. GDPR in Europe)
- So there’s a need for some kind of human oversight or verification
- But we don’t know how to do that yet
- Probably involves some form of interpretability

= opportunity for smart programmers!

Concluding thoughts

Interpretability probably needed for practical and ethical reasons

- But no one has figured out how to make them useful
- Seriously. At all.

Feature-based and example-based explanations have been the two basic types

Natural language explanations & mechanistic explanations more pertinent now

- But still unclear how to make them useful

Lots of work to do!