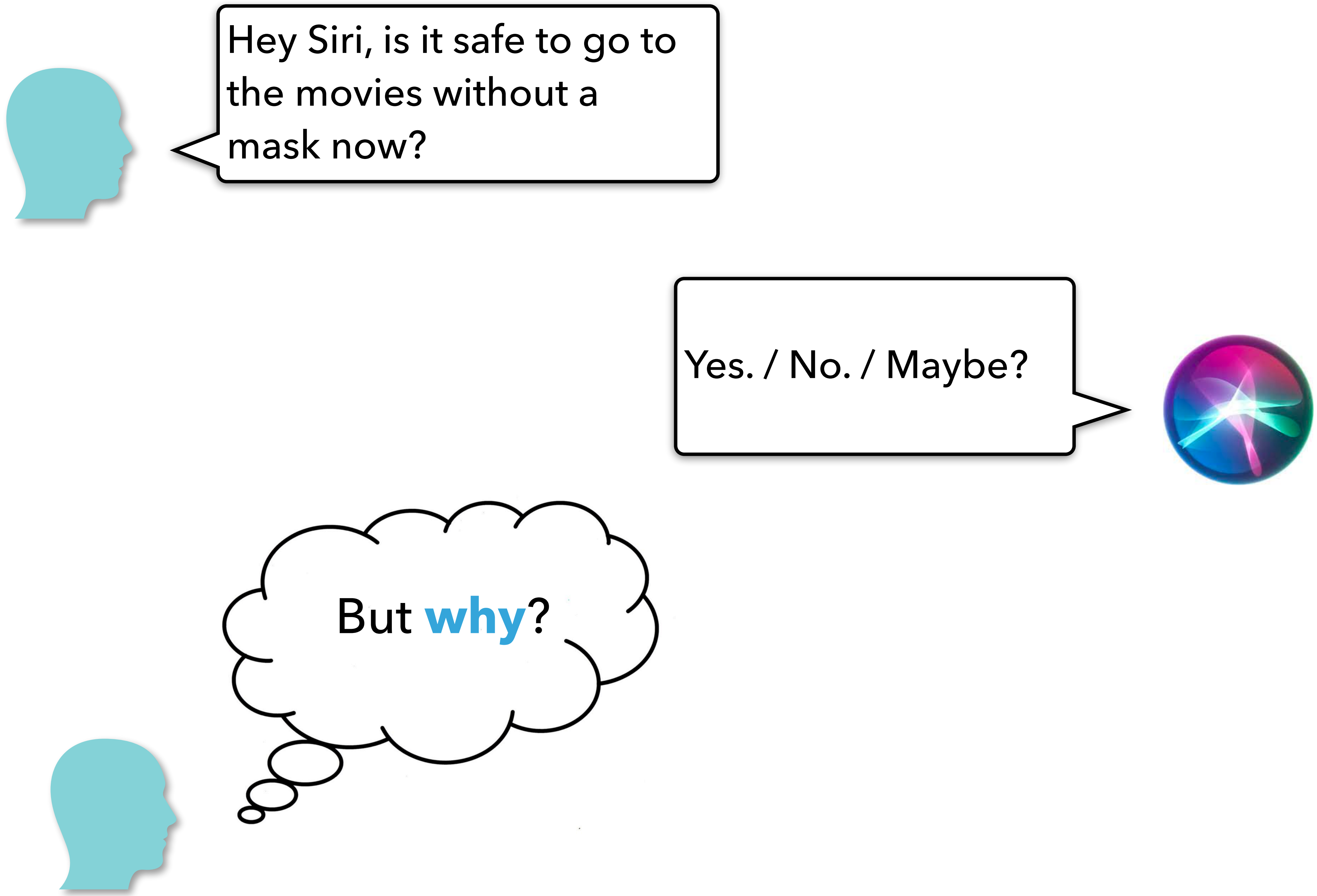
TOWARDS FAITHFUL MODEL EXPLANATION IN NLP:

A SURVEY

Veronica Qing Lyu

University of Pennsylvania


Hey Siri, is it safe to go to the movies without a mask now?
Yes. / No. / Maybe?
But **why**?

WE NEED EXPLAINABILITY

OUTLINE

INTRODUCTION

Background

- End-to-end Neural Networks (NNs) have achieved enormous success on a wide range of NLP tasks (e.g., GLUE/SuperGLUE benchmarks by Wang et al. 2018, 2019).
- But they largely remain a black-box to humans – lacking explainability.

What Is Explainability?

"The extent to which the internal mechanics of a model can be presented in understandable terms to a human."

(Lipton 2017; Murdoch et al. 2019; Barredo Arrieta et al. 2020)

What Is Explainability?

- *What* knowledge does the model encode?
- *Why* does the model make certain predictions? ★

"The extent to which the internal mechanics of a model can be presented in understandable terms to a human."

- Model developers
- Fellow researchers
- Industry practitioners
- End-users
- …

the target audience

(Lipton 2017; Murdoch et al. 2019; Barredo Arrieta et al. 2020)

What Is Explainability?

The extent to which *why a model makes certain predictions* can be presented in understandable terms to *some target audience*.

(Lipton 2017; Murdoch et al. 2019; Barredo Arrieta et al. 2020)

Why Is Explainability Important?

- Explainability allows us to …
  - Discover dataset artifacts
  - Diagnose a model's strengths and weaknesses, and debug it
  - Enhance user trust in high-stake applications

Properties of Explanations

- **Time**
  - post-hoc: Explanation is produced *after* the prediction.
  - built-in: Explanation produced *at the same time with* the prediction, i.e., the model is *self-explanatory*.

Properties of Explanations

- Time
- **Model accessibility**
  - black-box: Explanation method can *only* see the model's *input and output*.
  - white-box: Explanation method can *additionally* access the model *weights*.

Properties of Explanations

- Time
- Model accessibility
- **Scope**
    - local: Explains why a model makes a *single* prediction.
    - global: Explains the general reasoning mechanisms for the *entire data distribution*.

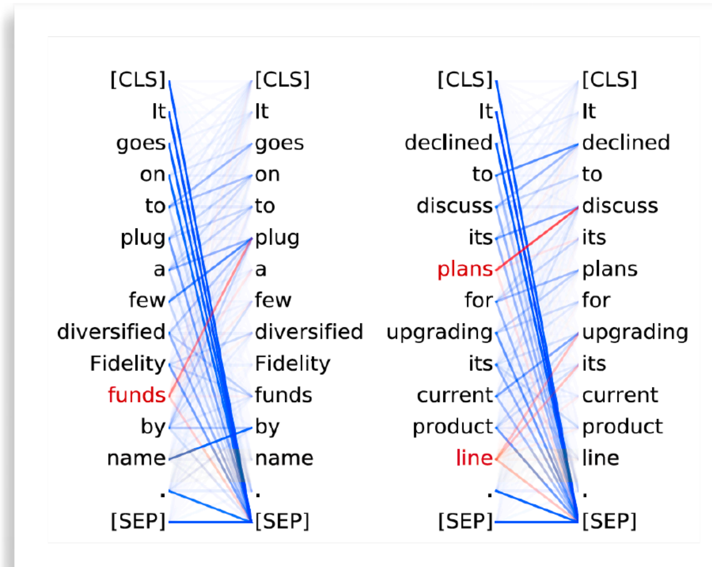
Properties of Explanations

- Time
- Model accessibility
- Scope
- **Unit of explanation**: what the explanation is in terms of
    - input features
    - examples
    - concepts[1]
    - feature interactions
    - combination
    - …

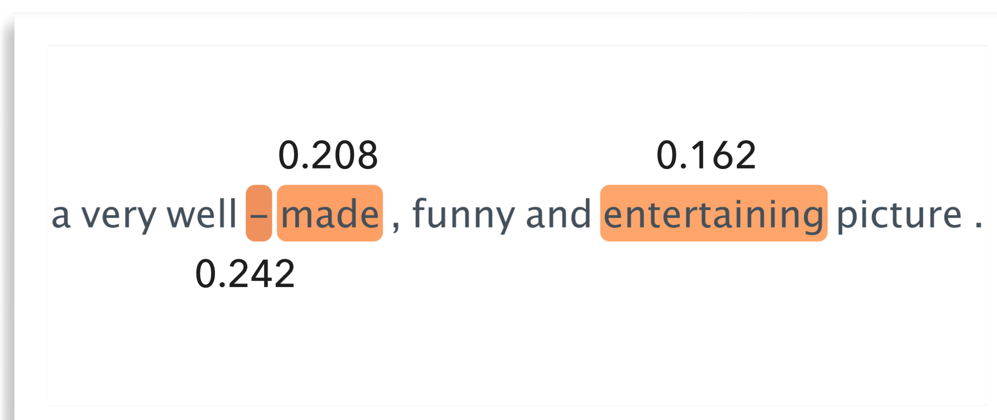
[1] Prior work has different definitions of *concepts*, including but not limited to phrases (Rajagopal et al. 2021) and high-level features (Jacovi et al. 2021).

Properties of Explanations

- Time
- Model accessibility
- Scope
- Unit of explanation
- **Form of explanation**
    - visualization
    - importance scores
    - natural language
    - causal graphs
    - ...



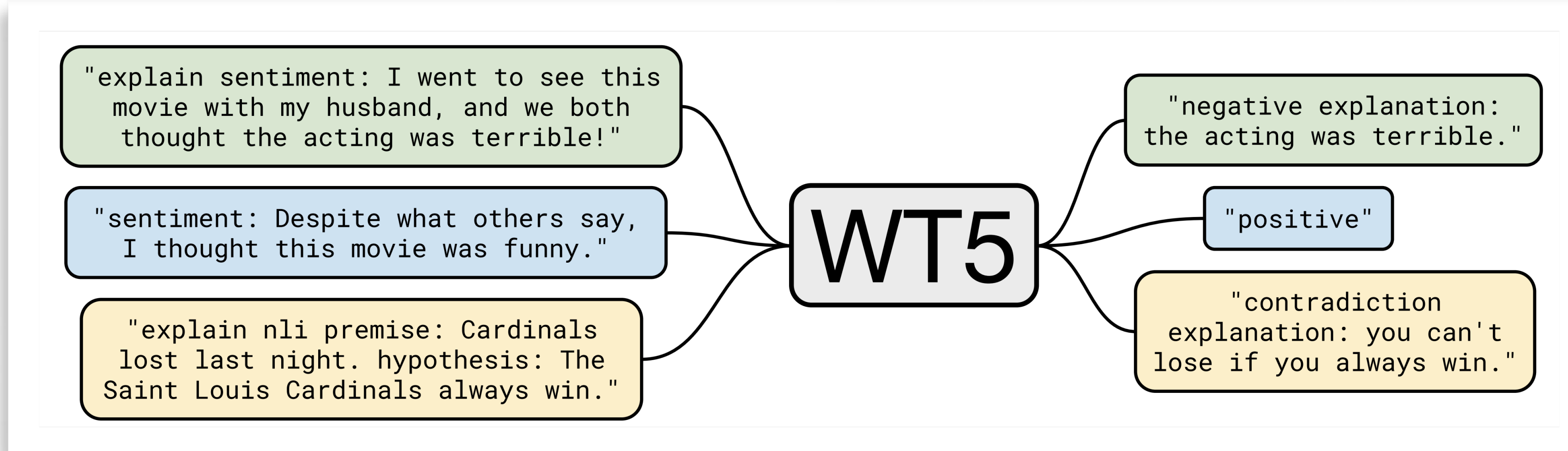
Visualization (Clark et al. 2019)



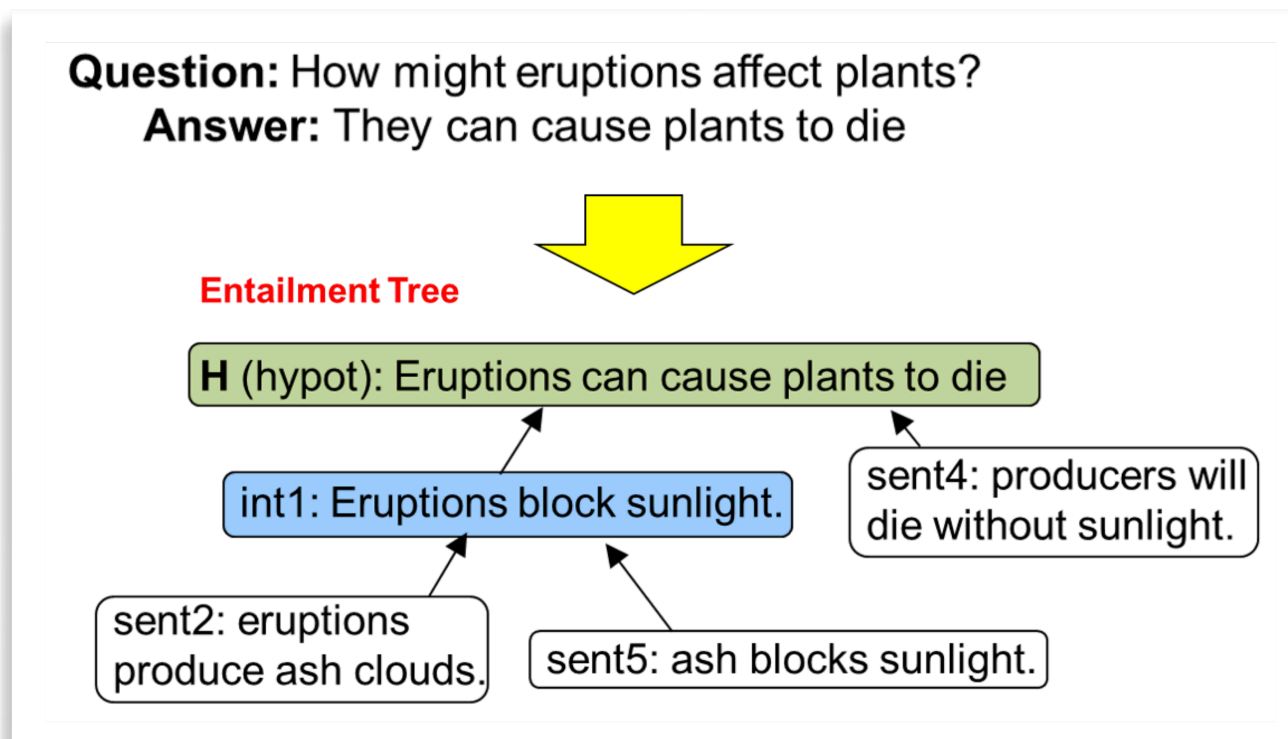
importance scores (AllenNLP Interpret)

Properties of Explanations

- Time
- Model accessibility
- Scope
- Unit of explanation
- **Form of explanation**
  - visualization
  - importance scores
  - natural language
  - causal graphs
  - ...



natural language (wT5, Narang et al. 2020)



causal graphs (EntailmentWriter, Dalvi et al. 2021)

Properties of Explanations

- Time
- Model accessibility
- Scope
- Unit of explanation
- Form of explanation
- **Target audience**
  - Model developers
  - Fellow researchers
  - Industry practitioners
  - End-users
  - …

Properties of Explanations

Don't worry; we'll elaborate later!

- A quick preview of what we'll cover:

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Similarity methods	post-hoc	white-box	local	examples, concepts	importance scores
Analysis of model-internal structures	post-hoc	white-box	local, global	features, interactions	visualization, importance scores
Backpropagation -based methods	post-hoc	white-box	local	features, interactions	visualization, importance scores
Counterfactual intervention	post-hoc	black-box, white-box	local, global	features, examples, concepts	importance scores
Self-explanatory models	built-in	white-box	local, global	features, examples, concepts	importance scores, natural language, causal graphs

Table 1: Comparison of different model explanation methods in terms of their properties.

Principles of Explanations

- Faithfulness
- Plausibility
- Input Sensitivity
- Model Sensitivity
- Completeness
- Minimality
- …

See §1.1.4

What Is Faithfulness?

(aka. fidelity, reliability)

i.e., it can't lie

An explanation should accurately reflect the reasoning process behind the model's prediction.

(Harrington et al. 1985 ; Ribeiro, Singh, and Guestrin 2016 ; Jacovi and Goldberg 2020)

What Is Plausibility?

(aka. persuasiveness, understandability)

An explanation should be understandable and convincing to the target audience.

(Herman 2019 ; Jacovi and Goldberg 2020)

Faithfulness vs. Plausibility

- Commonality: No established formal definition for either principle yet.
- Tension:

  - Plausibility doesn’t imply Faithfulness; and vice versa.
    (They are not necessarily incompatible, though.)

Why Is Faithfulness Important?

- **Faithfulness establishes causality**
  - "what is encoded" ≠ "what is used "

    correlational causal

  - LMs encode linguistic features even when they are irrelevant to the end task labels (Ravichander et al., 2021)

Why Is Faithfulness Important?

- Faithfulness establishes causality
- **An unfaithful explanation can be dangerous**
  - Especially if it is plausible (i.e., appealing to humans)!
  - Humans would still trust the model, even if it does not work in the way we want
    - e.g. Attention-based explanations can be deceiving to users, by hiding the model's gender bias (Pruthi et al., 2020)

How Do We Measure Faithfulness?

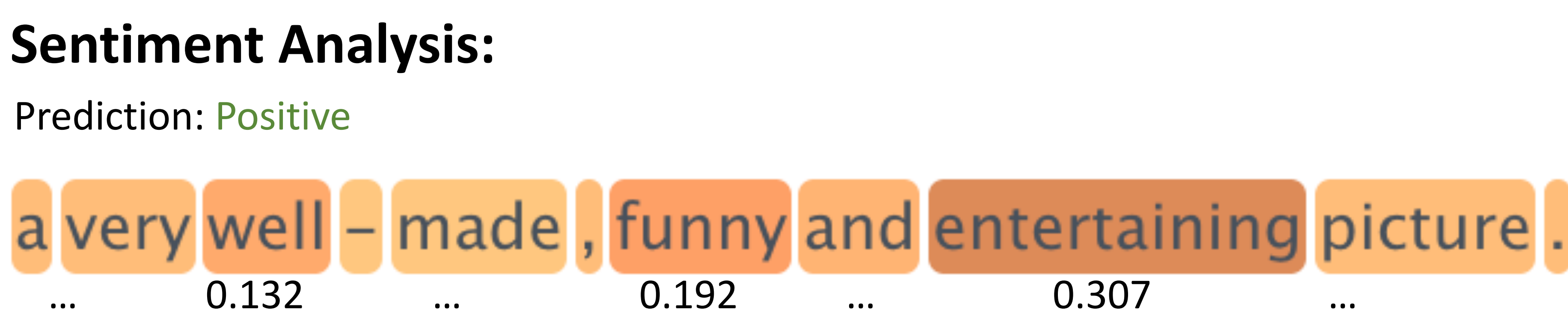
No established consensus yet!

- (a) Axiomatic evaluation ✅
- (b) Predictive power evaluation ✅
- (c) Robustness evaluation
- (d) Perturbation-based evaluation ✅
- (e) White-box evaluation ✅
- (f) Human perception evaluation

✅: recommened (with caveat — see §1.2.4 for more details)

How Do We Measure Faithfulness?

- Perturbation-based evaluation
  - Given a feature importance ranking, generated by an explanation method



  - Remove a fixed proportion of features from the input, based on the ranking
    - most important features are first removed → we expect a larger change in model prediction
    - least important features are first removed → we expect a smaller change in model prediction
    - random features are first removed → we expect the change to be somehwere in the middle

PRIOR ATTEMPTS AT FAITHFUL EXPLANATION

Five Categories

- Similarity methods
- Analysis of model-internal structures
- Backpropagation-based methods
- Counterfactual intervention
- Self-explanatory models

We'll only elaborate on **a few representative works** in each category
See §2 for a total of 90+

Running Example

Sentiment Analysis:

"The movie is great. I love it."

Prediction: Positive

Our goal:

What **features** (e.g., tokens) are **most important** for the model's prediction?

Five Categories

- **Similarity methods**
- Analysis of model-internal structures
- Backpropagation-based methods
- Counterfactual intervention
- Self-explanatory models

Similarity methods

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Similarity methods	post-hoc	white-box	local	examples, concepts	importance scores

- For a given test example, find its most similar training examples in the model's learned representation space to justify the current prediction

NOT the input feature space!

- Akin to how humans justify their actions by analogy

Similarity methods

- Running example

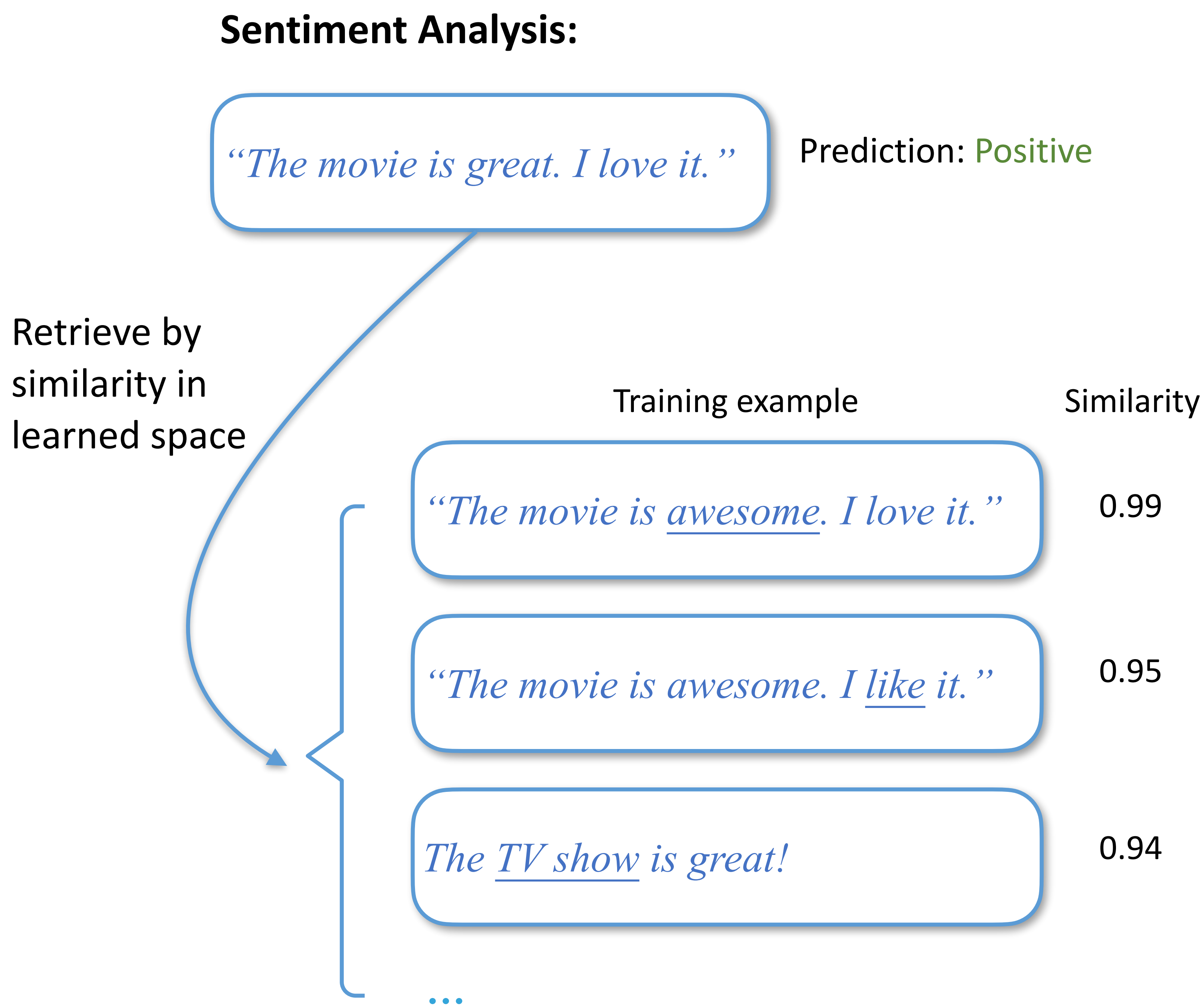


Figure 1: Visualization of a similarity method on the running example.

Similarity methods

- Past work[2]:
  - Caruana et al. (1999): **theoretically formalize** the earliest similarity method, searching for test example's k-Nearest Neighbors (**kNN**) in the training set
  - Wallace et al. (2018): **replace** the original model's final softmax classifier with a kNN classifier at test time
  - Rajagopal et al. (2021): find most similar ***concepts*** (phrases in this case) instead of whole examples in the training set

[2] See detailed distinctions in §2.2.2.

Similarity methods

- **Advantages**
  - (a) Intuitive to understand
  - (b) Easy to implement, as no re-training or data manipulation is needed
  - (c) Highly model-agnostic and metric-agnostic

Similarity methods

- **Disadvantages**
  - (a) only provide *the outcome of the model's reasoning process* (i.e., which examples are similar in the learned space), but not *how the model reasons* (i.e., how the space is learned).
  - (b) Evaluated mostly with Plausibility, but rarely with Faithfulness
    - No guarantee that the model reasons in a similar way for similar examples!

Five Categories

- Similarity methods
- **Analysis of model-internal structures**
- Backpropagation-based methods
- Counterfactual intervention
- Self-explanatory models

Analysis of model-internal structures

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Analysis of model-internal structures	post-hoc	white-box	local, global	features, interactions	visualization, importance scores

- What **structures**?
  - neurons
  - layers
  - specific mechanisms e.g., convolution, attention, etc.
- How to **analyze**?
  - visualization: activation heatmaps, information flow, …
  - clustering: neurons with similar functions, inputs with similar activation patterns, …
  - correlation analysis: between neurons and linguistic properties
  - …

Analysis of model-internal structures

- Running example

Sentiment Analysis:

"The movie is great. I love it."

Prediction: Positive

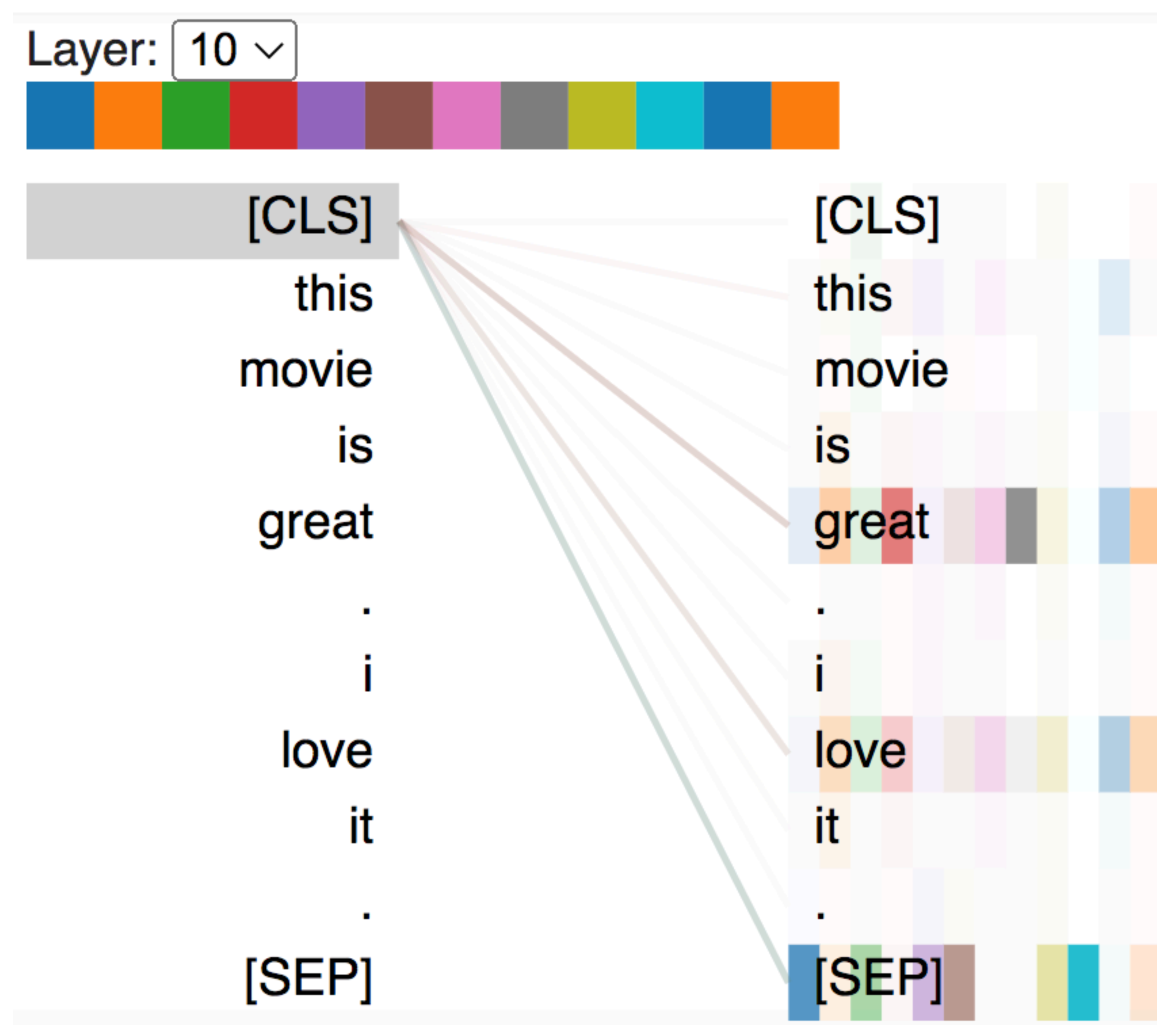


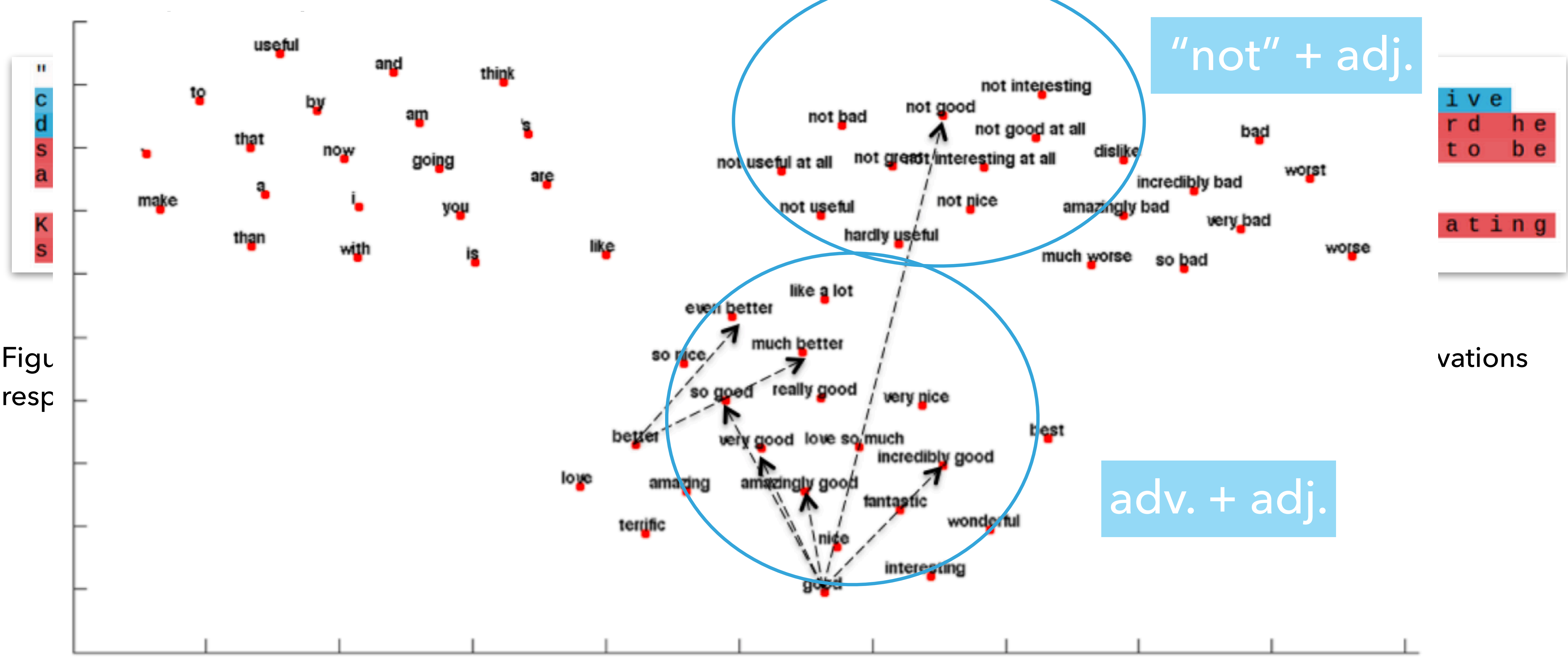
Figure 2: Attention weight visualization on the running example (generated with BertViz).

Analysis of model-internal structures

- **Past work**
  - Pre-attention era
  - Post-attention era

Analysis of model-internal structures

- **Pre-attention era**
  - Neurons with “specific purposes”: (Karpathy et al. 2015), (Strobelt et al. 2018)
  - Inputs with similar activation patterns: (Li et al. 2016), (Poerner et al. 2018),



Figu
resp
vations

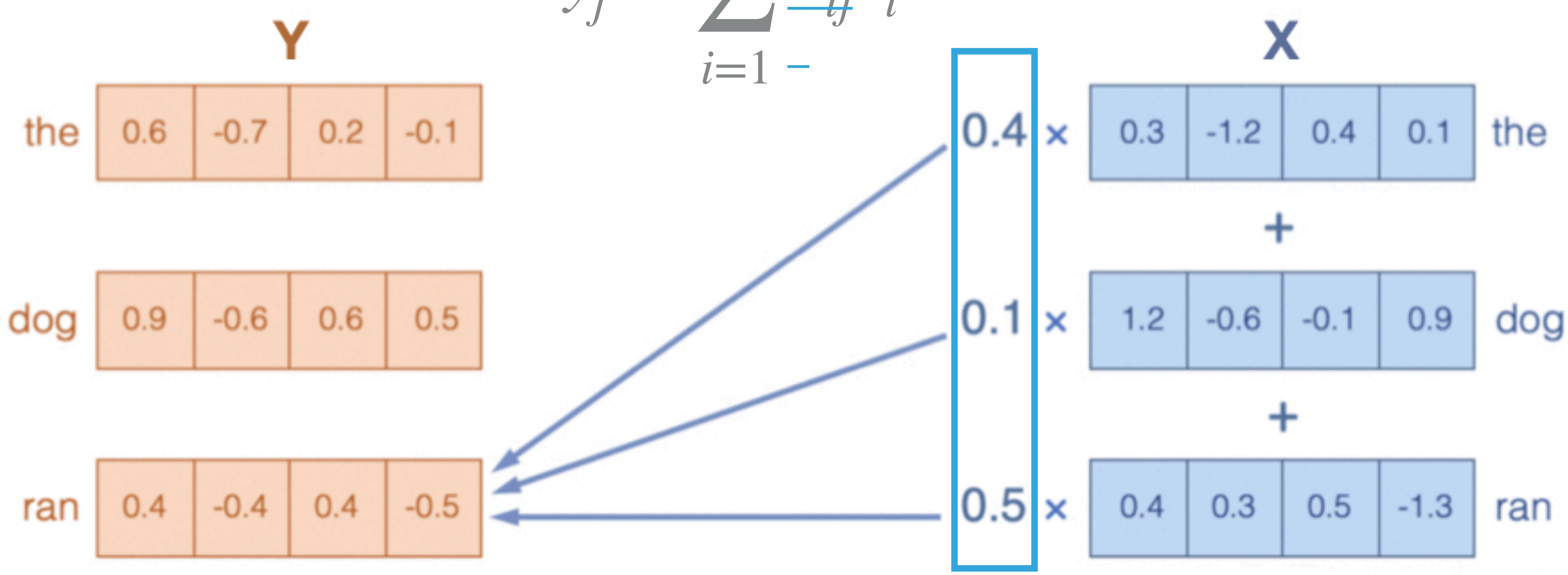
Figure 4: t-SNE visualization on latent representations for intensifications and negations (Li et al. 2016).

Analysis of model-internal structures

- **Post-attention era**
  - the core of **Transformers** (Vaswani et al. 2017)
  - The Attention mechanism (Bahdanau et al. 2015)
    - A sequence-to-sequence function
      Input: $x = < x_1, x_2, \ldots, x_n >$
      Output: $y = < y_1, y_2, \ldots, y_n >$

attention weights[3]

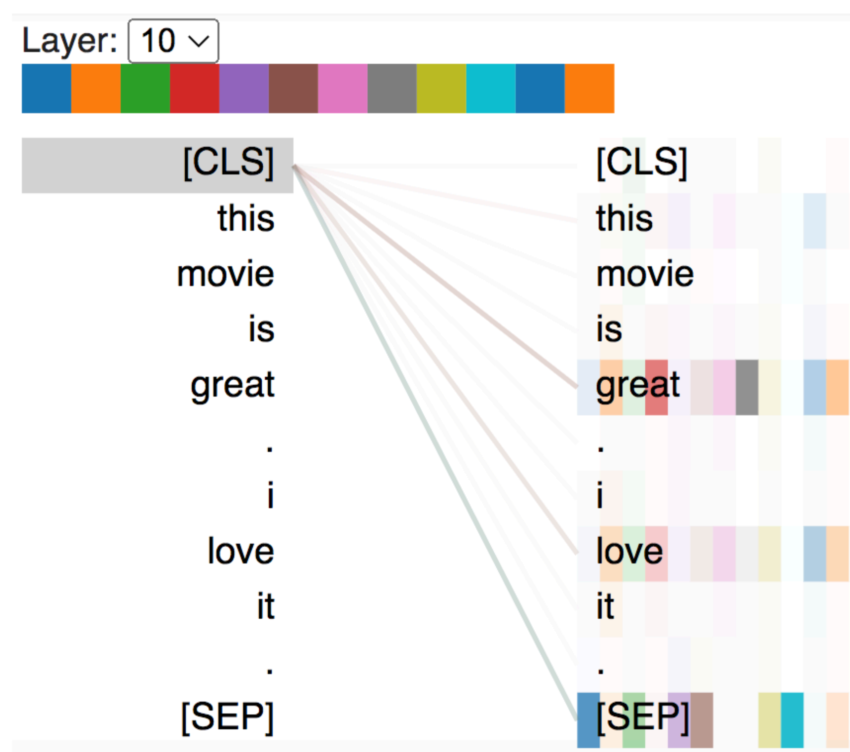
$$y_j = \sum_{i=1}^{n} \underline{a_{ij}} x_i$$



[3] Computed with a *compatibility function*.

Analysis of model-internal structures

- **Post-attention era**
  - Attention weight a_{ij}: how much the output "attends to" each input feature representation x_i
    - This is often intuitively seen as an explanation of feature importance for model prediction (Xu et al., 2015; Choi et al.,2016; Lei et al., 2017; Martins and Astudillo 2016; Xie et al. 2017; Mullenbach et al. 2018; …)



"These are the most important tokens!"

🤔 But is it **really** so?

Analysis of model-internal structures

- **Post-attention era**
  - Debate on Faithfulness
    - "Attention is **not** explanation" (Jain and Wallace 2019)
      - One can construct "adversarial attention distribution": *maximally* different from the original distribution, but *minimally* influence the prediction

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

original α

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

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

adversarial $\tilde{\alpha}$

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

Figure 6: A sentiment analysis model's original and adversarial attention distribution over words in a negative movie review.

Analysis of model-internal structures

- **Post-attention era**
  - Debate on Faithfulness
    - "Attention is **not not** explanation" (Wiegreffe and Pinter 2019)
      - Adversarial distributions are not adversarial weights: it's hard for the model to converge to these adversarial distributions through natural training
    - many, many followups …
      - Well, it's not that hard (Pruthi et al. 2020)
      - It's possible to remedy attention towards a more faithful explanation (Tutek and Snajder, 2020; Hao et al. 2021)
      - See §2.3.2 for more details

Analysis of model-internal structures

- **Advantages**
  - (a) Intuitive to understand
  - (b) Easily accessible and computationally efficient
  - (c) Many interactive tools available, helping the user form hypotheses
  - (d) Attention can capture the interaction between features, while many other methods only capture flat importance scores of individual features

Analysis of model-internal structures

- **Disadvantages**
  - (a) Questionable Faithfulness
  - (b) Attention weights on are hidden states (≠input features), which already incorporates contextual information
  - (c) Only captures what happens at a single time step, w/o taking the whole computation path into account

Five Categories

- Similarity methods
- Analysis of model-internal structures
- **Backpropagation-based methods**
- Counterfactual intervention
- Self-explanatory models

Backpropagation-based methods

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Backpropagation -based methods	post-hoc	white-box	local	features, interactions	visualization, importance scores

- Two subcategories:
  Gradient methods & **Propagation methods**
  - **Commonality**: Both identify the contribution of input features via a backward pass, propagating the *importance* (or *relevance*) from the output to the input layer
  - **Difference**: The former follow standard backpropagation (BP) rules, while the latter define custom backpropagation rules depending on each layer type

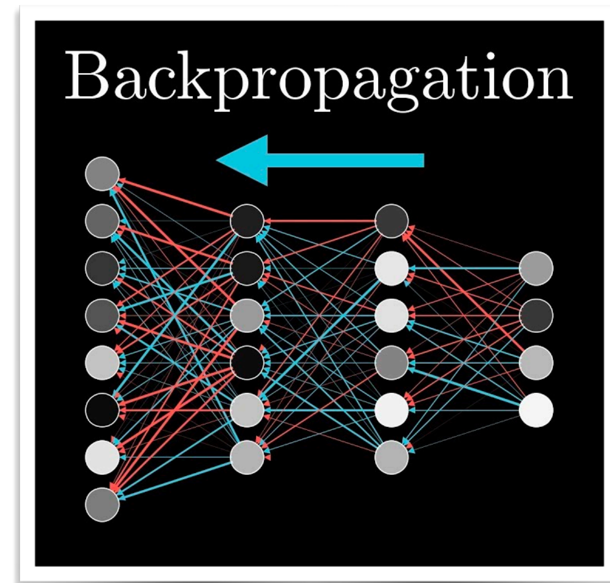


Figure from 3Blue1Brown

Backpropagation-based methods

- Running example

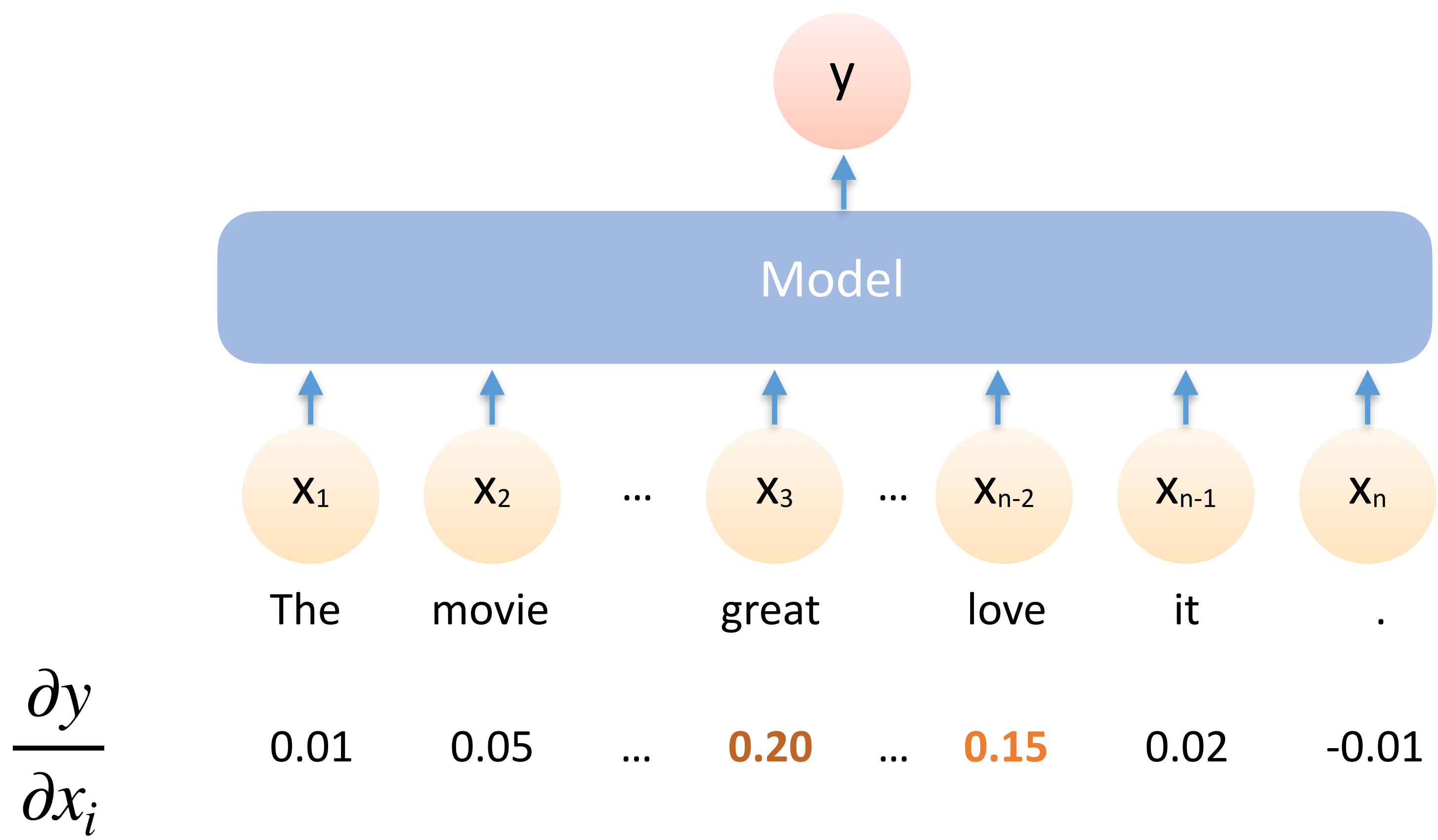


Figure 7: Visualization of a backpropagation-based method (Simple Gradients) on the running example.

Backpropagation-based methods

- Most ideas of this family originated in Computer Vision (CV).
- Notations:
  - x : input example
  - x_i : input features
  - M: the model
  - $y = M(x)$: the model's prediction
  - $r_i(x)$: the *relevance* of each feature x_i to y
  - $\bar{x}$ (optional): *baseline* input to compare against x (e.g., all-black image, all-zero sentence)

Please keep these in mind as we're going to use them later!

Backpropagation-based methods

- **Gradient methods**
  - Follow standard BP rules ⇒ treat the gradient (or some variant of it) of the model output w.r.t each input feature as its relevance
  - Intuition: gradient represents *how much difference a tiny change in the input will make to the output*
  - Specific gradient methods differ in how they calculate $r_i(x)$, the *relevance* of each feature x_i

Backpropagation-based methods

- **Simple Gradients / Vanilla Gradients**
  - The relevance is just the gradient itself:

$$r_i(x) = \frac{\partial M(x)}{\partial x_i}, \ \|\frac{\partial M(x)}{\partial x_i}\|_1, \text{ or } \|\frac{\partial M(x)}{\partial x_i}\|_2$$

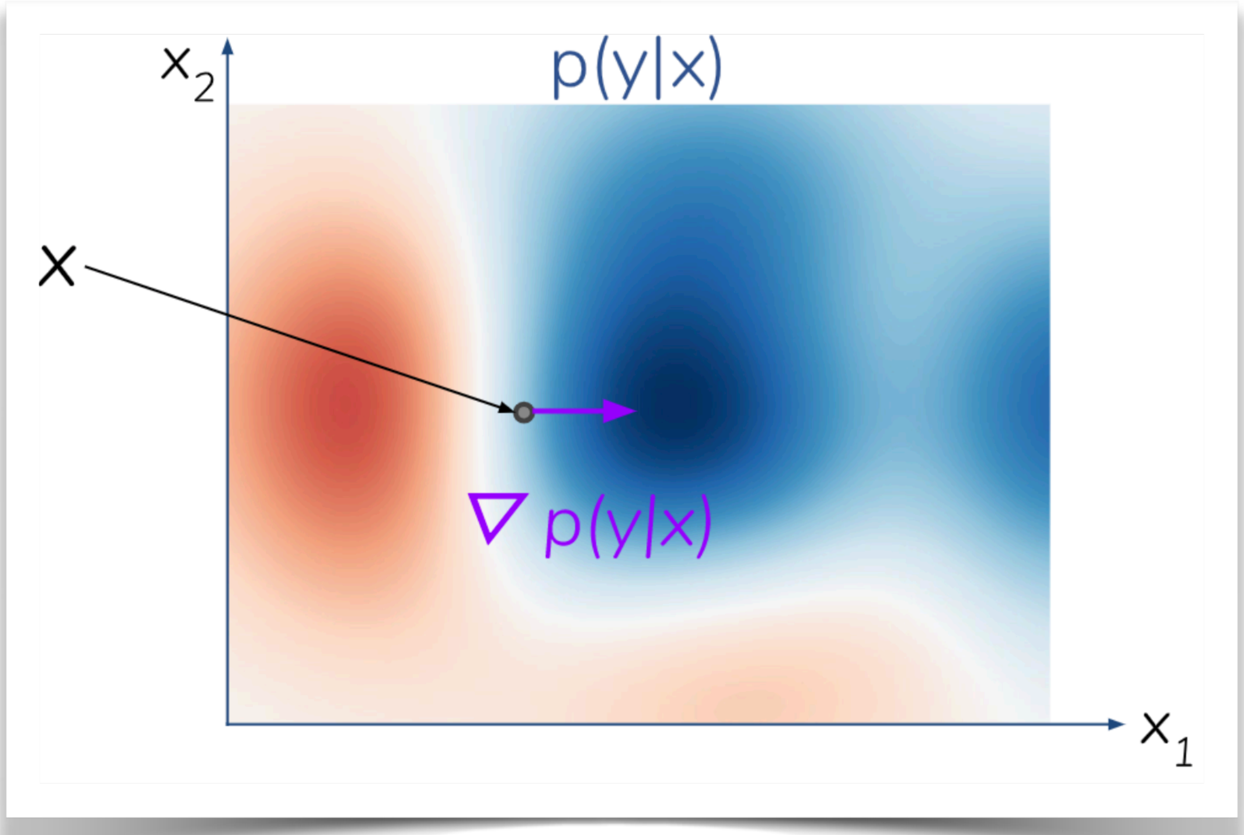


Figure from EMNLP 2020 interpretability tutorial

(Baehrens et al. 2010; Simonyan, Vedaldi, and Zisserman 2014)

Backpropagation-based methods

- **Simple Gradients / Vanilla Gradients**
  - Problems:
    - Only measures the *sensitivity* of the output w.r.t changes in the feature, but not the *contribution* of the feature to the output
      - e.g., saturation
    - Too "local": the gradient can change drastically with subtle changes in the input

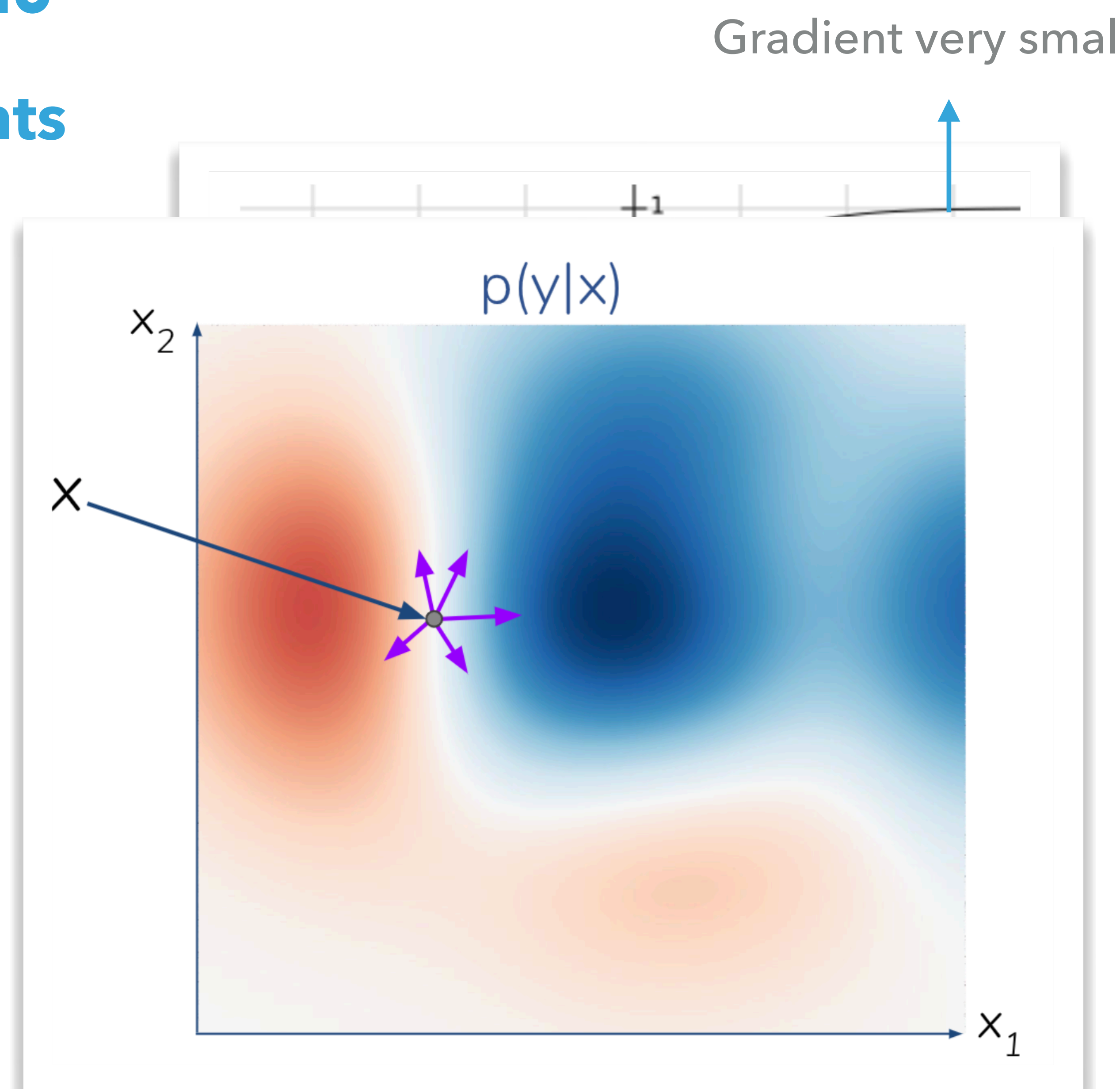


Figure from EMNLP 2020 interpretability tutorial

Backpropagation-based methods

- **Gradient×Input**
  - The relevance is the inner product of gradient & input:

$$r_i(x) = x_i \odot \frac{\partial M(x)}{\partial x_i}$$

  - This is to measure the *contribution* of the feature to the output, instead of the *sensitivity* of the output to changes in the feature

(Denil et al. 2015)

Backpropagation-based methods

- **Gradient×Input**
  - Problems
    - Fails the Input Sensitivity test (Sundararajan et al. 2017) (cf. § 1.1.4): If two inputs differ only at one feature and lead to different model predictions, then the explanation should assign non-zero importance to the feature.
      - e.g.[4] Suppose the model is

$$M(x) = 1 - \max(0, 1-x).$$

Then we have

$$M(0) = 0,$$
$$M(2) = 1.$$

However,

$$\text{Gradient}\times\text{Input}(0) = 0,$$
$$\text{Gradient}\times\text{Input}(2) = 0$$

[4] Example from (Sundararajan et al. 2017)

Backpropagation-based methods

- **Integrated Gradients**
  - Average gradients along path from baseline to input:

$$r_i(x) = \underbrace{(x_i - \overline{x}_i)}_{4} \odot \underbrace{\int_{\alpha=0}^{1} \overbrace{\frac{\partial M(\overbrace{\overline{x}+\alpha(x-\overline{x})}^{1})}{\partial x_i}}^{2} d\alpha}_{3}$$

1. Interpolate points between baseline $\overline{x}$ and input x
2. Compute gradient for each interpolated point
3. Compute integral (approximated by summation)
4. Rescale

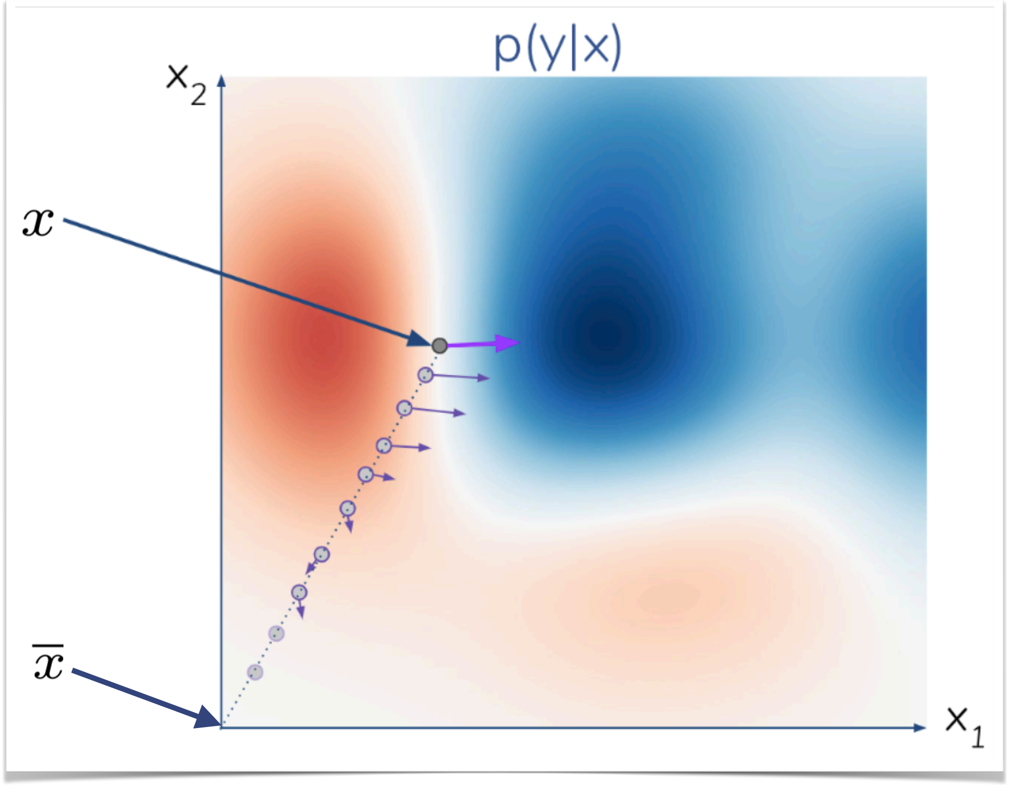


Figure from EMNLP 2020 interpretability tutorial

(Sundararajan et al. 2017)

Backpropagation-based methods

- **Integrated Gradients**
  - Problems
    - still visually noisy …
      - maybe due to the "too local" problem? (Smilkov et al. 2017)

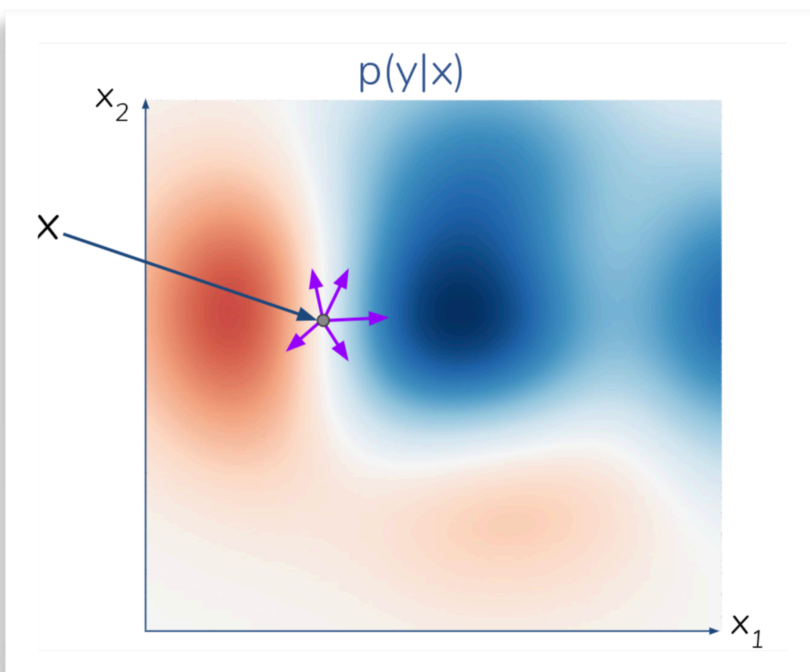
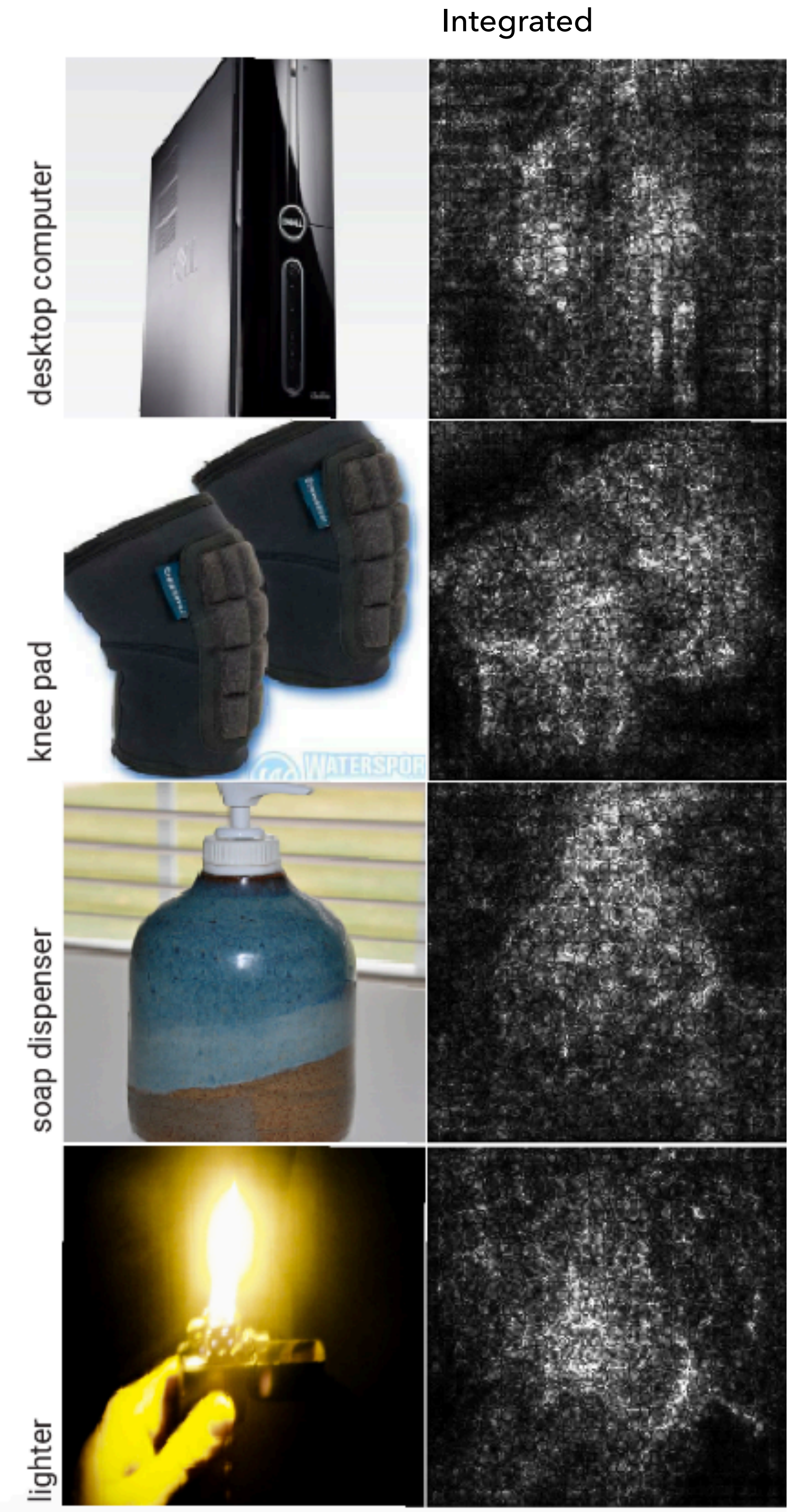


Figure from EMNLP 2020 interpretability tutorial

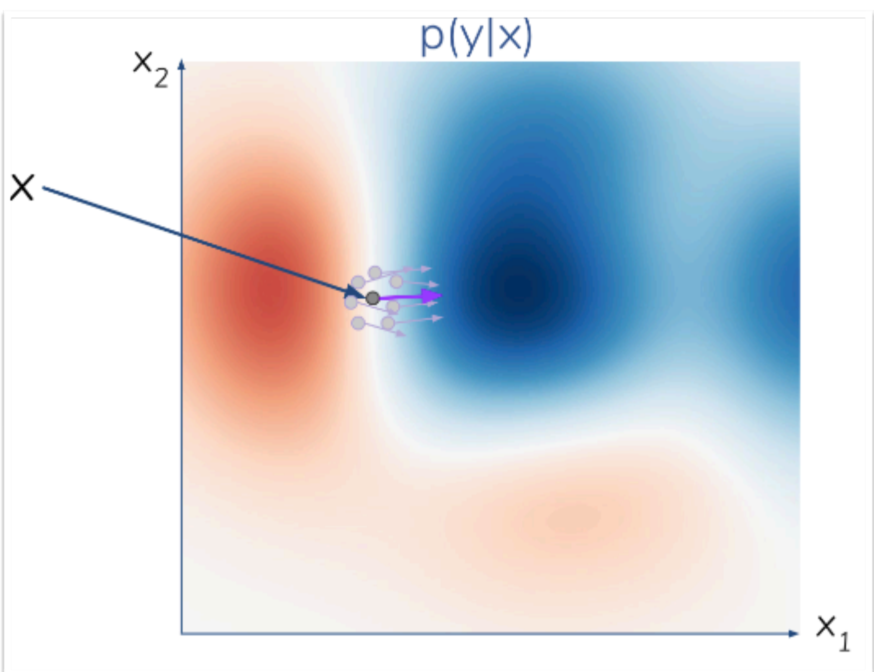


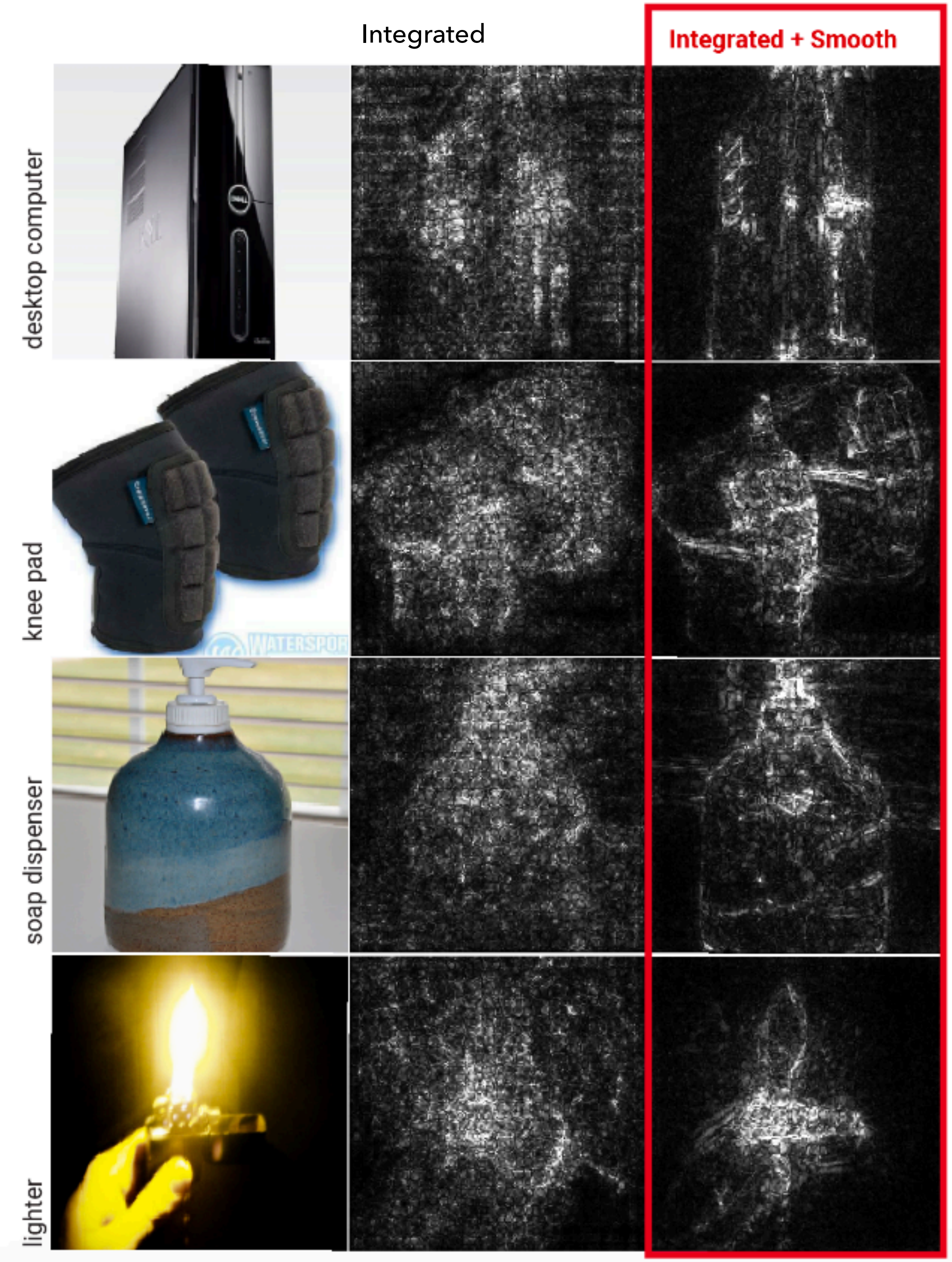
Backpropagation-based methods

- **SmoothGrad**
  - Add Gaussian noise to the input and average the gradients:

$$r_i(x) = \frac{1}{m} \sum_1^m \hat{r}_i(x)(x + \mathcal{N}(0, \sigma^2))$$

where $\hat{r}_i(x)$ is any other relevance computation





(Smilkov et al. 2017)

Backpropagation-based methods

- **Gradient methods**

Method	**Computation of $r_i(x)$**
Simple Gradients	$\frac{\partial M(x)}{\partial x_i}$, $\lVert\frac{\partial M(x)}{\partial x_i}\rVert_1$, or $\lVert\frac{\partial M(x)}{\partial x_i}\rVert_2$
Gradient×Input	$x_i \odot \frac{\partial M(x)}{\partial x_i}$
Integrated Gradients	$(x_i - \overline{x}_i) \odot \int_{\alpha=0}^{1} \frac{\partial M(\overline{x}+\alpha(x-\overline{x}))}{\partial x_i} d\alpha$ approximated by $(x_i - \overline{x}_i) \odot \sum_{\alpha=0}^{1} \frac{\partial M(\overline{x}+\alpha(x-\overline{x}))}{\partial x_i}$
SmoothGrad	$\frac{1}{m} \sum_{1}^{m} \hat{r}_i(x)(x + \mathcal{N}(0, \sigma^2))$ where $\hat{r}_i(x)$ is any other relevance computation

Table 2: Summary of Gradient methods in terms of how they compute $r_i(x)$.

Backpropagation-based methods

- **Gradient methods: in NLP**

"This is said to be the best movie of the year, but I was almost asleep."

Prediction: Positive (prob = 0.52 🤔)

Simple Gradients

This is said to **be** the best movie of the year , but **I** was almost asleep **.**

be: 0.100, I: 0.175, .: 0.108

Integrated Gradients

This is **said** to be the **best** movie of the year , but I was almost **asleep** .

said: 0.093, best: 0.271, asleep: 0.146

SmoothGrad

This is **said** to be the **best** movie of the year , but I was almost **asleep** .

said: 0.133, best: 0.094, asleep: 0.205

Figure 9: A visualization of different gradient methods on a sentiment classification example predicted as Positive by a GLoVe-LSTM model (generated with AllenNLP Interpret[5]). Darker shades indicate higher relevance for the prediction.

[5] Gradient×Input isn't available in the toolkit.

Backpropagation-based methods

- **Advantages**
  - (a) Relatively easy to compute
  - (b) In terms of Faithfulness, gradients (and variants) are intrinsically tied to the influence of input features on the prediction
    Empirically, certain above-mentioned methods are shown to be more faithful than existing baselines via perturbation-based evaluation
  - (c) Takes the entire computation path into account, as opposed to a snapshot

Backpropagation-based methods

- **Disadvantages**
  - (a) Mostly target low-level features, e.g., pixels / input tokens
  - (b) Not obvious how to apply to non-classification tasks
  - (c) The explanation can be unstable, i.e., minimally different inputs can lead to drastically different relevance maps (Ghorbani et al. 2019; Feng et al. 2018)
  - (d) In terms of Faithfulness, many methods still do not report empirical evaluation results. Actually, there is negative evidence:
    - Certain methods are shown to be only doing input recovery, ignorant of the model's behavior (Nie, Zhang, and Patel 2018)
    - See more in §2.4.4

Five Categories

- Similarity methods
- Analysis of model-internal structures
- Backpropagation-based methods
- **Counterfactual intervention**
- Self-explanatory models

Counterfactual intervention

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Counterfactual intervention	post-hoc	black-box, white-box	local, global	features, examples,	importance scores

- *Counterfactual* reasoning (from social science)

Given two occurring events A and B, A is said to **cause** B if, under some hypothetical **counterfactual** case that A did not occur, B would not have occurred.

in machine learning: example/ feature/ neuron … model output

(Roese and Olson 1995; Winship and Morgan 1999; Lipton 2017)

Counterfactual intervention

- Running example

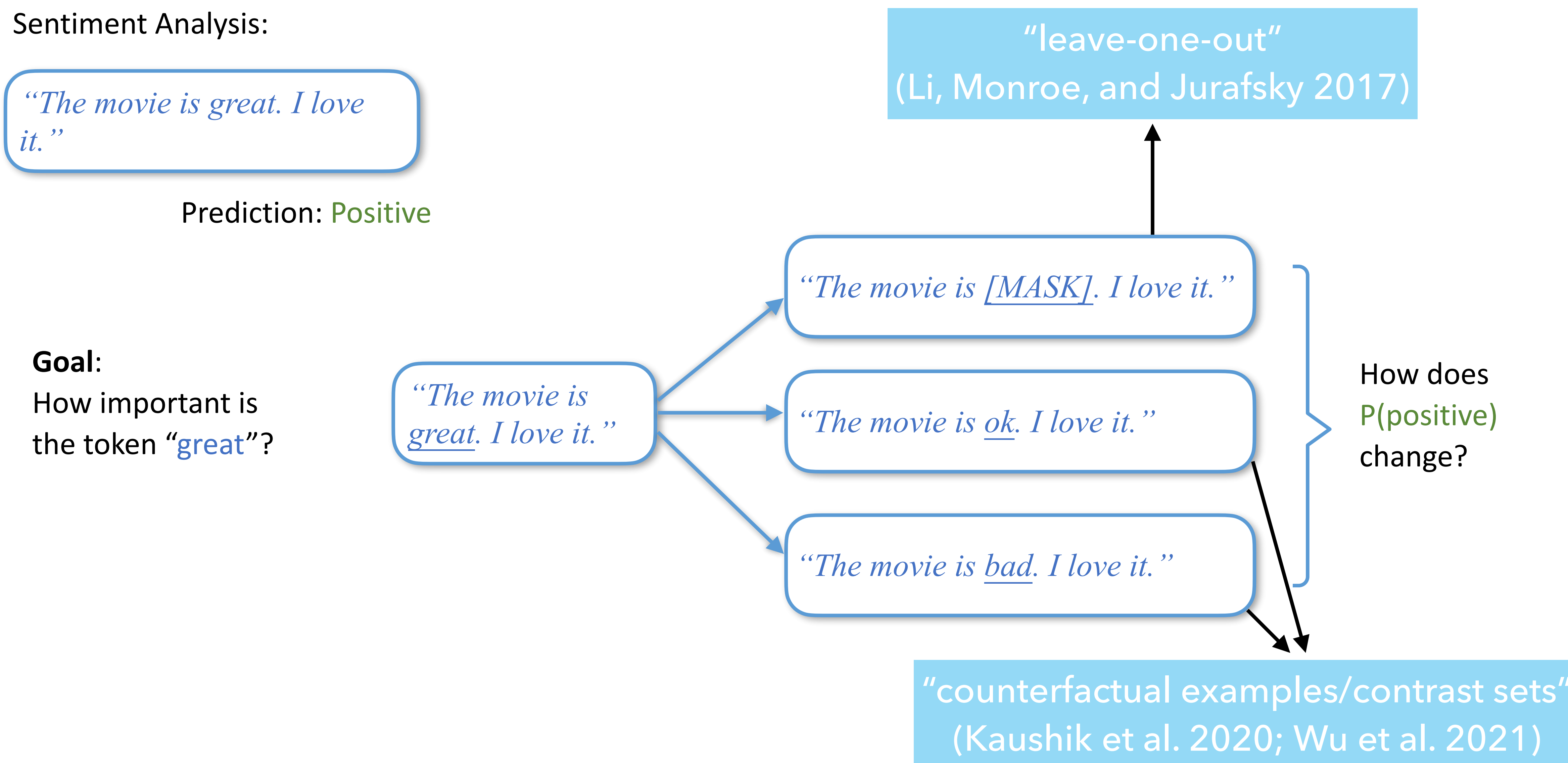
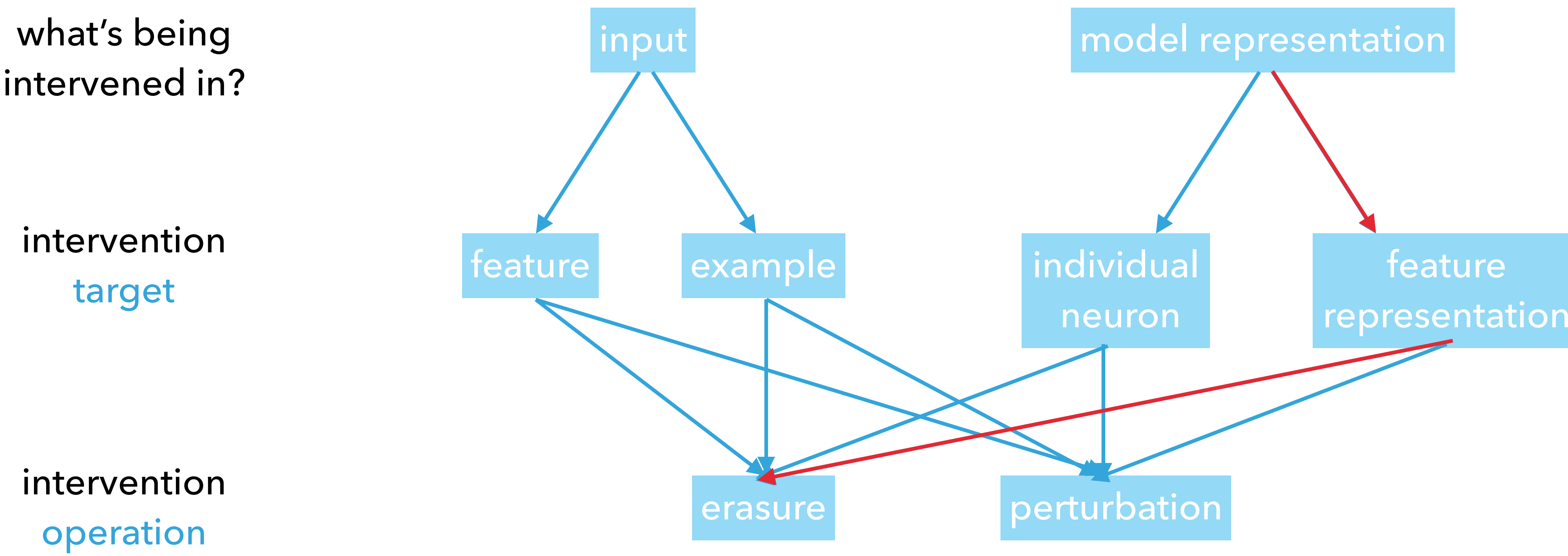


Figure X: Visualization of simple counterfactual intervention methods on the running example.

Counterfactual intervention

- Past work



- Each path is a different type of counterfactual intervention.
- We'll elaborate on one path here: feature-representation-targetd erasure
  (See more in §2.5.2)

Counterfactual intervention

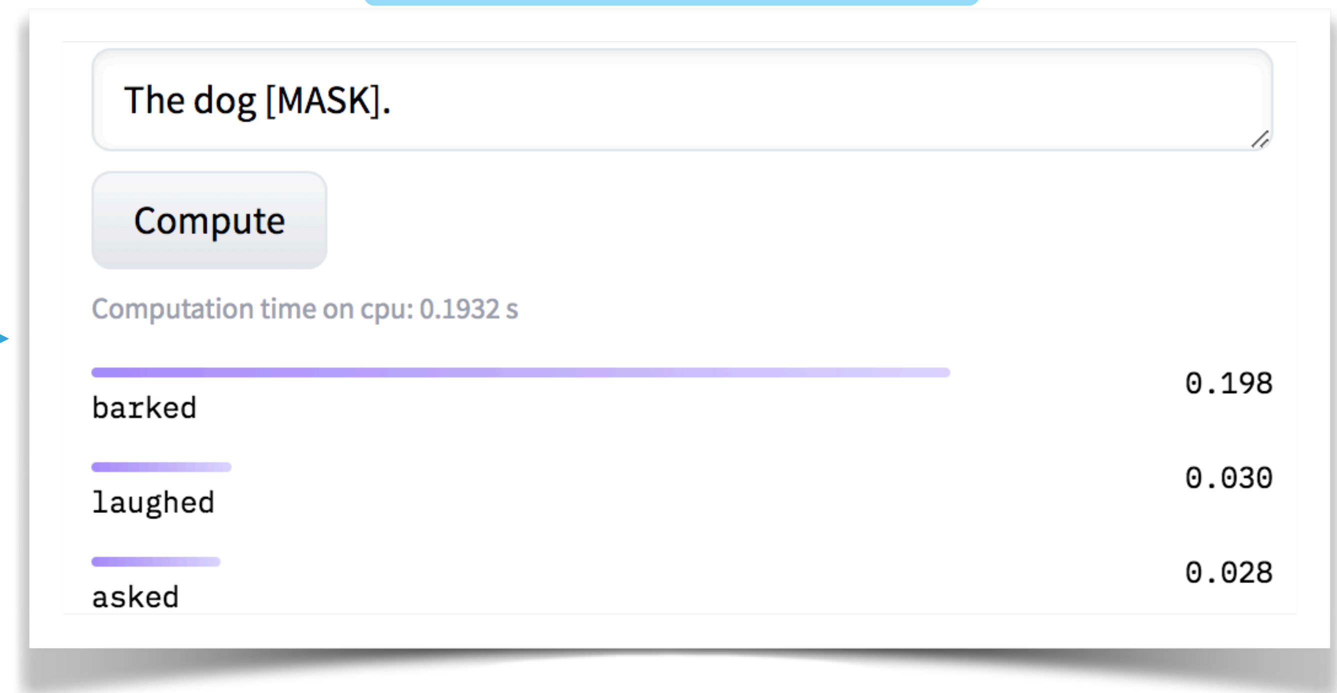
- Feature-representation-targetd erasure
  - **Goal**: Is some feature **used** by the model in some task?

e.g. part-of-speech (POS)

e.g. word prediction

"[MASK]" should be a VERB

used?



  - **Intuition**: If we erase the POS feature from the model representation, how would the word prediction performance change?

Counterfactual intervention

- Amnesic Probing (Elazar et al. 2021)

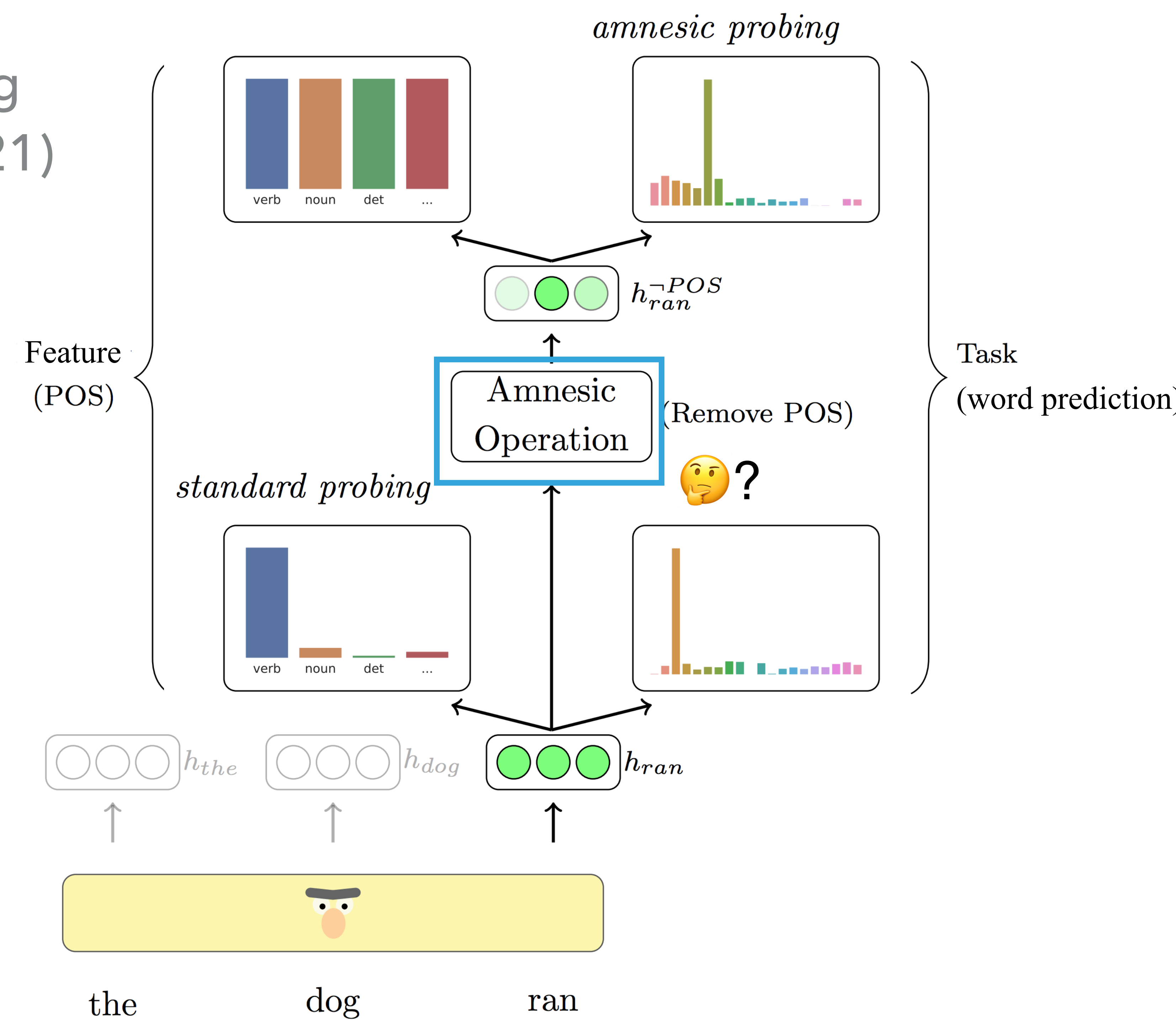
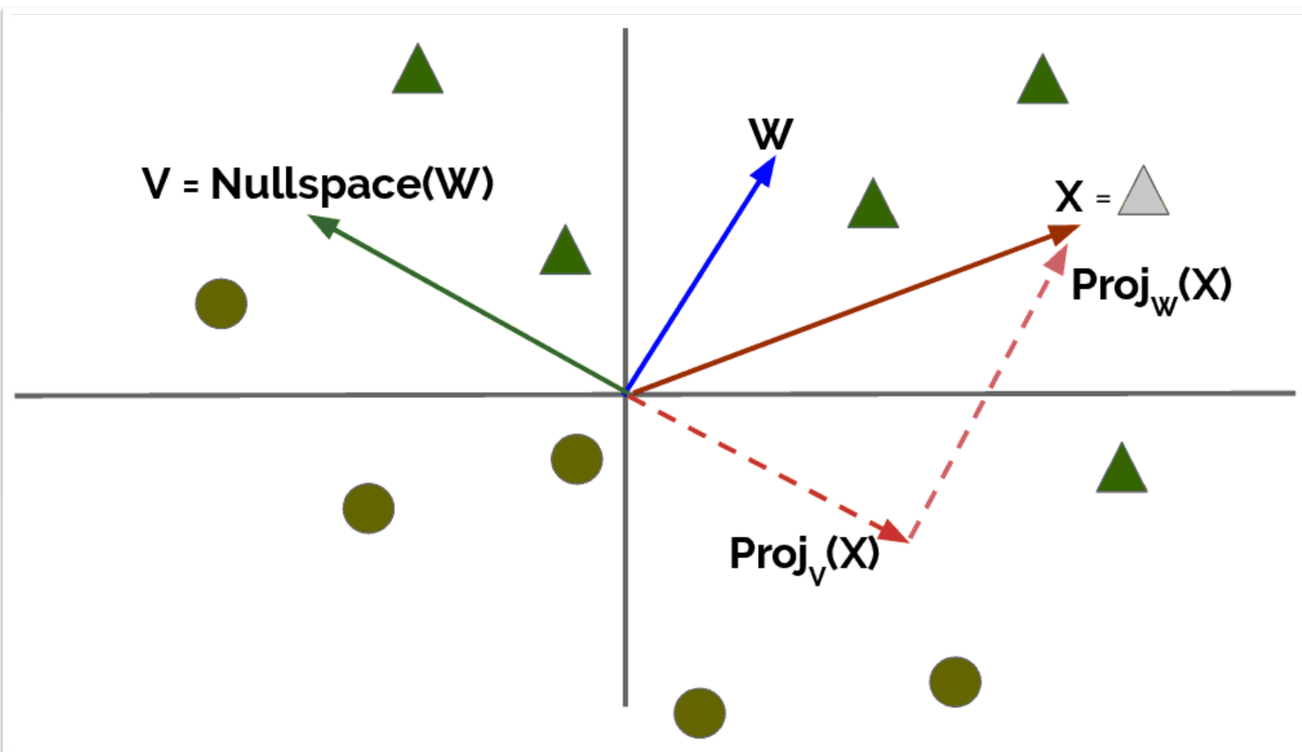


Figure 10: Visualization of Amnesic Probing (figure from Elazar et al. 2021).

Counterfactual intervention

Amnesic Operation : Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020)



Suppose ▲: VERB ●: NOUN

x: an input word representation
W: a linear classifier

Goal: remove the ▲ / ● feature from the model representation x

Method:

1. Train a linear classifier W to predict the target feature.
2. Project x onto V, the nullspace of W.

W has **no effect** on the projected space now! i.e. We've **removed** the target feature **linearly encoded by W**.

3. Repeat 1-2 until there's no such W with above random performance.

→ We've removed the target feature linearly

Counterfactual intervention

- Amnesic Probing (Elazar et al. 2021):
  - **Findings**:
    - *POS*, *dependency tree*, and *named entity* are used in word prediction!
    - But *constituent boundary* seems not.
  - **Faithfulness**:
    - Faithful by construction?
    - Sanity check:
      - Have we removed only the target feature? ✅ most of the time

Counterfactual intervention

- **Advantages**
  - (a) Rooted in the causality literature, and is designed to capture causal instead of mere correlational effects between inputs and outputs
  - (b) Compared to other methods, counterfactual intervention methods are more often explicitly evaluated in terms of Faithfulness
  - (c) Several methods capture the contribution of high-level features beyond input tokens

Counterfactual intervention

- **Disadvantages**
  - (a) Erasure-based intervention can result in nonsensical inputs
  - (b) Intervening in a single feature relies on the assumption that features are independent
    - e.g. "*This movie is <u>mediocre</u>, maybe even <u>bad</u>*"
  - (c) Interventions are often overly specific to the particular example
  - (d) Counterfactual intervention may suffer from hindsight bias

See more in §2.5.4

Five Categories

- Similarity methods
- Analysis of model-internal structures
- Backpropagation-based methods
- Counterfactual intervention
- **Self-explanatory models**

Self-explanatory models

Method	Time	Model accessibility	Scope	Unit of explanation	Form of explanation
Self-explanatory models	built-in	white-box	local, global	features, examples, concepts	importance scores, natural language, causal graphs

- Explaining existing models might be unfaithful …
  What if we just train a model that can explain itself?
- Self-explanatory models output the end task prediciton along with the explanation
- We can supervise the end task and the explanation

Self-explanatory models

- Running example

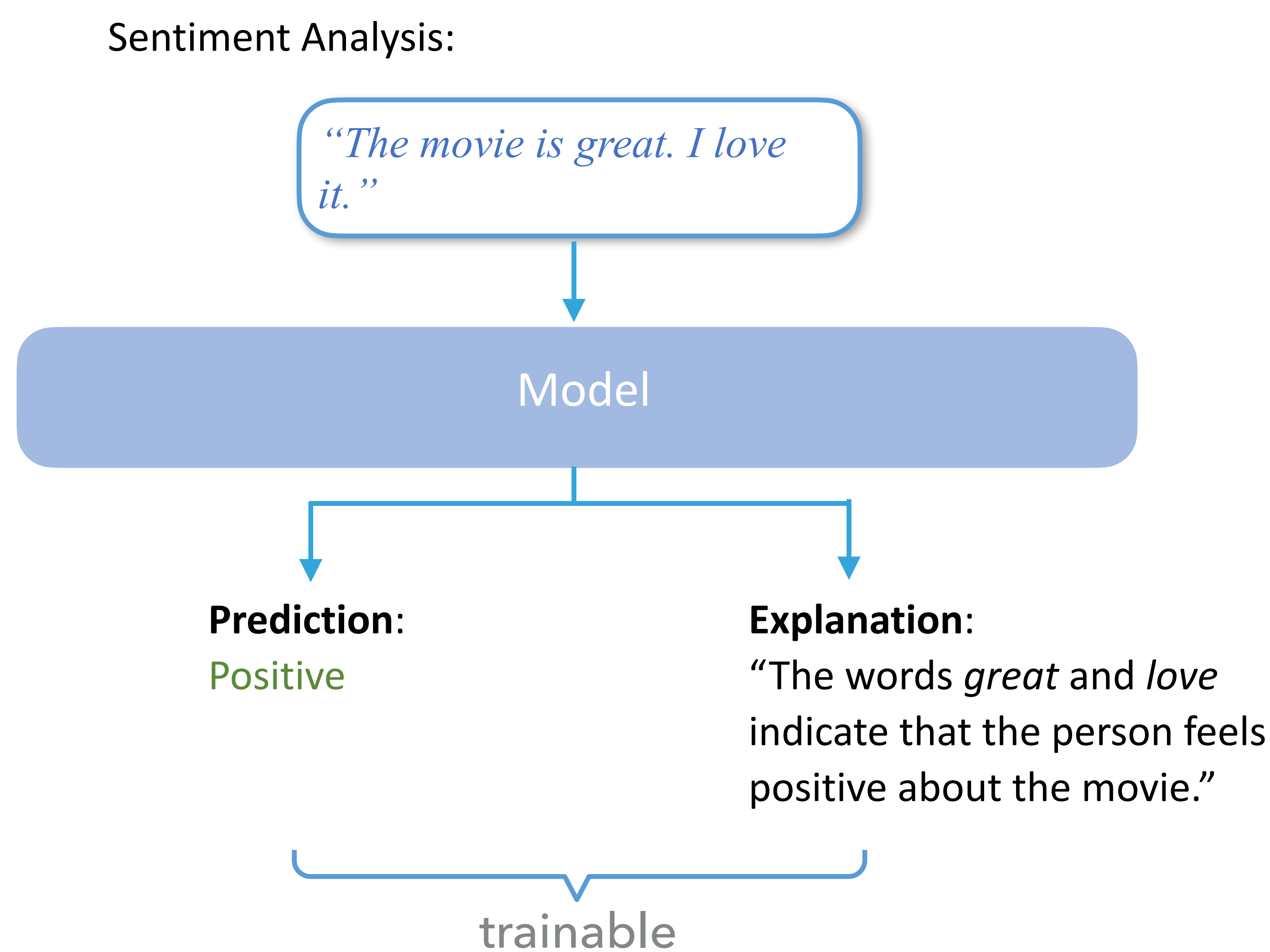
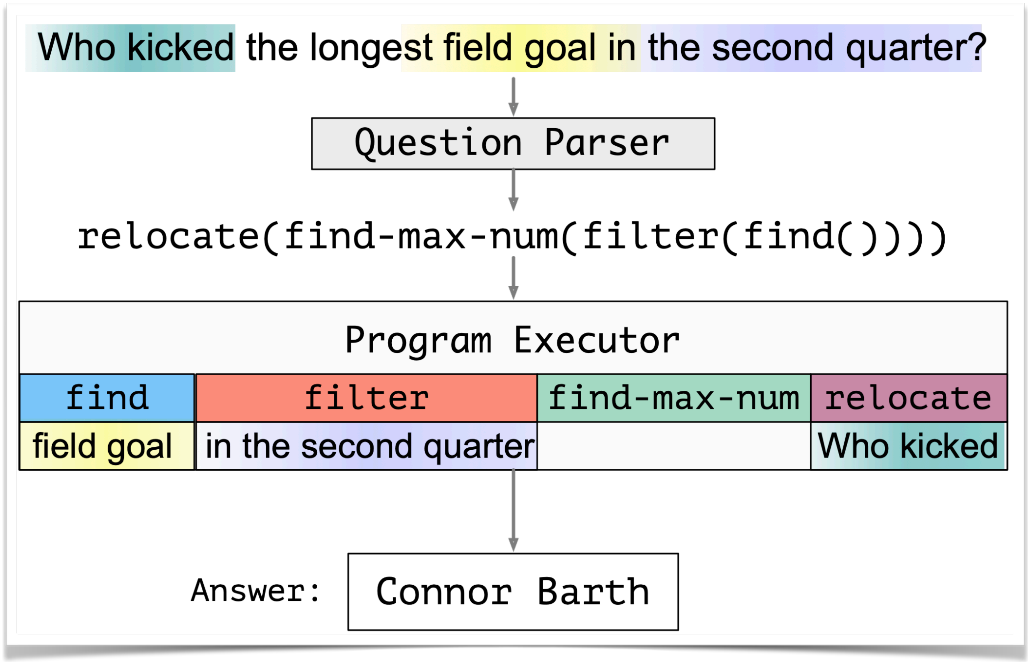


Figure 11: A schematic visualization of self-explanatory models the running example.

Self-explanatory models

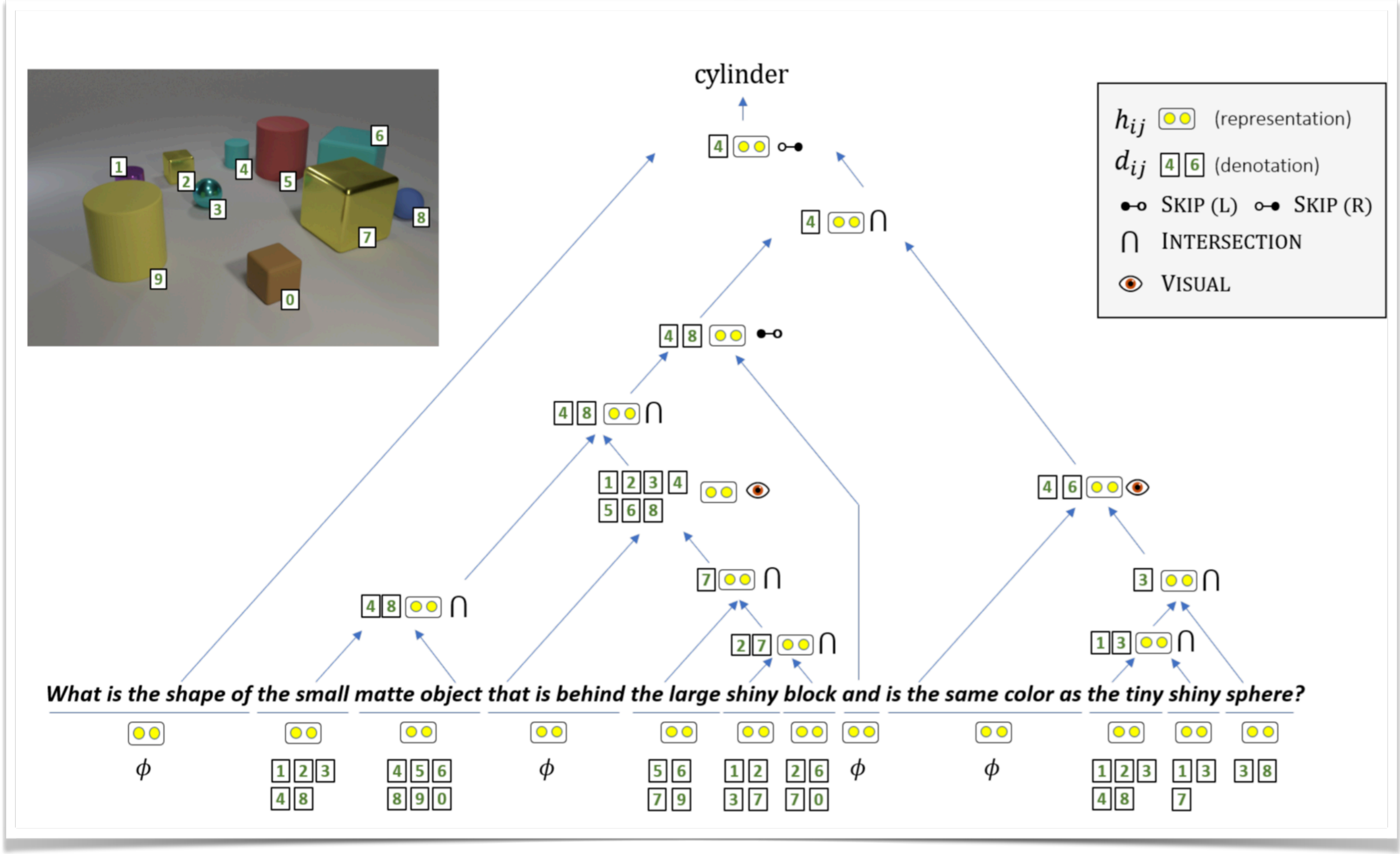
- Past work
  - Explainable architecture
    - Neural Module Networks
    - Neural-Symbolic Models
    - Models with constraints
  - Generating explanations
    - predict-then-explain
    - explain-then-predict
    - jointly-predict-and-explain



(Gupta et al. 2019)

Self-explanatory models

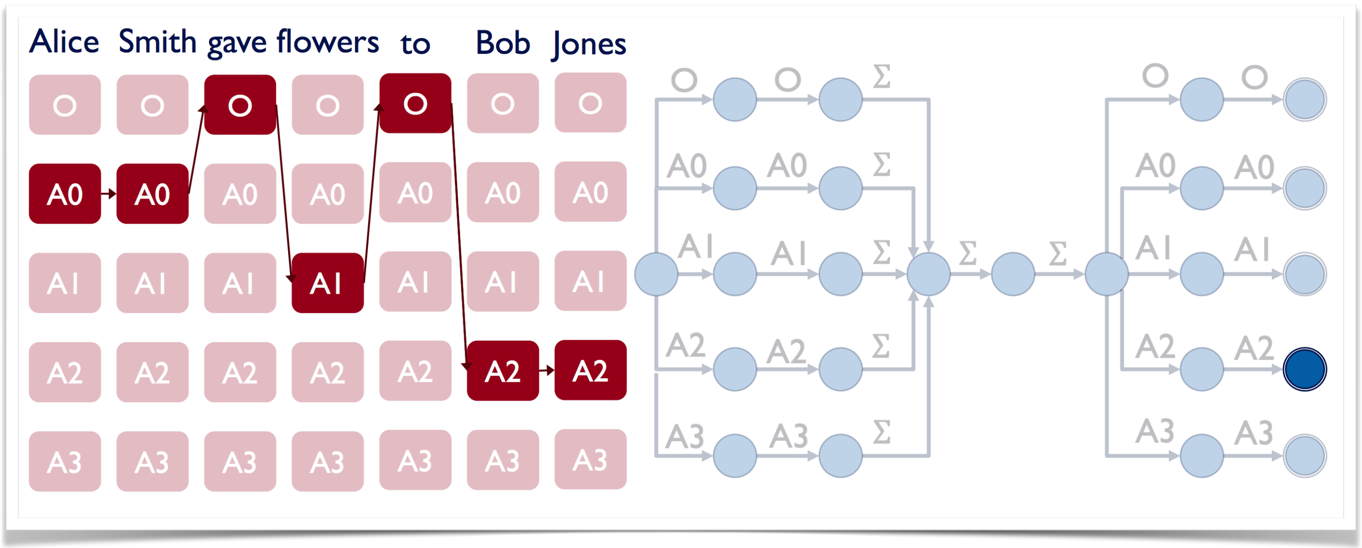
- Past work
  - Explainable architecture
    - Neural Module Networks
    - Neural-Symbolic Models
    - Models with constraints
  - Generating explanations
    - predict-then-explain
    - explain-then-predict
    - jointly-predict-and-explain



(Bogin et al. 2021)

Self-explanatory models

- Past work
  - Explainable architecture
    - Neural Module Networks
    - Neural-Symbolic Models
    - Models with constraints
  - Generating explanations
    - predict-then-explain
    - explain-then-predict
    - jointly-predict-and-explain



(Deutsch et al. 2019)

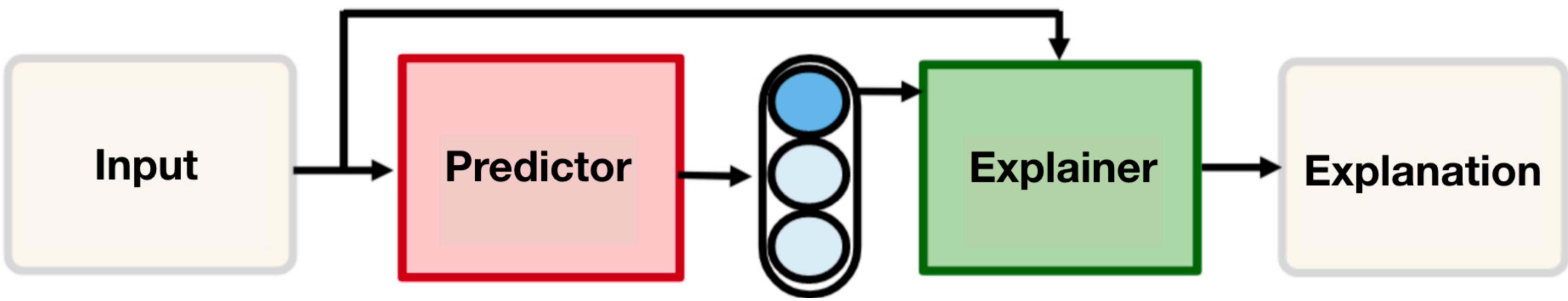
Self-explanatory models

- Past work
  - Explainable architecture
    - Neural Module Networks
    - Neural-Symbolic Models
    - Models with constraints
  - Generating explanations
    - predict-then-explain
    - explain-then-predict
    - jointly-predict-and-explain

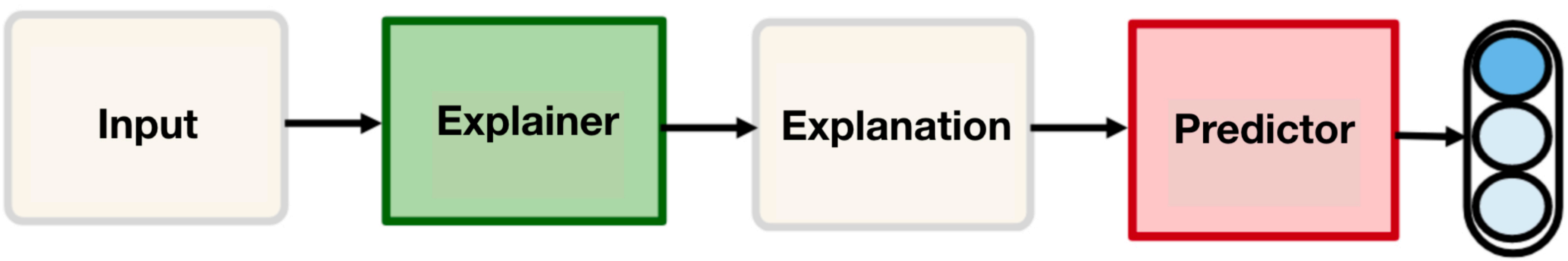
See more in §2.6.2!

Self-explanatory models

- predict-then-explain:



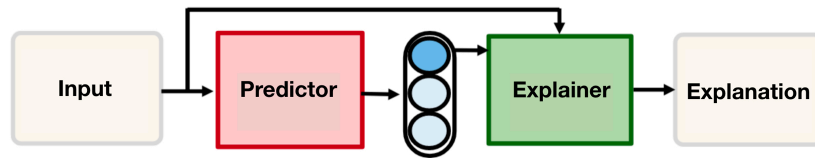
- explain-then-predict:



(figures adapted from Kumar and Talukdar 2020)

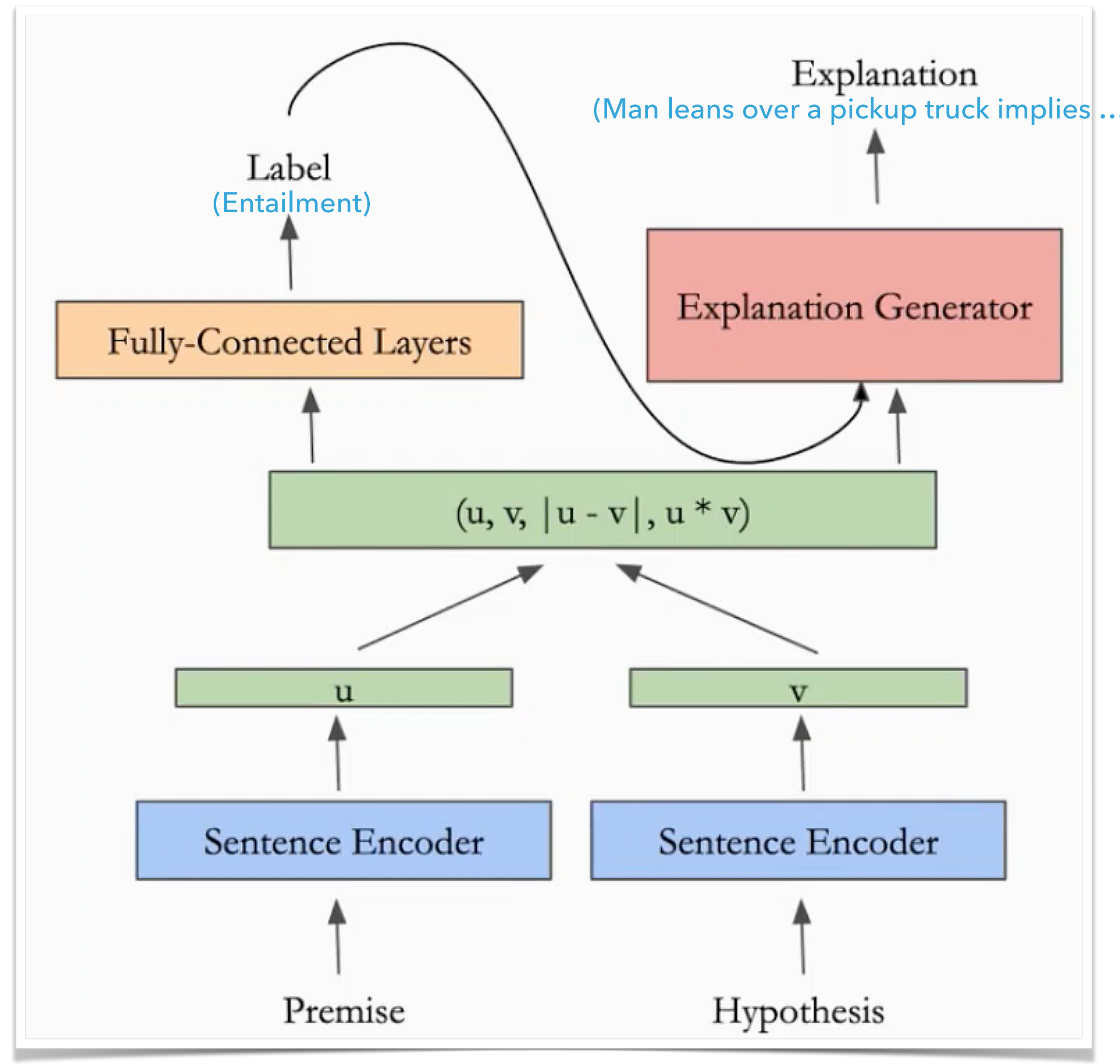
Self-explanatory models

- predict-then-explain:



  - (Camburu et al. 2018)
    - **Task**: Natural Language Inference (NLI)
    - **Data**: e-SNLI
      Stanford Natural Language Inference dataset (SNLI) with human-provided explanations
      - Example:

Premise: A man in an orange vest leans on a pickup truck.
Hypothesis: A man is touching a truck.
Label: Entailment
Explanation: Man leans on a pickup truck implies that he is touching it.

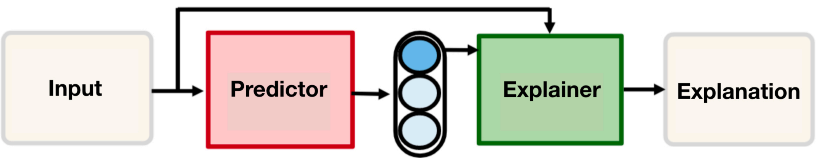


(figure from Oana-Maria Camburu's talk)

    - Train the predictor + the explainer

Self-explanatory models

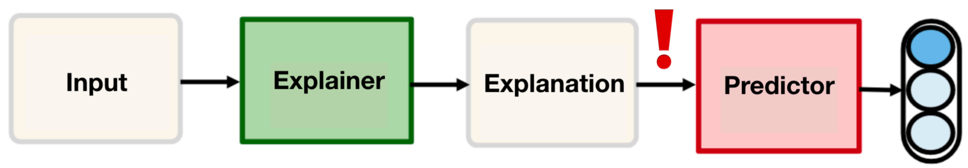
- predict-then-explain:



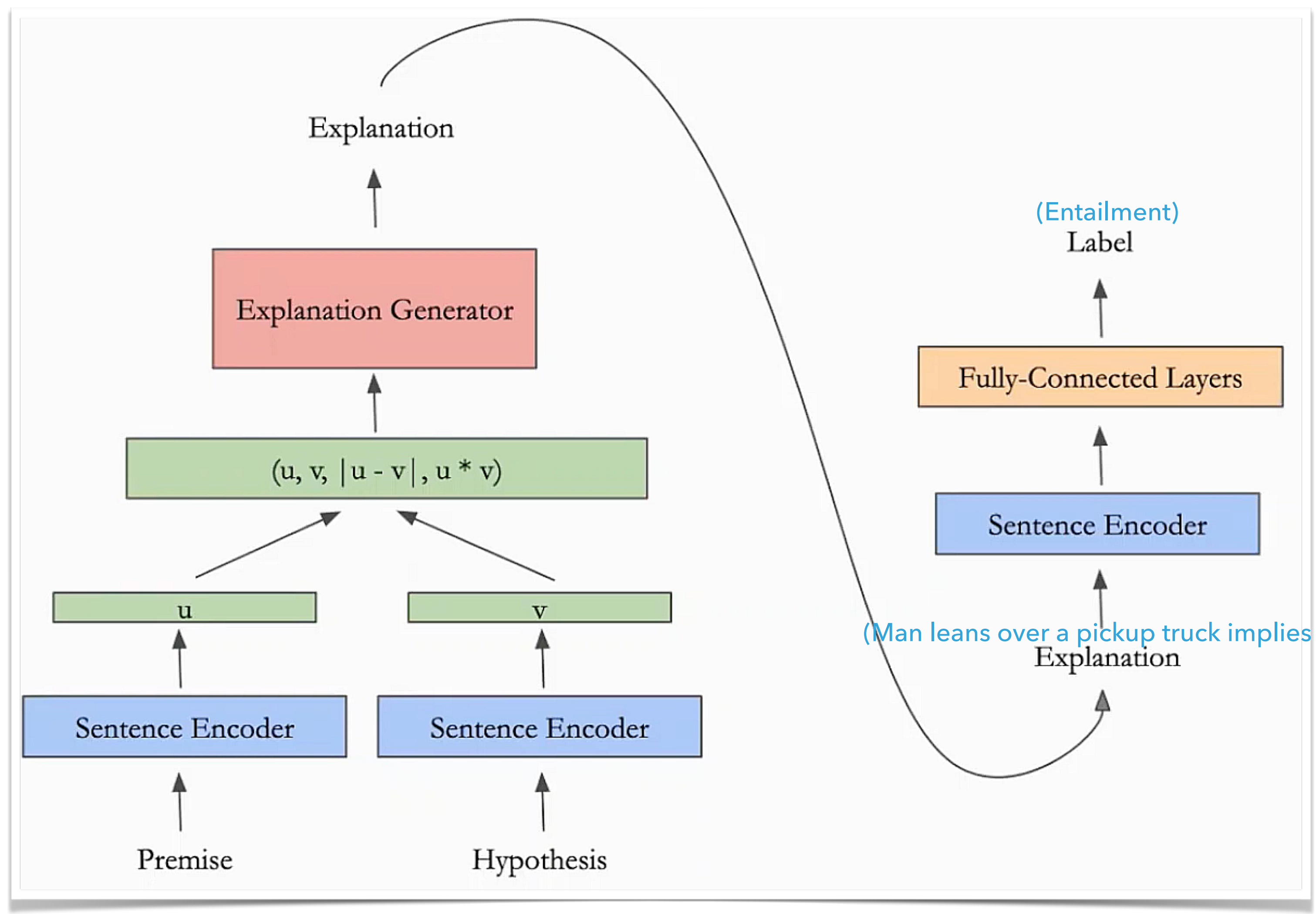
  - Problems:
    - Is the Explainer faithful 🤔?
      - The Predictor doesn’t depend on the Explainer.
        The Explainer suffers from the same Faithfulness challenge as previous post-hoc methods …

Self-explanatory models

- explain-then-predict:



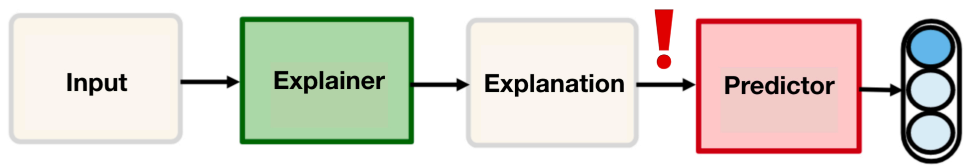
  - Predictor can only access the explanation, but not the input
    - why?
  - Still (Camburu et al. 2018): compared to predict-then-explain, slightly worse label accuracy, but better explanation plausibility



(figure from Oana-Maria Camburu's talk)

Self-explanatory models

- explain-then-predict:



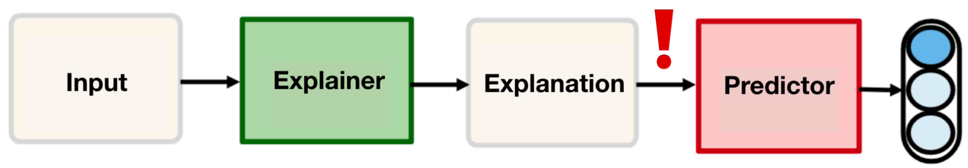
  - faithful by construction?
    - But the explanation may contain spurious cues to the label … e.g.

"X is a type of Y"
"X implies Y"
"X is the same as Y" … → Entailment

"not all X are Y"
"not every X is Y" … → Neutral

"X is not the same as Y" … → Contradiction

Self-explanatory models

- explain-then-predict: 

  - Fix: let the Explainer generate an explanation for every label?
  - (Kumar and Talukdar 2020): Natural language Inference over Label-specific Explanations (**NILE**)[6]

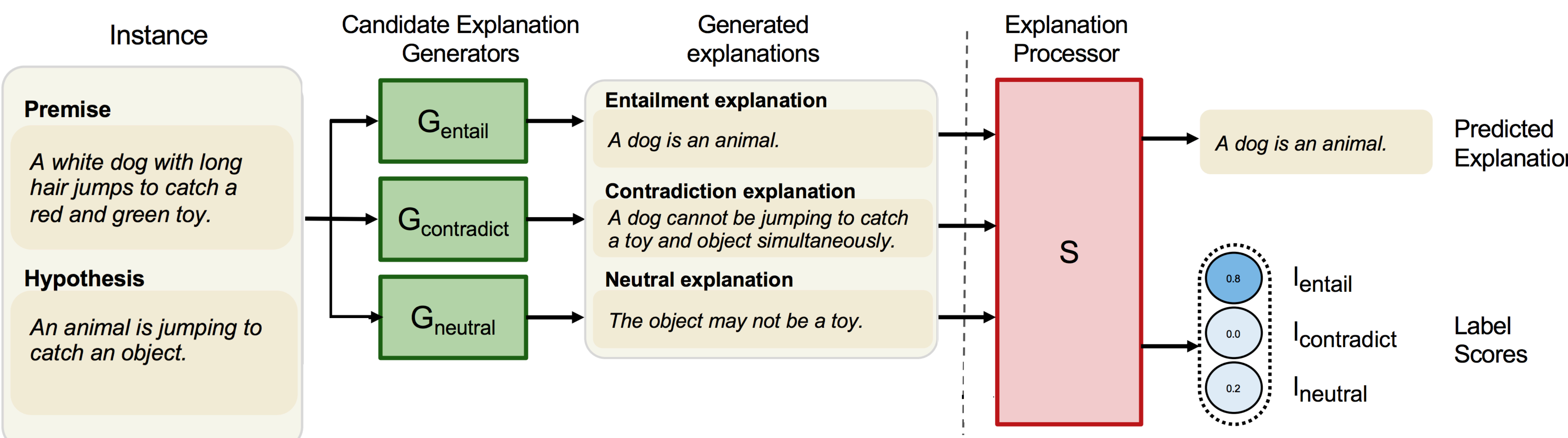


figure from (Kumar and Talukdar 2020)

[6] NILE also has a joint-predict-and-explain variant; see §2.6.2 for more details

Self-explanatory models

- **Advantages**
  - (a) No need for post-hoc explanations
  - (b) Flexible form of explanation: model architecture, input features, natural language, causal graphs …
  - (c) Possible to supervise the explainer with human-provided explanations, thus encouraging the model to rely on desired human-like reasoning mechanisms instead of spurious cues
  - (d) Certain self-explanatory models (see §2.6.3 for examples) are faithful by construction (we should be extra cautious about this claim, though)

Self-explanatory models

- **Disadvantages**
  - (a) Still, many self-explanatory models cannot guarantee Faithfulness (see examples in §2.6.4)
  - (b) Interpretability can come at the cost of task performance (Camburu et al. 2018; Subramanian et al. 2020; inter alia)
  - (c) Large-scale human supervision on explanations can be costly and noisy (Dalvi et al. 2021)
  - (d) Hard to automatically evaluate the quality of model-generated explanations given the reference human explanations

Five Categories

- ~~Similarity methods~~
- ~~Analysis of model-internal structures~~
- ~~Backpropagation-based methods~~
- ~~Counterfactual intervention~~
- ~~Self-explanatory models~~

DISCUSSION

Virtues

A. Explainability research is conducive to **bridging the gap between competence and performance** in language models.

(unconscious) **knowledge** of a language ≠ actual **use** of the knowledge

B. There has been i**ncreasing awareness** of **Faithfulness** and other principles of explanation methods.

C. Usually, the **form** of explanation (importance scores, visualization, natural language, or causal graphs) is **intuitive** to understand, even for lay people.

D. There are a plethora of **model-agnostic** explanation methods, especially faor classification tasks.

E. Many studies **draw insights from work in vision** and develop adaptable methods in language.

F. Numerous **toolkits** have been developed to help users apply explanation methods to their own models.[7]

[7] See §2.3, 2.4, and 2.5 for more details.

Challenges and Future Work

A. Many methods still lack **objective quality evaluation**, especially in terms of Faithfulness. (§1.2.4)
→ We need a **universal evaluation framework**, which is fundamental to measuring the progress of any research in thais area.

B. Most methods provide explanations in terms of **surface-level features**, e.g., pixels in vision and tokens in language. (§2.4)
→ Future work should explore how to capture the contribution of **higher-level features** in a task, including **linguistic** (case, gender, part-of-speech, semantic role, syntax dependency, …), and **extra-linguistic** (commonsense and world knowledge, …) ones.

C. Most methods only capture **importance scores of individual features** to the prediction. (§2.3, 2.4)
→ Future work can focus on **more flexible forms of explanation**, e.g., feature interactions or causal graphs.

Challenges and Future Work

D. Existing work mostly focuses on **limited task formats**, e.g., classification.
→ Future work can study **alternative task formats** such as language generation and structured prediction, or even better, develop generalizable methods across tasks.

E. It is not always obvious whether insights from explanations are **actionable**. How should the user go about **fixing** a discovered problem (through the data, model architecture, training procedure, hyper- parameters, …)? How should they **communicate** with the model?
→ **Interactive** explanations will be a fruitful area for future study.

F. There has been a tension between model performance and interpretability, especially evident in self-explanatory models.
→ It will be helpful to have a **theoretical understanding** of whether the tension is intrinsic or avoidable.

CONCLUSION

Conclusion

A. This survey provides **an extensive tour of recent advances** in NLP explainability, through the lens of **Faithfulness**.

B. We first discuss the notion of **Faithfulness** – despite being a fundamental principle of model explanation methods, Faithfulness does **not** have a well-established definition or evaluation framework.

C. We present a critical review of **five categories** of existing model explanation methods: similarity methods, analysis of model-internal structures, backpropagation-based methods, counterfactual intervention, and self-explanatory models.

D. We summarize all methods by discussing their common **virtues and challenges** and outline **future research directions**.

E. We hope that this survey provides an overview of the area for **researchers interested in interpretability**, as well as **developers aiming at better understanding their own models**.

THANKS
FOR LISTENING!
QUESTIONS?

References

1. Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, and Katja Hansen. 2010. How to Explain Individual Classification Decisions. Journal of Machine Learning Research, page 29.
2. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
3. Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins,Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58:82–115.
4. Bogin, Ben, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. 2021. Latent Compositional Representations Improve Systematic Generalization in Grounded Question Answering. Transactions of the Association for Computational Linguistics, 9:195–210. Place: Cambridge, MA Publisher: MIT Press.
5. Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc.
6. Caruana, R., H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson. 1999. Case-based explanation of non-case-based learning methods. Proceedings of the AMIA Symposium, pages 212–215.
7. Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Association for Computational Linguistics, Florence, Italy.
8. Dalvi, Bhavana, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining Answers with Entailment Trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7358–7370, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.
9. Denil, Misha, Alban Demiraj, and Nando de Freitas. 2015. Extraction of Salient Sentences from Labelled Documents. arXiv:1412.6815 [cs]. ArXiv: 1412.6815.
10. Deutsch, Daniel, Shyam Upadhyay, and Dan Roth. 2019. A General-Purpose Algorithm for Constrained Sequential Inference. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 482–492, Association for Computational Linguistics, Hong Kong, China.
11. Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics, 9:160–175.
12. Feng, Shi, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3719–3728, Association for Computational Linguistics, Brussels, Belgium.
13. Ghorbani, Amirata, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, pages 3681–3688, AAAI Press, Honolulu, Hawaii, USA.
14. Gupta, Nitish, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2019. Neural Module Networks for Reasoning over Text.

References

15. Harrington, L. A., M. D. Morley, A. Šcedrov, and S. G. Simpson. 1985. Harvey Friedman's Research on the Foundations of Mathematics. Elsevier. Google-Books-ID: 2plPRR4LDxIC.
16. Herman, Bernease. 2019. The Promise and Peril of Human Evaluation for Model Interpretability. arXiv:1711.07414 [cs, stat]. ArXiv: 1711.07414.
17. Jacovi, Alon and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
18. Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive Explanations for Model Interpretability. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1597–1611, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.
19. Jain, Sarthak and Byron C. Wallace. 2019. Attention is not Explanation. arXiv:1902.10186 [cs]. ArXiv: 1902.10186.
20. Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and Understanding Recurrent Networks. arXiv:1506.02078 [cs]. ArXiv: 1506.02078.
21. Kaushik, Divyansh, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. arXiv:1909.12434 [cs, stat]. ArXiv: 1909.12434.
22. Kumar, Sawan and Partha Talukdar. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
23. Lei, Tao. 2017. Interpretable neural models for natural language processing. Thesis, Massachusetts Institute of Technology. Accepted: 2017-05-11T19:59:27Z.
24. Li, Jiwei, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 681–691, Association for Computational Linguistics, San Diego, California.
25. Li, Jiwei, Will Monroe, and Dan Jurafsky. 2017. Understanding Neural Networks through Representation Erasure. arXiv:1612.08220 [cs]. ArXiv: 1612.08220.
26. Lipton, Zachary C. 2017. The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat]. ArXiv: 1606.03490.
27. Martins, Andre and Ramon Astudillo. 2016. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In Proceedings of The 33rd International Conference on Machine Learning, pages 1614–1623, PMLR. ISSN: 1938-7228.
28. Mullenbach, James, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1101–1111, Association for Computational Linguistics, New Orleans, Louisiana.
29. Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44):22071–22080. Publisher: Proceedings of the National Academy of Sciences.
30. Narang, Sharan, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. arXiv:2004.14546 [cs]. ArXiv: 2004.14546.
31. Nie, Weili, Yang Zhang, and Ankit Patel. 2018. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In Proceedings of the 35th International Conference on Machine Learning, pages 3809–3818, PMLR. ISSN: 2640-3498.

References

32. Pruthi, Danish, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to Deceive with Attention-Based Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4782–4793, Association for Computational Linguistics, Online.
33. Rajagopal, Dheeraj, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 836–850, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.
34. Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7237–7256, Association for Computational Linguistics, Online.
35. Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3363–3377, Association for Computational Linguistics, Online.
36. Ribeiro, Marco Tulio, 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, KDD '16, pages 1135–1144, Association for Computing Machinery, New York, NY, USA.
37. Roese, Neal J. and James M. Olson. 1995. Counterfactual thinking: A critical overview. In What might have been: The social psychology of counterfactual thinking. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pages 1–55.
38. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In In Workshop at International Conference on Learning Representations.
39. Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825 [cs, stat]. ArXiv: 1706.03825.
40. Strobelt, Hendrik, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. 2018. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. IEEE Transactions on Visualization and Computer Graphics, 24(1):667–676. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
41. Subramanian, Sanjay, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining Faithful Interpretations from Compositional Neural Networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5594–5608, Association for Computational Linguistics, Online.
42. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 3319–3328, JMLR.org, Sydney, NSW, Australia.
43. Wallace, Eric, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting Neural Networks with Nearest Neighbors. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 136–144, Association for Computational Linguistics, Brussels, Belgium.
44. Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc.
45. Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Association for Computational Linguistics, Brussels, Belgium

References

46. Wiegreffe, Sarah and Yuval Pinter. 2019. Attention is not not Explanation. arXiv:1908.04626 [cs]. ArXiv: 1908.04626.
47. Winship, Christopher and Stephen L. Morgan. 1999. The Estimation of Causal Effects from Observational Data. Annual Review of Sociology, 25(1):659–706. _eprint: https://doi.org/10.1146/annurev.soc.25.1.659.
48. Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6707–6723, Association for Computational Linguistics, Online.
49. Xie, Qizhe, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An Interpretable Knowledge Transfer Model for Knowledge Base Completion. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 950–962, Association for Computational Linguistics, Vancouver, Canada.
50. Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning, pages 2048–2057, PMLR. ISSN: 1938-7228.