Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg
Bar Ilan University and Allen Institute for AI
yoav.goldberg@gmail.com

Introduction

- Challenge: the interpretation of neural models and their decisions.
- How did a model reach its decision?

Introduction

- Challenge: the interpretation of neural models and their decisions.
- How did a model reach its decision?
- A lot of research activity and interest, **but the foundations are still shaky.**

Introduction

- Challenge: the interpretation of neural models and their decisions.
- How did a model reach its decision?
- A lot of research activity and interest, **but the foundations are still shaky.**
- Core questions of the field, currently:
  - **What counts as an explanation** of a model's decision?
  - **How do we evaluate** the quality of an explanation?

Introduction

- Challenge: the interpretation of neural models and their decisions.
- How did a model reach its decision?
- A lot of research activity and interest, **but the foundations are still shaky.**
- Core questions of the field, currently:
  - **What counts as an explanation** of a model's decision?
  - **How do we evaluate** the quality of an explanation?
- **This work:**

  Trying to make sense of where things stand w.r.t **faithful explanations.**

Introduction

We cover:

1. **Guidelines:** Pitfalls to avoid when evaluating models for faithfulness

2. **Survey**: Three assumptions underlying current literature on faithful explanations

3. **Opinion:** Is faithful interpretation doomed to fail? And what should we do about it?

Background

What makes an interpretation useful?

Literature* formalizes multiple attributes we may want in an interpretation:

* references in the paper.

What makes an interpretation useful?

Literature* formalizes multiple attributes we may want in an interpretation:

Readability

Is the explanation intuitive and easy to understand?

* references in the paper.

What makes an interpretation useful?

Literature* formalizes multiple attributes we may want in an interpretation:

Readability

Is the explanation intuitive and easy to understand?

Plausibility

Is it *convincing* as an explanation to the interpreted process?

* references in the paper.

What makes an interpretation useful?

Literature* formalizes multiple attributes we may want in an interpretation:

Readability

Is the explanation intuitive and easy to understand?

Plausibility

Is it *convincing* as an explanation to the interpreted process?

Faithfulness

Does the explanation accurately describes the true reasoning process of the model?

* references in the paper.

What makes an interpretation useful?

Literature formalizes multiple attributes we may want in an interpretation:

Readability

Is the explanation intuitive and easy to understand?

Plausibility

Is it *convincing* as an explanation to the interpreted process?

Faithfulness

Does the explanation accurately describes the true reasoning process of the model?

There is a **trade-off** between them:

e.g., **raw activations** of a neural network are **faithful**, but **not-readable**.

What makes an interpretation useful?

We focus on the faithfulness property

- **Faithfulness is key to useful interpretations.**
- **Faithfulness is challenging to achieve:** how does one remain faithful while being a readable and simplified version?

Faithfulness

Does the explanation accurately describes the true reasoning process of the model?

Pitfalls to avoid when evaluating for faithfulness

Guidelines on evaluating faithfulness

- We propose multiple guidelines on how to **not** evaluate faithfulness.
- These are inspired by common pitfalls we observed in the literature.

Guidelines on evaluating faithfulness

- We propose multiple guidelines on how to **not** evaluate faithfulness.
- These are inspired by common pitfalls we observed in the literature.
- We highlight some of them here. More in paper.

Guideline 1:

faithfulness ≠ plausibility

- Many evaluations **conflate** evaluating **faithfulness** and evaluating **plausibility.**
  - This is bad and should be avoided. **Be explicit.**

Guideline 1:

faithfulness ≠ plausibility

- Many evaluations **conflate** evaluating **faithfulness** and evaluating **plausibility.**
  - This is bad and should be avoided. **Be explicit.**
- A *plausible* but *unfaithful* interpretation is **akin to lying,** and can be dangerous.
  - For example, convincing a user that the model decided based on (a very plausible) X while in fact it decided based on Y.

Guideline 2:

model decision process ≠ human decision process

- Many works evaluate faithfulness by comparing model-provided explanations to human-provided explanations, or asking humans to rate explanation quality.

Guideline 2:

model decision process ≠ human decision process

- Many works evaluate faithfulness by comparing model-provided explanations to human-provided explanations, or asking humans to rate explanation quality.
  - This is not evaluating faithfulness. Don't do this.

Guideline 2:

model decision process ≠ human decision process

- Many works evaluate faithfulness by comparing model-provided explanations to human-provided explanations, or asking humans to rate explanation quality.
  - This is not evaluating faithfulness. Don't do this.
- We (humans) cannot understand models that need interpretation. (otherwise, why research this?)

Guideline 2:

model decision process ≠ human decision process

- Many works evaluate faithfulness by comparing model-provided explanations to human-provided explanations, or asking humans to rate explanation quality.
  - This is not evaluating faithfulness. Don't do this.
- We (humans) cannot understand models that need interpretation. (otherwise, why research this?)
  - ⇒ Humans cannot judge if an interpretation is faithful.

Guideline 2:

model decision process ≠ human decision process

- Many works evaluate faithfulness by comparing model-provided explanations to human-provided explanations, or asking humans to rate explanation quality.
  - This is not evaluating faithfulness. Don't do this.
- We (humans) cannot understand models that need interpretation. (otherwise, why research this?)
  - ⇒ Humans cannot judge if an interpretation is faithful.
  - ⇒ Evaluating interpretations using humans input is evaluating plausibility, not faithfulness.

Guideline 3:

claims are just claims until tested

Don't trust untested claims of "inherent interpretability" of models.

A model which is believed to be "inherently interpretable" should be rigorously tested *just the same* as post-hoc methods.

A claim is a claim until proven, even if it seems reasonable.

The three assumptions behind faithfulness

How do we currently define faithfulness?

- **How does the community currently define and evaluate faithfulness?**
- Papers on the subject often propose ad-hoc evaluation methods distinct to each paper, **seemingly unrelated**.

How do we currently define faithfulness?

- **How does the community currently define and evaluate faithfulness?**
- Papers on the subject often propose ad-hoc evaluation methods distinct to each paper, **seemingly unrelated**.
- We uncover **three implicit assumptions** shared among the current methods.

How do we currently define faithfulness?

- **How does the community currently define and evaluate faithfulness?**
- Papers on the subject often propose ad-hoc evaluation methods distinct to each paper, **seemingly unrelated**.
- We uncover **three implicit assumptions** shared among the current methods.
- The paper aligns the available literature along these assumptions.
    - See the paper for references and descriptions.

How do we currently define faithfulness?

- **How does the community currently define and evaluate faithfulness?**
- Papers on the subject often propose ad-hoc evaluation methods distinct to each paper, **seemingly unrelated**.
- We uncover **three implicit assumptions** shared among the current methods.
- The paper aligns the available literature along these assumptions.
    - See the paper for references and descriptions.
- We **do not necessarily endorse** these assumptions.
    - We offer a survey and meta-analysis of the literature today.

Survey: how do we evaluate faithfulness?

***The Model Assumption*:**

Two models make the same predictions

they use the same reasoning process.

⇒ If two models behave similarly, their interpretations should be similar.

⇒ If the interpretation is in itself a model, it should mimic the decisions of the original model.

E.g., decision trees or rule lists.

Survey: how do we evaluate faithfulness?

***The Prediction Assumption*:** On similar inputs,

the model makes similar decisions	iff ⟺	its reasoning is similar.

⇒ On similar inputs/decisions, interpretations should be similar.

Survey: how do we evaluate faithfulness?

***The Linearity Assumption*:** Certain parts of the input can be significant to the decision independently from other parts.

⇒ Heat-maps can be faithful under certain circumstances.

E.g., attention scores, saliency maps.

Is faithful interpretation doomed?

against the all-or-nothing approach

How are the assumptions being used?

Using these assumptions, current work focuses on proving that interpretation methods are not faithful.

Assumptions make it **easy to show** by **counter-example** that an interpretation is **not faithful**:

How are the assumptions being used?

Using these assumptions, current work focuses on proving that interpretation methods are not faithful.

Assumptions make it **easy to show** by **counter-example** that an interpretation is **not faithful**:

- Finding another model that behaves similarly, with different interpretations.

How are the assumptions being used?

Using these assumptions, current work focuses on proving that interpretation methods are not faithful.

Assumptions make it **easy to show** by **counter-example** that an interpretation is **not faithful**:

- Finding another model that behaves similarly, with different interpretations.
- Finding very similar inputs, with very similar predictions, yet very dissimilar interpretations.

How are the assumptions being used?

Using these assumptions, current work focuses on proving that interpretation methods are not faithful.

Assumptions make it **easy to show** by **counter-example** that an interpretation is **not faithful**:

- Finding another model that behaves similarly, with different interpretations.
- Finding very similar inputs, with very similar predictions, yet very dissimilar interpretations.
- And so on.

Position: against a fatalistic approach

Assumptions make it **easy to show** by **counter-example** that an interpretation is **not faithful**.

We argue that this is **unproductive**.

We believe that almost any proposal can be ruled out this way, because a simplification process (interpretation) is always lossy in some aspect.

But is a ***completely*** faithful interpretation truly necessary?

Are our criteria ***too strict***?

Position: a possible way forward

Domain restrictions

It is easy to find counter-examples, but we care about **natural input spaces** and **specific tasks**.

(1) We should care about how the interpretation method behaves **on these inputs / tasks.**

Targeted interpretations

An interpretation method may be unfaithful in general, but work only on **specific examples**.

(2) Can we define testable conditions on input examples that guarantee faithfulness of a method on them?

Challenge

Domain restrictions

Targeted interpretations

Your suggestions?

We challenge the community to define measures of faithfulness that will allow interpretations to be **"faithful enough"** to be useful, along these axes, or others.