Rethinking Attention-Model Explainability through Faithfulness Violation Test

Yibing Liu   Haoliang Li   Yangyang Guo
Chenqi Kong   Jing Li   Shiqi Wang

The 39th International Conference on Machine Learning (ICML 2022)

Motivation

Attention weights can be always non-negative.

$$A = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_h}}) \quad \text{(Vaswani et al., 2017)}$$

Motivation

Attention weights can be always non-negative.

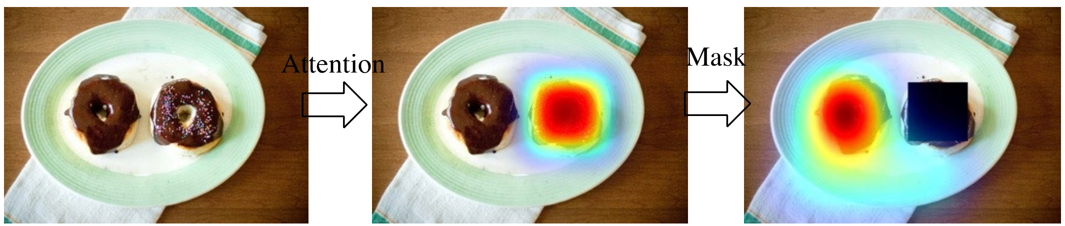
$$A = \text{softmax}(\frac{QK^T}{\sqrt{d_h}}) \quad \text{(Vaswani et al., 2017)}$$

But do **positive** attention weights indicate that features **contribute** to model predictions?

Motivation

Do positive attention weights indicate contribution effects? No!

Question1: What are colorful pieces on the doughnut?



Pred: powder
(Confidence 16%↓) ✓

Question2: What is the girl eating?

Pred: donut
(Confidence 12%↑) ✗

How to evaluate the explanation faithfulness?

Evaluating two properties in explanation weights

– Importance Correlation:

Magnitude <-> Feature Importance

– Polarity Consistency:

Sign <-> Polarity of Feature Impact

How to evaluate the explanation faithfulness?

Evaluating two properties in explanation weights

- Importance Correlation:

  Magnitude <-> Feature Importance

- Polarity Consistency:

  Sign <-> Polarity of Feature Impact

Previous Work

How to evaluate the explanation faithfulness?

Evaluating two properties in explanation weights

– Importance Correlation:

Magnitude <-> Feature Importance

– Polarity Consistency:

Sign <-> Polarity of Feature Impact

Ours Work

Method: Faithfulness Violation Test

Idea: measure the ratio of test samples violating polarity consistency.

Method: Faithfulness Violation Test

Idea: measure the ratio of test samples violating polarity consistency.

Steps: given a test sample x and an explanation method $w(\cdot)$:

Method: Faithfulness Violation Test

Idea: measure the ratio of test samples violating polarity consistency.

Steps: given a test sample x and an explanation method $w(\cdot)$:

1. Find the most influential feature $x^* = \mathrm{argmax}_{x_i \in x} ||w(x_i)||$.

Method: Faithfulness Violation Test

Idea: measure the ratio of test samples violating polarity consistency.

Steps: given a test sample x and an explanation method $w(\cdot)$:

1. Find the most influential feature $x^* = \operatorname{argmax}_{x_i \in x} ||w(x_i)||$.
2. Estimate the feature impact of x^* based on the perturbation test

$$\Delta \mathrm{C}(x, x^*) = f(x)_{\hat{y}} - f(x \backslash x^*)_{\hat{y}}.$$

Method: Faithfulness Violation Test

Idea: measure the ratio of test samples violating polarity consistency.

Steps: given a test sample x and an explanation method $w(\cdot)$:

1. Find the most influential feature $x^* = \operatorname{argmax}_{x_i \in x} ||w(x_i)||$.
2. Estimate the feature impact of x^* based on the perturbation test

$$\Delta \mathrm{C}(x, x^*) = f(x)_{\hat{y}} - f(x \backslash x^*)_{\hat{y}}.$$

3. Check if the explanation weight aligns with the feature impact.

$$\text{Violation} = \mathbb{1}_{\operatorname{sign}(w(x^*) \cdot \Delta \mathrm{C}(x, x^*)) < 0}$$

Experiments

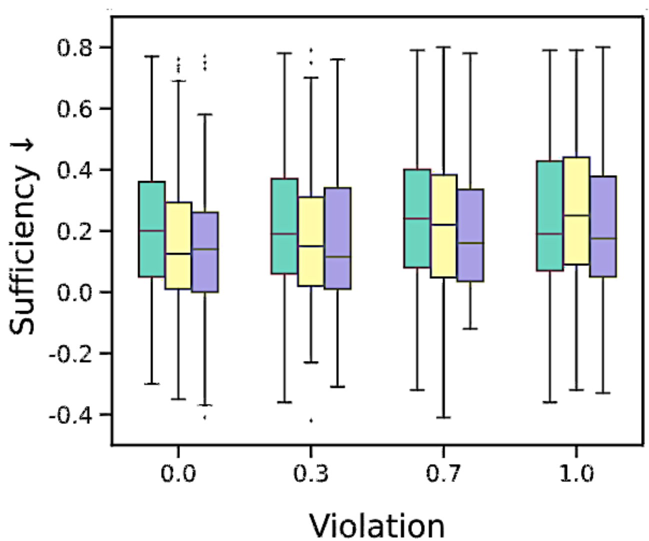
- RQ1: Why we need the faithfulness violation test?
- RQ2: How existing methods perform on faithfulness?
- RQ3: What factors dominate the faithfulness violation issue?

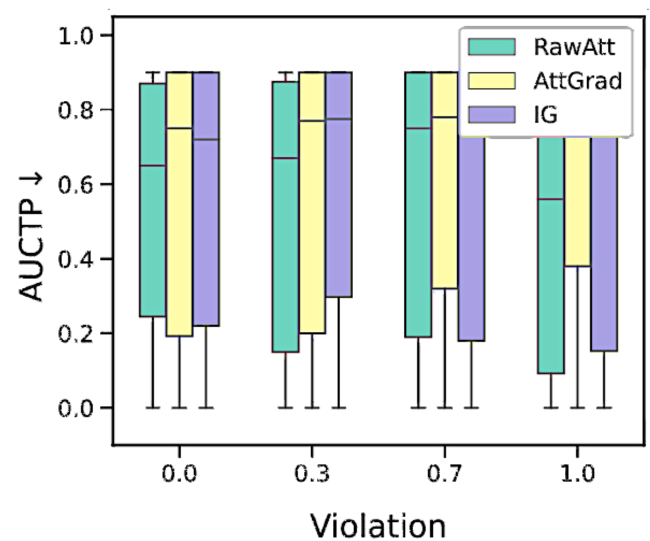
Method	Denoted	Basis
Generic attention-based explanation methods		
Inherent Attention Explanation	RawAtt	α
Attention $\odot$ Gradient	AttGrad	$\alpha \odot \nabla\alpha$
Attention $\odot$ InputNorm	AttIN	$\alpha \odot \|\|v(x)\|\|$
Transformer-based explanation methods		
Partial LRP	PLRP	R^{α}
Attention Rollout	Rollout	α
Transformer Attention Attribution	TransAtt	$\nabla\alpha \odot \mathrm{R}^{\alpha}$
Generic Attention Attribution	GenAtt	$\alpha \odot \nabla\alpha$
Gradient-based attribution methods		
Input $\odot$ Gradient	InputGrad	$x \odot \nabla x$
Integrated Gradients	IG	$x \odot \nabla x$

(More results are provided in our paper)

Comparison with Existing Metrics (RQ1)

Existing metrics are incapable of examining the polarity consistency!

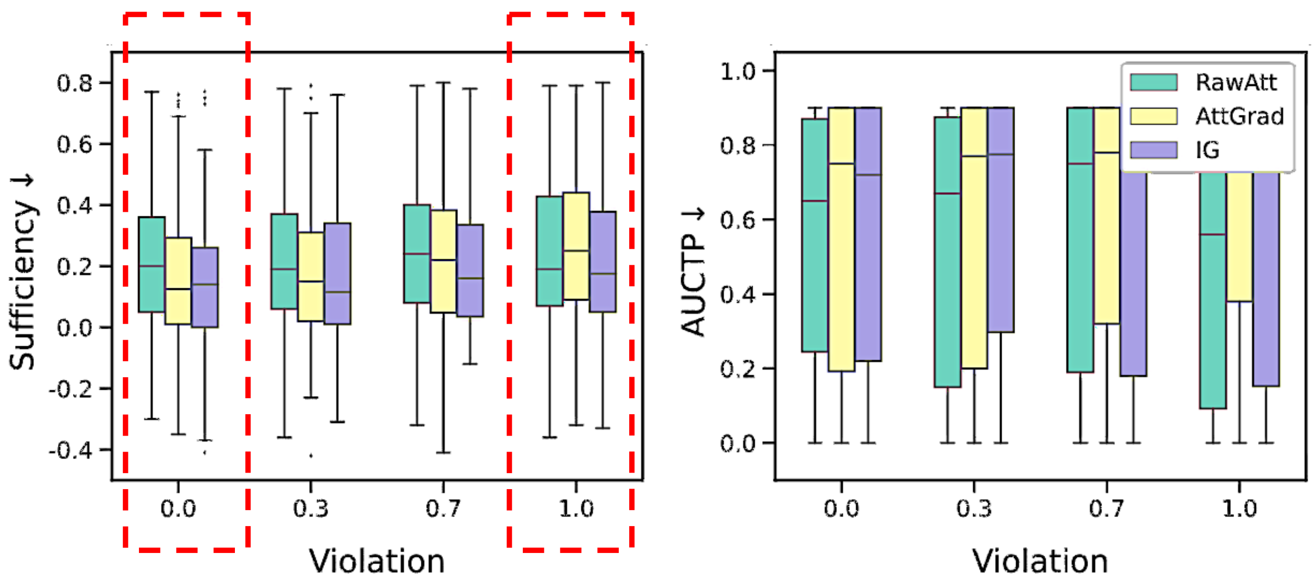




(More results are provided in our paper)

Comparison with Existing Metrics (RQ1)

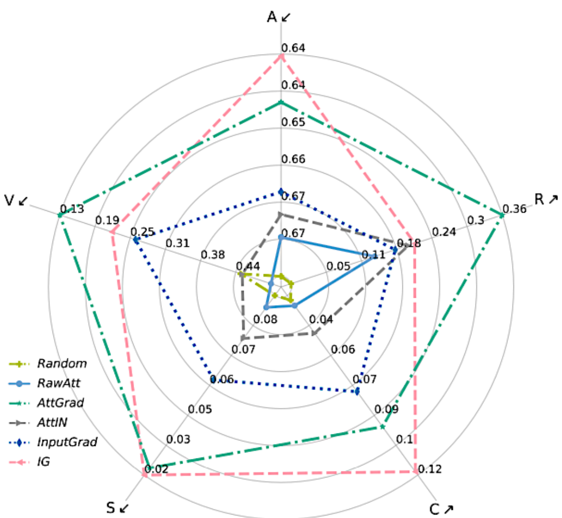
Existing metrics are incapable of examining the polarity consistency!



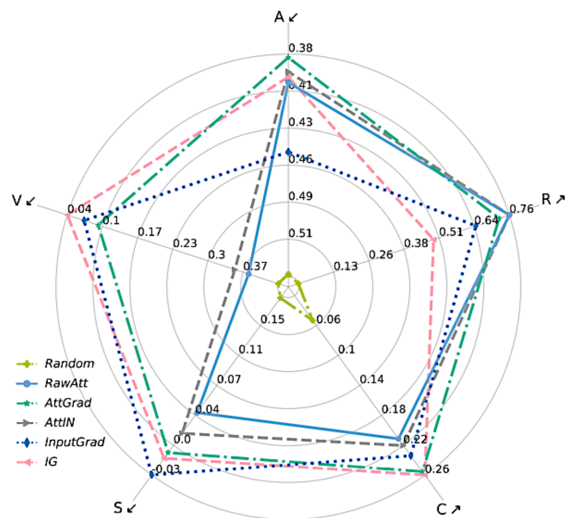
(More results are provided in our paper)

Sanity Faithfulness Evaluation (RQ2)

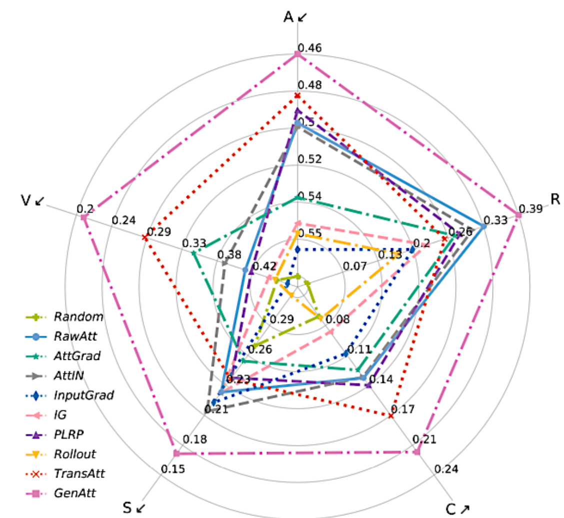
Most tested explanation methods suffer from the faithfulness violation issue regarding polarity consistency.



(a) LSTM+DotAtt on QQP dataset



(b) BUTD on VQA 2.0 dataset

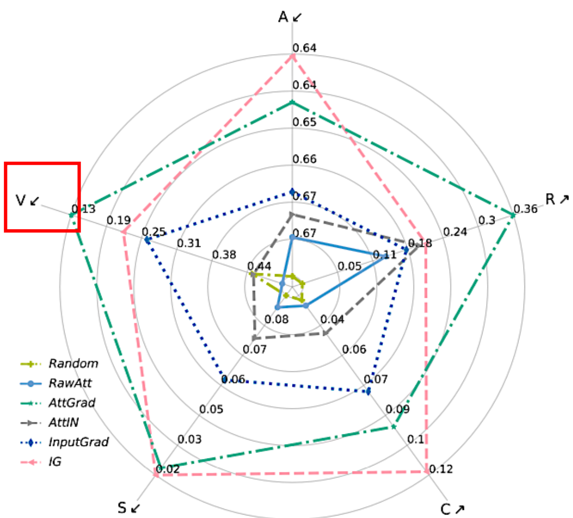


(c) LXMERT on GQA dataset

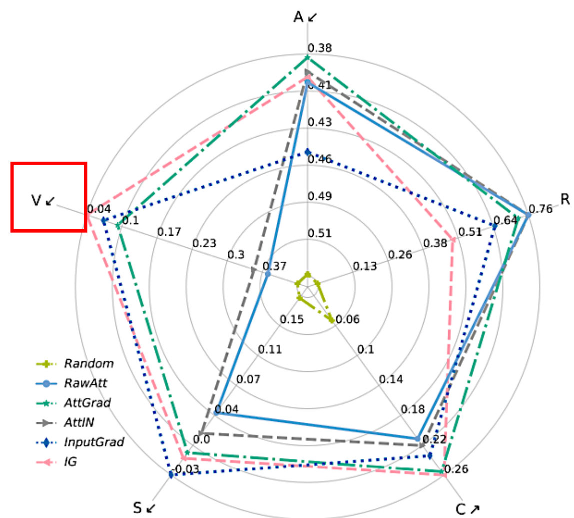
(More results are provided in our paper)

Sanity Faithfulness Evaluation (RQ2)

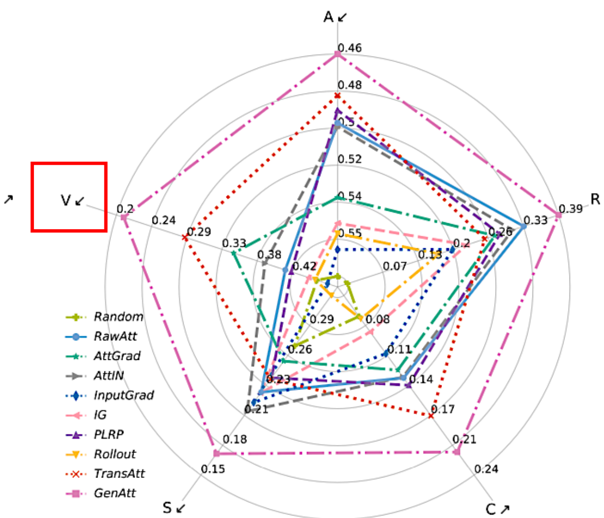
Most tested explanation methods suffer from the faithfulness violation issue regarding polarity consistency.



(a) LSTM+DotAtt on QQP dataset



(b) BUTD on VQA 2.0 dataset



(c) LXMERT on GQA dataset

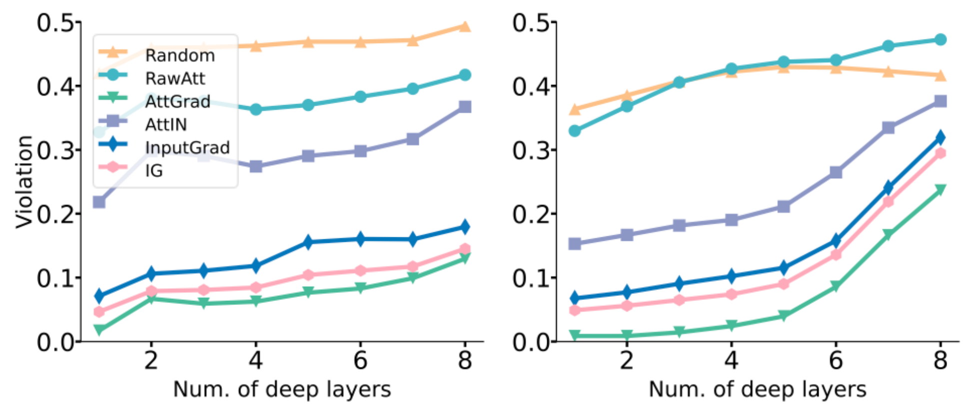
(More results are provided in our paper)

Factor Analysis (RQ3)

Two dominant factors

- The capability to identify polarity
- The complexity of model architectures

Method	Yelp	AgNews	VQA 2.0
α	0.31	0.28	0.40
$\alpha \odot \nabla\alpha$	**0.02**	**0.03**	**0.06**
$\alpha \odot \lvert\nabla\alpha\rvert$	0.15	0.07	0.25
$\alpha \odot \mathrm{sign}(\nabla\alpha)$	0.16	0.18	0.27



(a) Yelp dataset (b) SST dataset

Thank you!

Paper

Code