Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining

Andreas Madsen[1,2] **Nicholas Meade**[1,3,*] **Vaibhav Adlakha**[1,3,*] **Siva Reddy**[1,3,4]

[1] Mila - Quebec AI Institute [2] Polytechnique Montreal [3] McGill University [4] Facebook CIFAR AI Chair

Introduction

Importance measures, such as attention, are frequently used to explain which words are important. But is the explanation true to the model?

Measuring the truthfulness of an explanation is called faithfulness. This is non-trivial because the true explanation is unknown. Therefore different works use proxy-measures to determine faithfulness. These works have important limitations, which often make them speculative.

In this work, we introduce the proxy-measure Recursive ROAR, an improvement on ROAR, which can determine if an explanation is either faithful or non-faithful.

Recursive ROAR - A Faithfulness Test

ROAR principle: If information (tokens) is truly important, then removing (masking) it and retraining the model should result in a worse model. Retraining is necessary to prevent out-of-distribution issues. (Hooker et al., 2019)

However, ROAR has an important problem which we solve with Recursive ROAR.

0%	The	movie	is	great	.	I	really	liked	it	.
10%	The	movie	is	[MASK]	.	I	really	liked	it	.
20%	The	[MASK]	is	[MASK]	.	I	really	liked	it	.

ROAR: Redundancies in the dataset can keep the performance high.

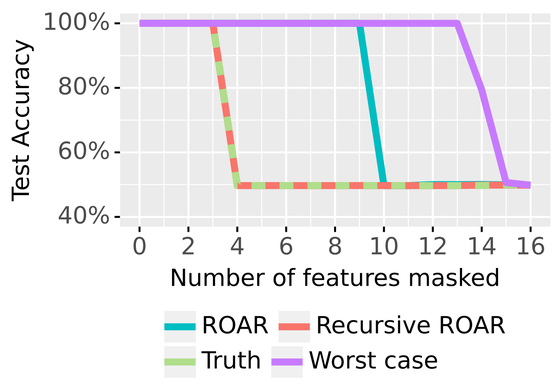
0%	The	movie	is	great	.	I	really	liked	it	.
10%	The	movie	is	[MASK]	.	I	really	liked	it	.
20%	The	movie	is	[MASK]	.	I	really	[MASK]	it	.

Recursive ROAR: By recursively re-evaluating the importance measures, the redundancies can be removed.

Theoretical validation:
We validate Recursive ROAR on the synthetic problem

$$\mathbf{x} = \frac{\mathbf{a}z}{10} + \mathbf{d}\eta + \frac{\epsilon}{10}, \quad y = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0 \end{cases}.$$

Because the optimal removal order is known. We can show that Recursive ROAR performs identical to the theoretical best case.



RACU (Relative Area Between Curves) - A Faithfulness Metric

The most faithful explanation is when the performance drops quickly, as allegedly important tokens are masked. This can be compared with removing random tokens.

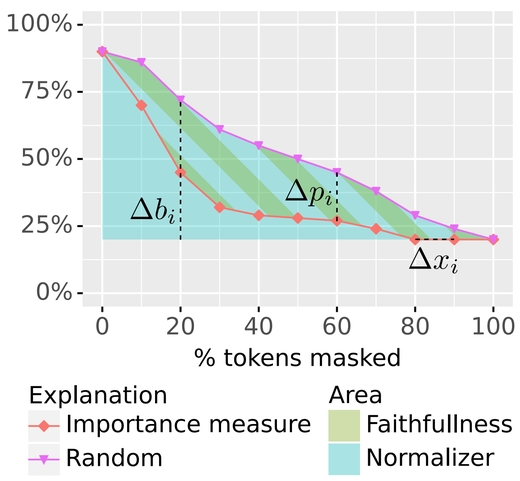
Faithfulness can therefore be quantified as an area-between-curves, of the importance measure ROAR curve and the random baseline ROAR curve. Additionally, this area can be compared with the theoretically-best explanation, which gets the performance of 100% masking immediately.

$$\text{RACU} = \frac{\sum_{i=1}^{I-1} \frac{1}{2}\Delta x_i(\Delta p_i + \Delta p_{i+1})}{\sum_{i=1}^{I-1} \frac{1}{2}\Delta x_i(\Delta b_i + \Delta b_{i+1})}$$

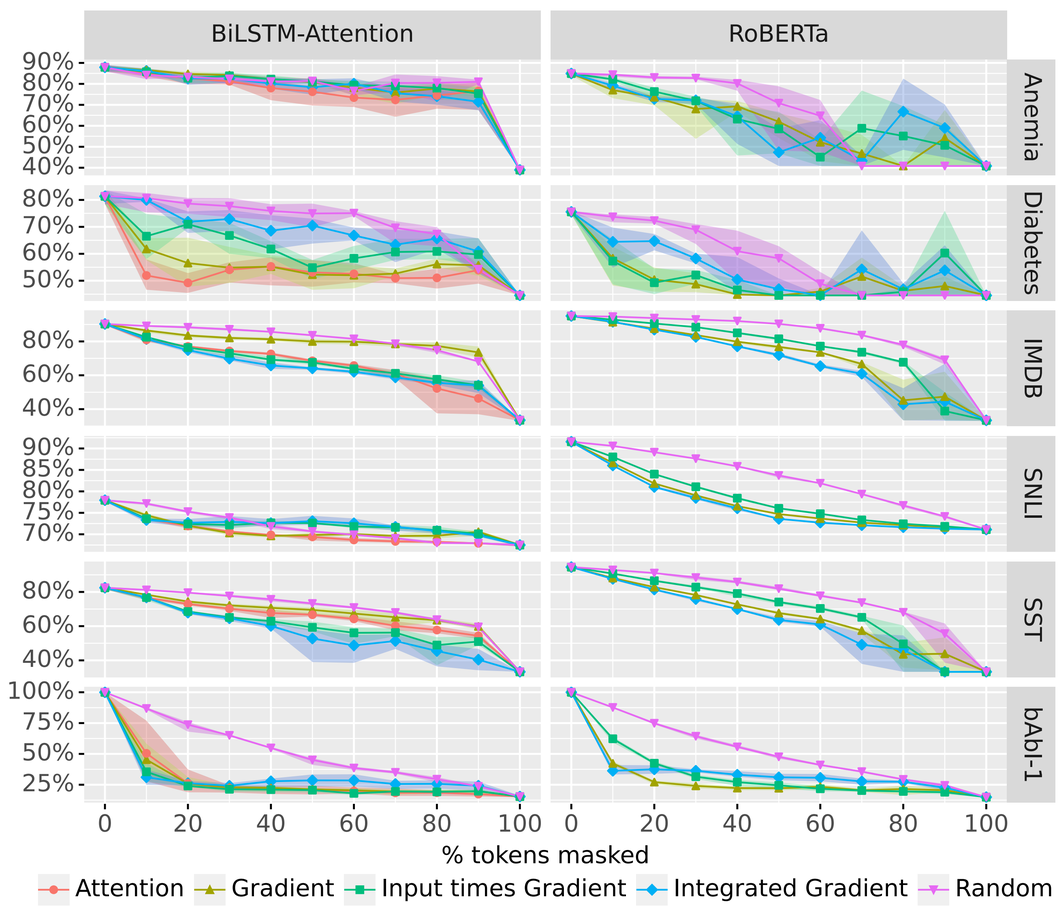
where $\Delta x_i = x_{i+1} - x_i$ *step size*

$\Delta p_i = b_i - p_i$ *performance delta*

$\Delta b_i = b_i - b_I$ *baseline delta*



Results



Qualitative: Shows the performance using Recursive ROAR. The more an importance measure is below the random baseline, the more faithful it is.

Dataset	Importance Measure	RACU Faithfulness [%] LSTM	RACU Faithfulness [%] RoBERTa
Anemia	Attention	$7.6^{+7.9}_{-6.8}$	–
	Gradient	$1.0^{+2.8}_{-4.1}$	$18.2^{+11.8}_{-13.8}$
	x ⊙ Gradient	$0.8^{+2.5}_{-3.5}$	$8.8^{+22.7}_{-22.8}$
	IG	$4.9^{+2.7}_{-1.8}$	$12.5^{+11.3}_{-7.0}$
Diabetes	Attention	$66.5^{+6.5}_{-13.0}$	–
	Gradient	$57.4^{+7.8}_{-7.0}$	$57.9^{+14.4}_{-19.8}$
	x ⊙ Gradient	$33.7^{+7.0}_{-15.7}$	$53.4^{+23.2}_{-29.3}$
	IG	$11.4^{+8.4}_{-15.0}$	$26.1^{+12.0}_{-25.1}$
IMDB	Attention	$29.8^{+5.0}_{-3.4}$	–
	Gradient	$3.1^{+2.4}_{-3.3}$	$25.4^{+3.1}_{-2.0}$
	x ⊙ Gradient	$28.4^{+1.0}_{-0.9}$	$16.9^{+1.1}_{-3.0}$
	IG	$32.5^{+0.9}_{-1.0}$	$35.1^{+2.4}_{-1.7}$
SNLI	Attention	$36.5^{+3.0}_{-3.5}$	–
	Gradient	$18.7^{+5.1}_{-3.5}$	$50.7^{+1.1}_{-0.8}$
	x ⊙ Gradient	$-10.7^{+6.1}_{-5.7}$	$41.0^{+0.4}_{-0.5}$
	IG	$-13.9^{+5.0}_{-5.0}$	$56.7^{+1.0}_{-1.1}$

Dataset	Importance Measure	RACU Faithfulness [%] LSTM	RACU Faithfulness [%] RoBERTa
SST	Attention	$15.7^{+2.4}_{-2.4}$	–
	Gradient	$7.6^{+2.3}_{-2.0}$	$26.1^{+1.6}_{-2.2}$
	x ⊙ Gradient	$28.0^{+5.6}_{-4.4}$	$18.6^{+4.1}_{-4.6}$
	IG	$37.8^{+4.6}_{-5.3}$	$32.9^{+1.8}_{-1.5}$
bAbI-1	Attention	$66.5^{+9.2}_{-9.2}$	–
	Gradient	$66.1^{+5.9}_{-6.5}$	$64.2^{+2.6}_{-2.6}$
	x ⊙ Gradient	$71.2^{+4.0}_{-4.2}$	$52.1^{+1.8}_{-3.7}$
	IG	$59.1^{+6.8}_{-7.4}$	$48.2^{+4.1}_{-5.7}$
bAbI-2	Attention	$75.4^{+4.9}_{-8.1}$	–
	Gradient	$66.3^{+4.2}_{-5.1}$	$57.8^{+2.0}_{-2.0}$
	x ⊙ Gradient	$66.7^{+8.0}_{-12.4}$	$48.1^{+3.2}_{-3.5}$
	IG	$34.6^{+13.4}_{-14.8}$	$42.0^{+3.8}_{-4.8}$
bAbI-3	Attention	$77.7^{+9.6}_{-8.1}$	–
	Gradient	$73.0^{+9.1}_{-7.6}$	$34.0^{+14.6}_{-15.1}$
	x ⊙ Gradient	$53.9^{+10.7}_{-24.1}$	$22.4^{+15.9}_{-12.4}$
	IG	$25.9^{+8.5}_{-9.1}$	$-27.9^{+18.0}_{-49.1}$

RACU: Shows the RACU score using Recursive ROAR. With a 95% confidence interval. *IG is Integrated Gradient and x ⊙ Gradient is Input times Gradient.*

Conclusions

1. In synthetic settings, Recursive ROAR matches the ground truth.
2. Faithfulness is both model and task-dependent.

Limitations

1. Computationally expensive to retrain models.
2. Can only comment on the model architecture and not the model instance.
3. Can leak the true class via masking and retraining.
4. Only measures faithfulness.

References

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In Advances in Neural Information Processing Systems, volume 32.

Mila · Polytechnique Montréal · McGill

Read full paper on arXiv
arxiv.org/abs/2110.08412