Pathologies of Neural Models Make Interpretations Difficult

Shi Feng[1] Eric Wallace[1] Alvin Grissom II[2] Mohit Iyyer[3,4] Pedro Rodriguez[1] Jordan Boyd-Graber[1]

[1]*University of Maryland* [2]*Ursinus College* [3]*UMass Amherst* [4]*Allen Institute for Artificial Intelligence*

Abstract

One way to interpret a neural text classifier is to highlight the most important words in the input. Existing methods of word importance estimate reply largely on model confidence—either directly by removing each word from the input, or indirectly with input gradient. To understand the limitations of confidence-based interpretation, we use input reduction, which iteratively removes the least important word. Input reduction produces nonsensical examples that trigger the same model prediction with high confidence. We explain this type of pathological behavior from the perspective of uncertainty estimate and adversarial examples. To mitigate the model deficiencies, we take the reduced examples and impose an entropy regularization. Fine-tuned models become more interpretable under input reduction without accuracy loss on regular examples.

Pathological Examples

Neural models are known to have issues...

SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	0.78 → 0.91

VQA

Original	What color is the flower ?
Answer	yellow
Reduced	**flower ?**
Confidence	0.827 → 0.819

SNLI

Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Reduced	**dancing**
Confidence	0.977 → 0.706

Input Reduction

How do we generate those examples?

Question	Confidence
What did Tesla ~~spend~~ Astor's money on ?	0.78
What did Tesla Astor's ~~money~~ on ?	0.74
What did Tesla Astor's ~~on~~ ?	0.76
~~What~~ did Tesla Astor's ?	0.80
did Tesla Astor's ~~?~~	0.87
did ~~Tesla~~ Astor's	0.82
did ~~Astor's~~	0.89
did	0.91

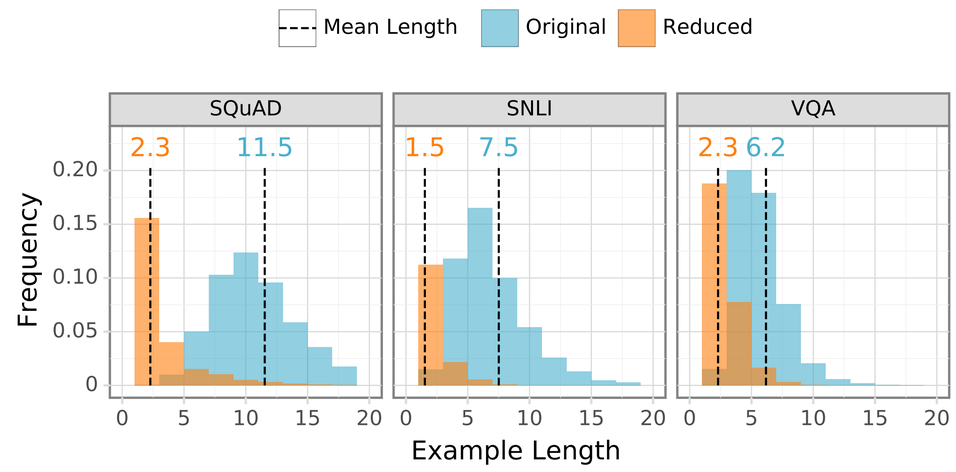
Leave-One-Out

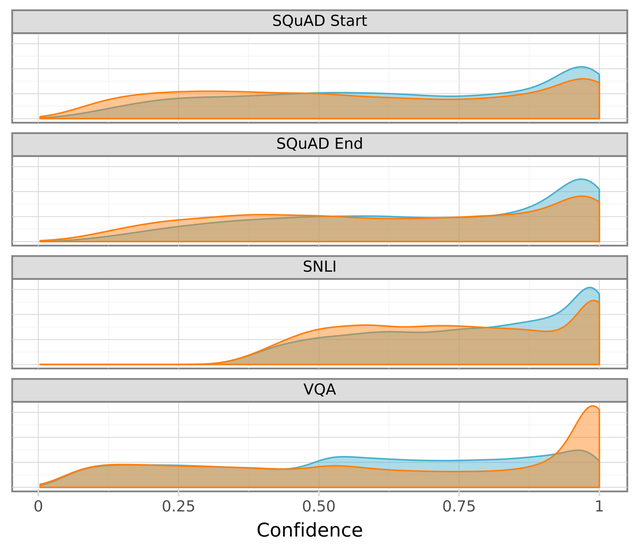
Input reduction is not stupid...

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What
What ~~did~~ Tesla spend Astor's money on ?	0.72	did
What did ~~Tesla~~ spend Astor's money on ?	0.66	Tesla
What did Tesla ~~spend~~ Astor's money on ?	0.74	spend
What did Tesla spend ~~Astor's~~ money on ?	0.76	Astor's
What did Tesla spend Astor's ~~money~~ on ?	**0.48**	money
What did Tesla spend Astor's money ~~on~~ ?	0.72	on
What did Tesla spend Astor's money on ~~?~~	0.73	?

On the Dataset Scale

This happens on the dataset scale...





Human Are Confused

The reduced examples are indeed nonsensical...

Dataset	Original	Reduced	vs. Random
SQuAD	80.58	31.72	53.70
SNLI-E	76.40	27.66	42.31
SNLI-N	55.40	52.66	50.64
SNLI-C	76.20	60.60	49.87
VQA	76.11	40.60	61.60

Heatmap Shifts

Importance measured individually ignores high-order correlation between words...

SQuAD

QuickBooks sponsored a "Small Business Big Game" contest, in which Death Wish Coffee had a 30-second commercial aired free of charge courtesy of QuickBooks. Death Wish Coffee beat out nine other contenders from across the United States for the free advertisement.

What	company	won	free	advertisement	due	to	QuickBooks	contest	?
What	company	won	free	advertisement	due	to	QuickBooks	?	
What	company	won	free	advertisement	due	to	?		
What	company	won	free	due	to	?			
What	won	free	due	to	?				

Mitigation

There is still hope...

$$\sum_{(\mathbf{x},y)} \log(f(y \mid \mathbf{x})) + \lambda \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \mathbb{H}\left(f(y \mid \tilde{\mathbf{x}})\right)$$

Conclusions

- Confidence of a neural model trained with MLE is not reliable for interpretation
- Existing interpretation methods ignore high-order correlation between words