Pathologies of Neural Models Make Interpretation 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

- We interpret text classifiers by highlighting important words in the input.
- The highlights are usually computed based on model confidence.
- We use input reduction to expose pathologies of model confidence;
- explain why this makes interpretation difficult;
- and propose a simple mitigation.

Pathological Examples

Neural model confidence is known to have issues. Here we show a particular case: models making the same predictions with high confidence even when the inputs are reduced to only a few words and appear non-sensical to humans.

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

Interpretation with Leave-One-Out

- Saliency maps indicate the "importance" of each word to the model's prediction.
- Simple importance function: remove each word and measure the confidence decrease.

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	?

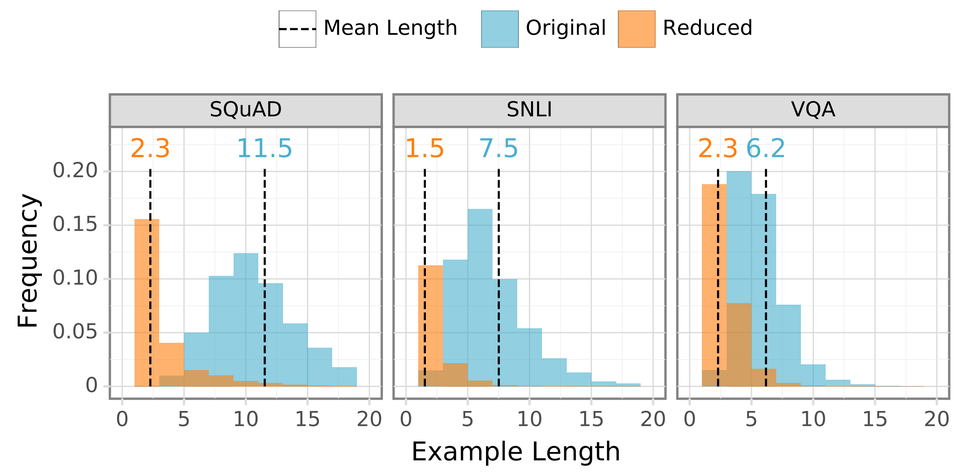
Input Reduction

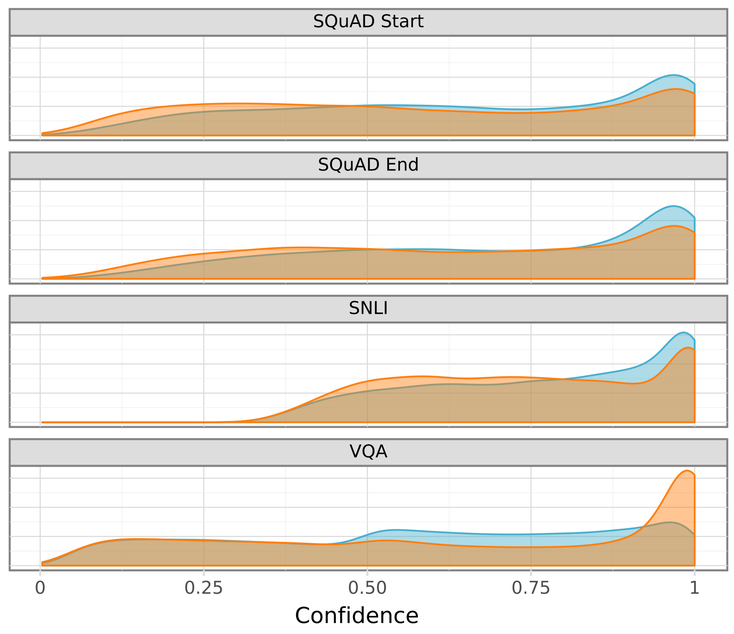
- The way we generate the pathological examples follow the same principle.
- We iteratively remove the least important words from the input.

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

Reduced Examples Are Extremely Short

- All examples in the validation set can be drastically reduced.
- Model confidence remains high.





Reduced Examples Are Non-sensical

- And the reduced examples are not just short but also non-sensical to humans.

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 is measured for each word individually.
- High-order correlations between between words are ignored.
- Removing an unimportant words can lead to a significant drop in the importance of an important word.

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

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

- Treat reduced examples as negative.
- Models should not confidently predict any label.
- Maximize the entropy on reduced examples.