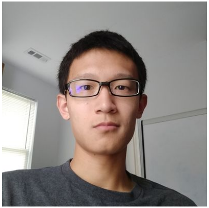
Pathologies of Neural Models Make Interpretation Difficult

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]AI2

November 14, 2018

Authors



Shi Feng
UMD

Eric Wallace
UMD

Alvin Grissom II
Ursinus College

Mohit Iyyer
UMass + AI2

Pedro Rodriguez
UMD

Jordan Boyd-Graber
UMD

- Neural networks make strong text classifiers.

- Neural networks make strong text classifiers.
- But, are they doing the “right” things?

Highlighting Important Words

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.
Question	What did Tesla spend Astor's money on ?
Highlights	What did Tesla spend Astor's money on ?

Importance of Words

Leave-one-out: remove a word and measure the decrease in confidence (Li et al., 2016)

Question	Confidence	Highlight
What did Tesla spend Astor’s money on ?	**0.78**	

Importance of Words

Leave-one-out: remove a word and measure the decrease in confidence (Li et al., 2016)

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What

Importance of Words

Leave-one-out: remove a word and measure the decrease in confidence (Li et al., 2016)

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

Importance of Words

Leave-one-out: remove a word and measure the decrease in confidence (Li et al., 2016)

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	?

Importance of Words

Leave-one-out: remove a word and measure the decrease in confidence (Li et al., 2016)

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	?

What did Tesla spend Astor's money on ?

Gradient-based Approximation

Approximate a word's removal using the input gradient (Simonyan et al., 2014):

$$\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial \boldsymbol{v}_i} \cdot \boldsymbol{v}_i$$

- Computes importance for all words in one backward pass.

Input Reduction

- What if we remove the unimportant words?

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

Question	Confidence
What did Tesla ~~spend~~ Astor's money on ?	0.78

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

Question	Confidence
What did Tesla ~~spend~~ Astor’s money on ?	0.78
What did Tesla Astor’s ~~money~~ on ?	0.74

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

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

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

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

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

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

Prediction remains the same.

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

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

What remains does not match what was considered important.

Input Reduction

- What if we remove the unimportant words?
- Keeping the model prediction the same.

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**

Model is confident when no reasonable prediction can be made.

SQuAD

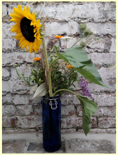
Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	0.78 → 0.91

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

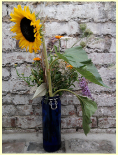


SQuAD

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

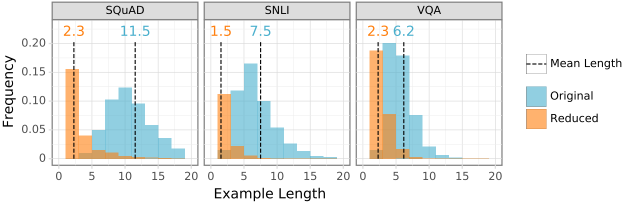
Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Reduced	**dancing**
Confidence	0.977 → 0.706

All Examples are Drastically Reduced

- Run input reduction for entire validation set.
- Keep model prediction the same.

All Examples are Drastically Reduced

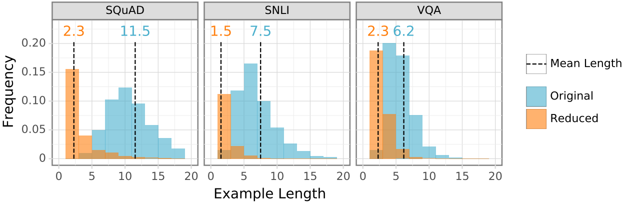
- Run input reduction for entire validation set.
- Keep model prediction the same.



- Consistently reduce examples to very short lengths without changing the model prediction.

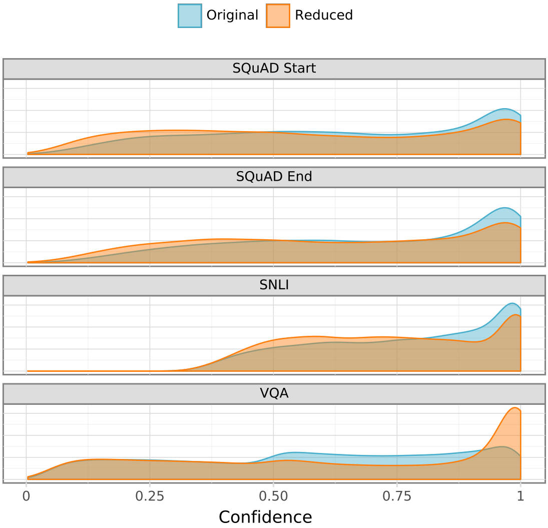
All Examples are Drastically Reduced

- Run input reduction for entire validation set.
- Keep model prediction the same.



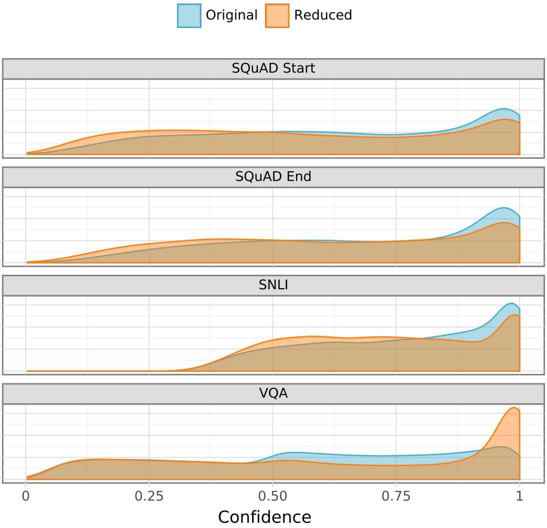
- Consistently reduce examples to very short lengths without changing the model prediction.
- **But how about the confidence?**

Confidence Remains High



- Model confidence remains high on reduced examples.

Confidence Remains High



- Model confidence remains high on reduced examples.

Humans Are Confused by Reduced Inputs

Dataset	Original	Reduced
SQuAD	80.58	31.72
SNLI-E	76.40	27.66
SNLI-N	55.40	52.66
SNLI-C	76.20	60.60
VQA	76.11	40.60

Humans Are Confused by Reduced Inputs

Dataset	Original	Reduced
SQuAD	80.58	31.72
SNLI-E	76.40	27.66
SNLI-N	55.40	52.66
SNLI-C	76.20	60.60
VQA	76.11	40.60

What did Tesla spend Astor's money on ?

did	spend ✔

Humans Are Confused by Reduced Inputs

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

What did Tesla spend Astor's money on ?

did	spend ✓

Humans Are Confused by Reduced Inputs

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

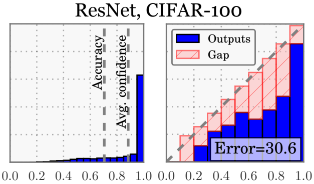
What did Tesla spend Astor's money on ?

did | spend ✔

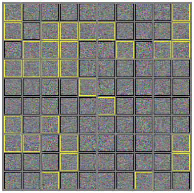
- Reduced examples are uninformative and appear random.
- **How did input reduction lead to this?**

Let’s Take a Step Back

Model Overconfidence
Guo et al. (2017)

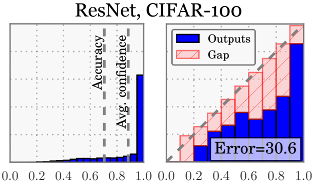


Rubbish Examples
Goodfellow et al. (2015)

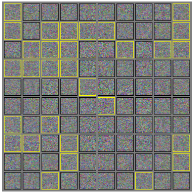


Let's Take a Step Back

Model Overconfidence
Guo et al. (2017)



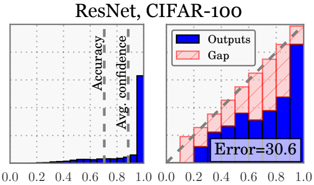
Rubbish Examples
Goodfellow et al. (2015)



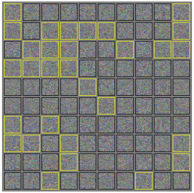
- Overconfidence does not cover non-sensical inputs.

Let’s Take a Step Back

Model Overconfidence
Guo et al. (2017)



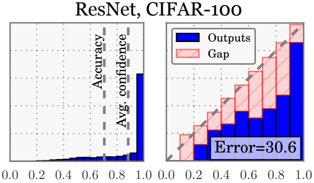
Rubbish Examples
Goodfellow et al. (2015)



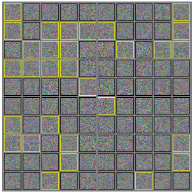
- Overconfidence does not cover non-sensical inputs.
- Reduced examples are rubbish examples.

Let's Take a Step Back

Model Overconfidence
Guo et al. (2017)



Rubbish Examples
Goodfellow et al. (2015)



- Overconfidence does not cover non-sensical inputs.
- Reduced examples are rubbish examples.
- **How did input reduction lead to rubbish examples?**

Issues of Linear, Confidence-based Interpretation

SQuAD

The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question	Confidence
Where did the Broncos practice for the Super Bowl ?	(0.90, 0.89)

Issues of Linear, Confidence-based Interpretation

SQuAD

The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question	Confidence
Where did the Broncos practice for the Super Bowl ?	(0.90, 0.89)
Where did the practice for the Super Bowl ?	(0.92, 0.88)

Issues of Linear, Confidence-based Interpretation

SQuAD

The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question	Confidence
Where did the Broncos practice for the Super Bowl ?	(0.90, 0.89)
Where did the practice for the Super Bowl ?	(0.92, 0.88)

- After the first reduction step, the input is already rubbish.

Issues of Linear, Confidence-based Interpretation

SQuAD

The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question	Confidence
Where did the Broncos practice for the Super Bowl ?	(0.90, 0.89)
Where did the practice for the Super Bowl ?	(0.92, 0.88)

- After the first reduction step, the input is already rubbish.
- Confidence remains high after the crucial word is removed.

Issues of Linear, Confidence-based Interpretation

SQuAD

The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question	Confidence
Where did the Broncos practice for the Super Bowl ?	(0.90, 0.89)
Where did the practice for the Super Bowl ?	(0.92, 0.88)

- After the first reduction step, the input is already rubbish.
- Confidence remains high after the crucial word is removed.
- Decrease in confidence does not align with importance.
  - Confidence $\neq$ Uncertainty

Issues of Linear, Confidence-based Interpretation

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 ?

Issues of Linear, Confidence-based Interpretation

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 ?

Issues of Linear, Confidence-based Interpretation

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 ?

Issues of Linear, Confidence-based Interpretation

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 ?

Issues of Linear, Confidence-based Interpretation

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 ?

- Independent word importance implicitly assumes bag-of-words.
- Higher-order correlations are ignored.

Mitigating Pathologies by Entropy Regularization

Mitigating Pathologies by Entropy Regularization

- Ideally, model should say “I don’t know”.
- Uniform distribution over classes.

Mitigating Pathologies by Entropy Regularization

- Ideally, model should say "I don't know".
- Uniform distribution over classes.
- Maximize the output entropy on **reduced examples**:

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

  where $\tilde{\mathcal{X}}$ is the set of reduced training examples.
- Fine-tune models with both MLE and entropy regularization.

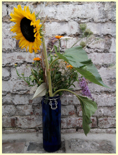
Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Before	**did**
After	**spend Astor money on ?**
Confidence	0.78 → 0.91 → 0.52

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 ?
Before	**did**
After	**spend Astor money on ?**
Confidence	0.78 → 0.91 → 0.52

Original	What color is the flower ?
Answer	yellow
Before	**flower ?**
After	**What color is flower ?**
Confidence	0.847 → 0.918 → 0.745

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 ?
Before	**did**
After	**spend Astor money on ?**
Confidence	0.78 → 0.91 → 0.52

Original	What color is the flower ?
Answer	yellow
Before	**flower ?**
After	**What color is flower ?**
Confidence	0.847 → 0.918 → 0.745



Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Before	**dancing**
After	**two man dancing**
Confidence	0.977 → 0.706 → 0.717

Input Reduction After Regularization

	Accuracy	
	Before	After
SQuAD	77.41	78.03
SNLI	85.71	85.72
VQA	61.61	61.54

Input Reduction After Regularization

	Accuracy		Reduced Lengths	
	Before	After	Before	After
SQuAD	77.41	78.03	2.27	4.97
SNLI	85.71	85.72	1.50	2.20
VQA	61.61	61.54	2.30	2.87

Input Reduction After Regularization

	Accuracy		Reduced Lengths	
	Before	After	Before	After
SQuAD	77.41	78.03	2.27	4.97
SNLI	85.71	85.72	1.50	2.20
VQA	61.61	61.54	2.30	2.87

- Human studies show examples are more meaningful.

Summary

- Neural models are overconfident → interpretation is difficult.
  - Poor uncertainty estimates from MLE training.
  - Entropy regularization on reduced examples helps mitigate.

Summary

- Neural models are overconfident → interpretation is difficult.
  - Poor uncertainty estimates from MLE training.
  - Entropy regularization on reduced examples helps mitigate.

- Gradient interpretations assume linear model (bag-of-words).
  - Neglects curvature (Hessian) and higher-order terms.

Summary

- Neural models are overconfident → interpretation is difficult.
  - Poor uncertainty estimates from MLE training.
  - Entropy regularization on reduced examples helps mitigate.

- Gradient interpretations assume linear model (bag-of-words).
  - Neglects curvature (Hessian) and higher-order terms.

- **Shameless Plugs:**
  - Competition on robust QA Models `www.qanta.org`
  - Take me as your student!

Reduced Examples Become More Meaningful

	Accuracy	
	Before	After
SQuAD	31.72	51.61
SNLI-E	27.66	32.37
SNLI-N	52.66	50.50
SNLI-C	60.60	63.90
VQA	40.60	51.85

Reduced Examples Become More Meaningful

	Accuracy		vs. Random	
	Before	After	Before	After
SQuAD	31.72	51.61	53.70	62.75
SNLI-E	27.66	32.37	42.31	50.62
SNLI-N	52.66	50.50	50.64	58.94
SNLI-C	60.60	63.90	49.87	56.92
VQA	40.60	51.85	61.60	61.88

- Input reduction leads to more meaningful examples after regularization.
- Entropy regularization helps mitigate the pathology.

References I

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Daniel Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *NAACL*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*.