Partial-input baselines show that NLI models can ignore context, but they don't

Neha Srikanth , Rachel Rudinger

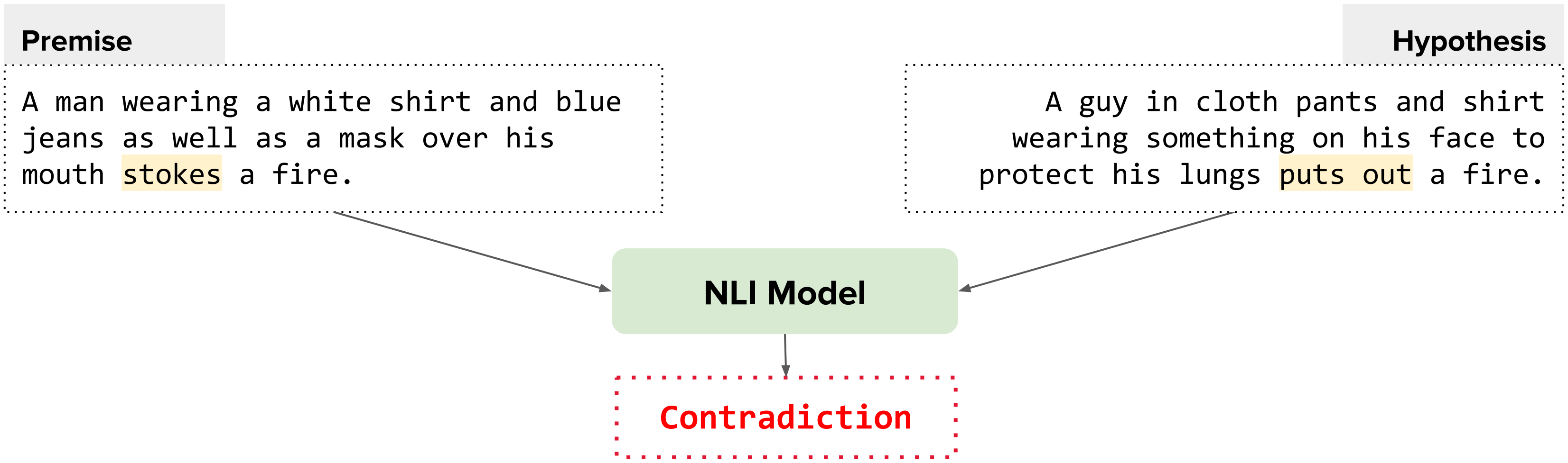
Department of Computer Science
University of Maryland, College Park

nehasrik@umd.edu, rudinger@umd.edu

Motivation

Natural language inference: predicting a *directional* relationship between pairs of text expressions

A necessary, but not sufficient, condition of true inferential reasoning is the ability for natural language inference (NLI) models to **utilize all parts of the example's input.**



Motivation

Recent work has illustrated the presence of **annotation artifacts**[1], or statistical biases, in parts of NLI instances (e.g. the hypothesis) that are predictive of the correct label.

Premise

`A man wearing a white shirt and blue jeans as well as a mask over his mouth stokes a fire.`

Hypothesis

`A guy in cloth pants and shirt wearing something on his face to protect his lungs puts out a fire.`

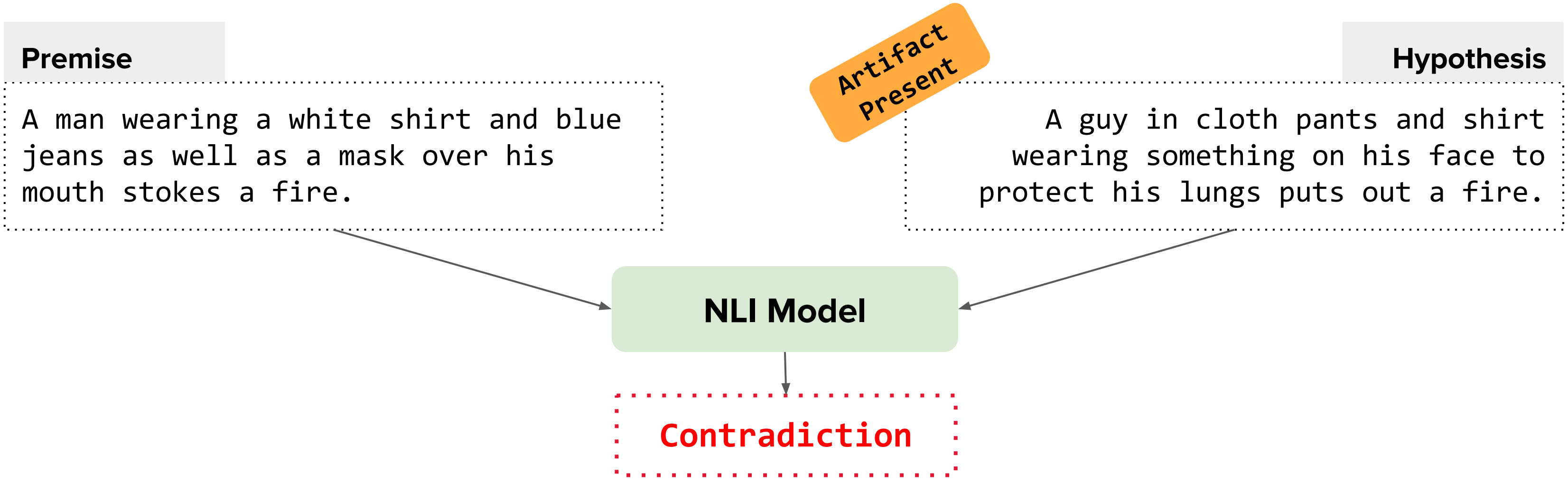
NLI Model

`Contradiction`

Motivation

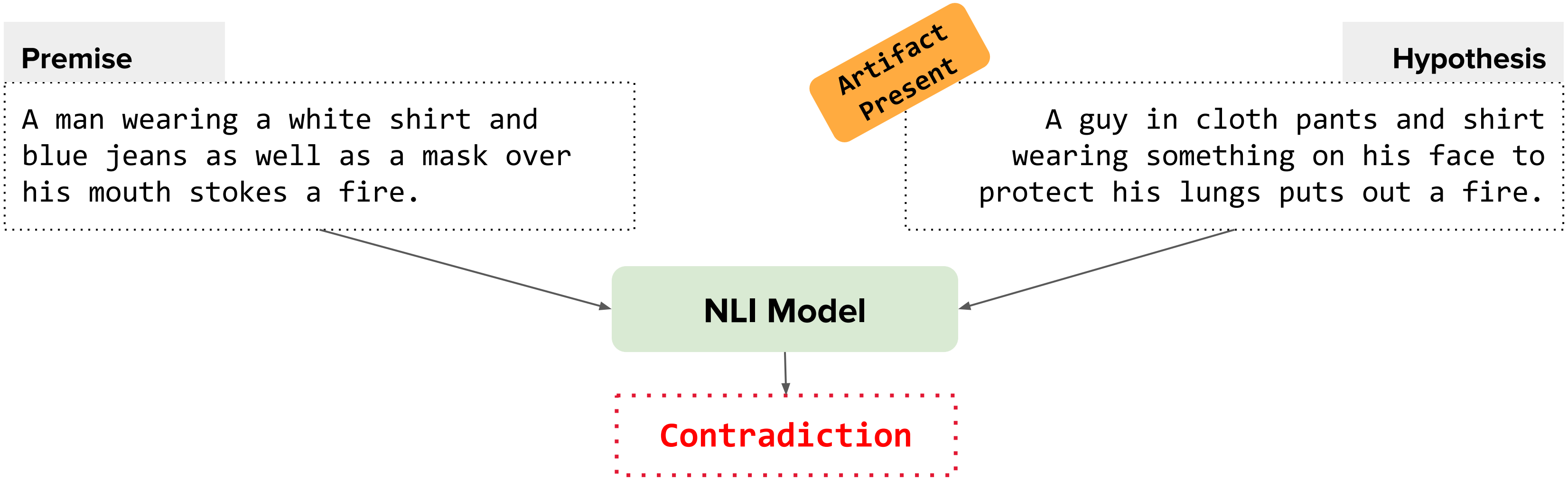
Recent work has illustrated the presence of **annotation artifacts**[1], or statistical biases, in parts of NLI instances (e.g. the hypothesis) that are predictive of the correct label.

Some suggest that the presence of such artifacts in datasets may in turn produce models that are incapable of learning to perform true reasoning.



Motivation

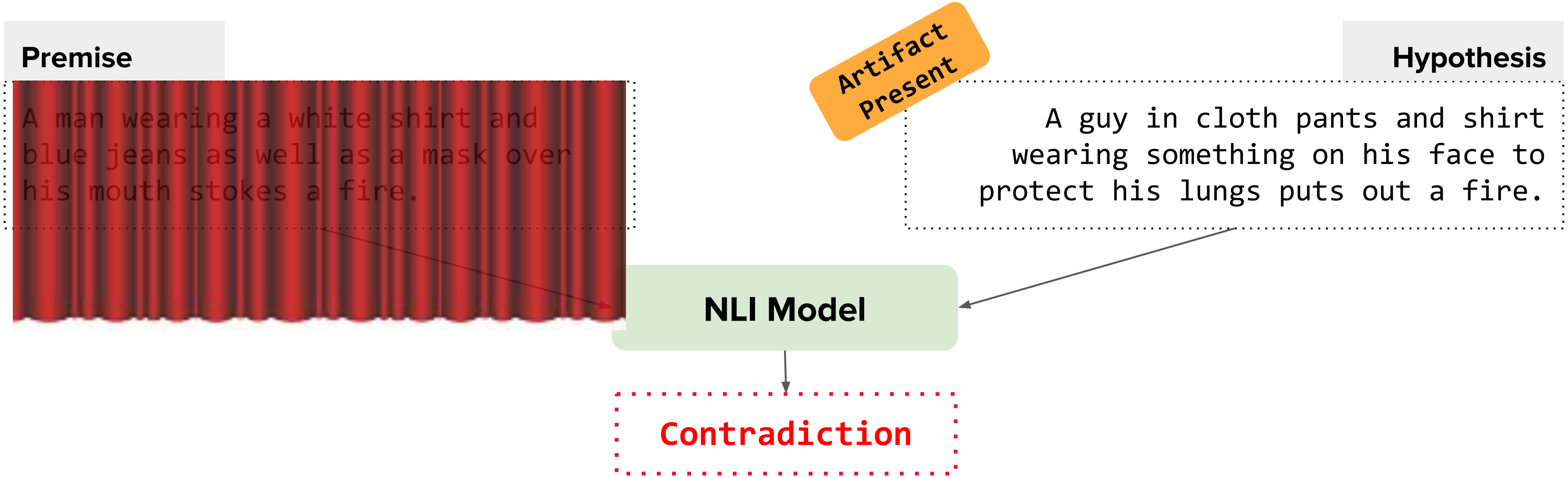
One way to detect the presence of artifacts in datasets through a **partial-input baseline**[2] in which only parts of NLI instances (e.g. only the hypothesis) are fed to a model trained to predict an entailment label.



Motivation

One way to detect the presence of artifacts in datasets through a **partial-input baseline**[2] in which only parts of NLI instances (e.g. only the hypothesis) are fed to a model trained to predict an entailment label.

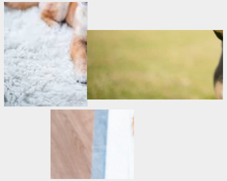
A strong partial-input baseline suggests that full input models can use “shortcuts” present in parts of the input to boost their performance.



Motivation

Dataset

Learn to do this task please! We're going to take away some context, though.



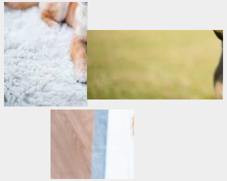
Accuracy

87%

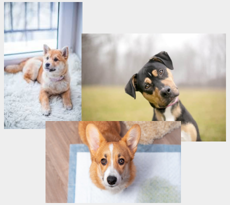
Motivation

Dataset

Learn to do this task please! We're going to take away some context, though.



Here's context back. Learn this task, please!

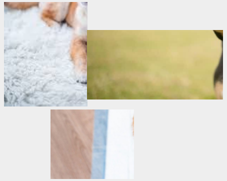


Accuracy
87%
92%

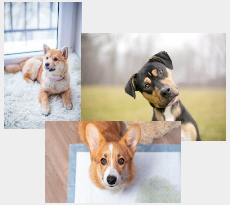
Motivation

Dataset

Learn to do this task please! We're going to take away some context, though.



Here's context back. Learn this task, please!



Accuracy
87%
92%

Central Question: Do NLI models **learn to condition on context** despite being trained on **artifact-ridden datasets?**

Contributions

We investigate the **role of context in NLI models** through two sets of experiments.

Experiment 1

Does **access to context strengthen a full-input model's confidence** in the correct label, despite a partial-input model's correct prediction?

Finding

Yes! Full-input models are more confident in the correct label than partial-input models.

Contributions

We investigate the **role of context in NLI models** through two sets of experiments.

Experiment 1: Context in NLI

Does **access to context strengthen a full-input model's confidence** in the correct label, despite a partial-input model's correct prediction?

Finding

Yes! Full-input models are more confident in the correct label than partial-input models.

Experiment 2: Context Editing

Are full-input models **sensitive to changes in non-target** components of the input (e.g. perturbations in the **premise**) when artifacts are present?

Finding

Yes! Full-input models are in fact sensitive to context modifications despite the artifacts in parts of the input.

Contributions

We investigate the **role of context in NLI models** through two sets of experiments.

Experiment 1: Context in NLI

Does **access to context strengthen a full-input model's confidence** in the correct label, despite a partial-input model's correct prediction?

Finding

Yes! Full-input models are more confident in the correct label than partial-input models.

Experiment 2: Context Editing

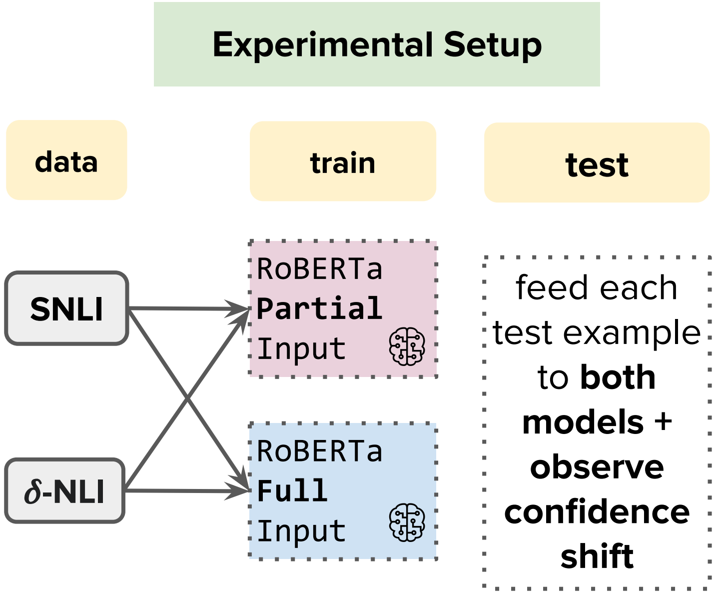
Are full-input models **sensitive to changes in non-target** components of the input (e.g. perturbations in the **premise**) when artifacts are present?

Finding

Yes! Full-input models are in fact sensitive to context modifications despite the artifacts in parts of the input.

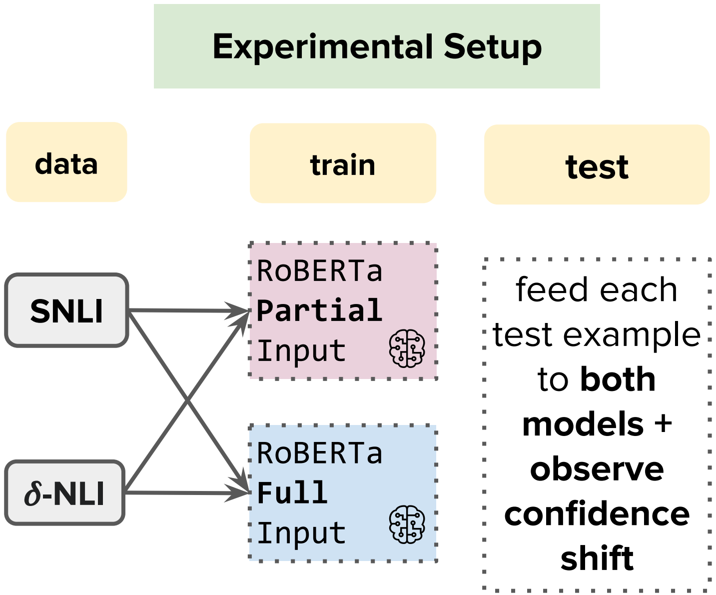
Experiment 1: Context in NLI

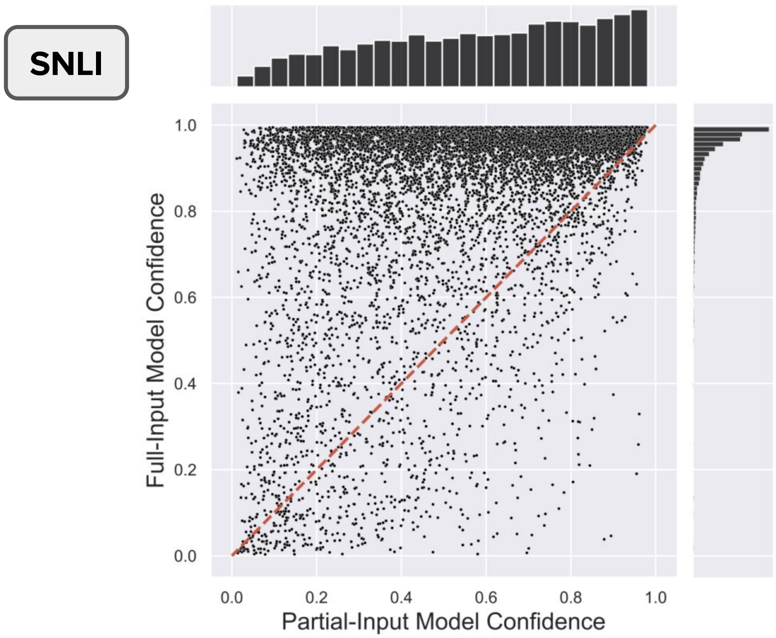
Does access to context **shift a full-input model's confidence** in the correct label?



Experiment 1: Context in NLI

Yes! Access to context **strengthens a full-input model's confidence** in the correct label.





Experiment 2: Context Editing

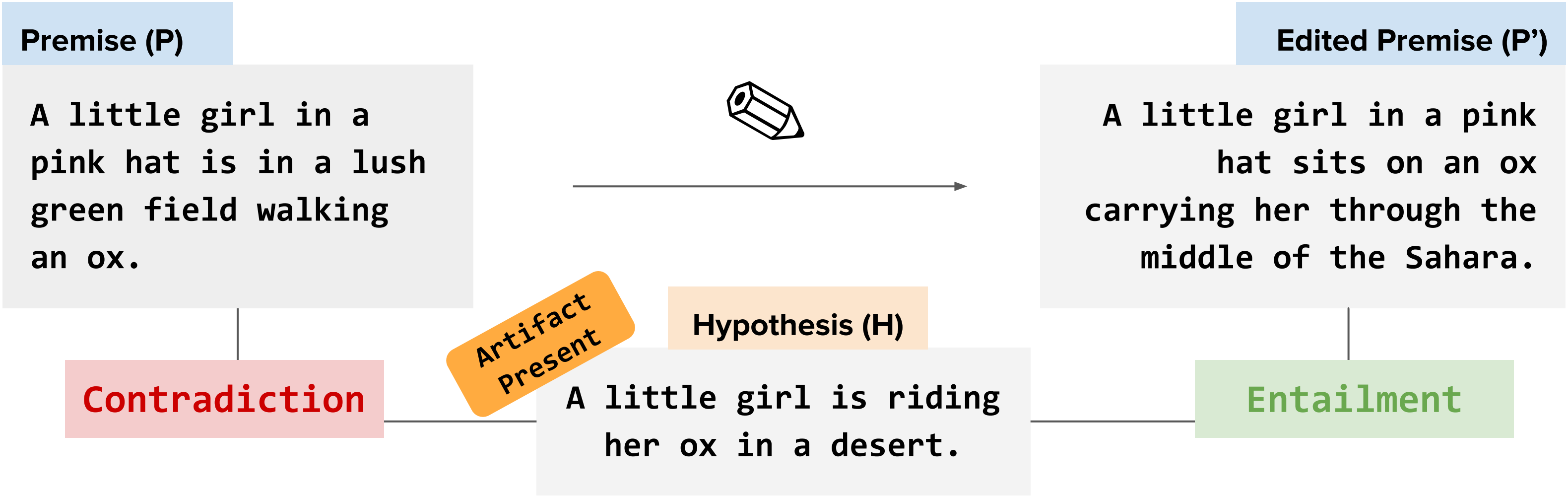
Are full-input models **sensitive to changes in non-target** components of the input (e.g. perturbations in the **premise**?)

We present an example modification scheme in which **we edit context sentences from examples where a model correctly predicts the label from the target alone**.

Context Editing

Are full-input models **sensitive to changes in non-target** components of the input (e.g. perturbations in the **premise**?)

We present an example modification scheme in which **we edit context sentences from examples where a model correctly predicts the label from the target alone**.



Experiment 2: Context Editing

Yes! Full-input models are **sensitive to changes in non-target** components of the input

SNLI		Edited Label (*l'*)		
		e	**n**	**c**
Original Label (*l*)	**e**	-	**0.76**	**0.76**
	n	0.42	-	**0.78**
	c	**0.90**	**0.78**	-

Consistent achievement of above 70% accuracy on edited examples illustrates full-input models are in fact sensitive to context modifications!

Experiment 2: Context Editing

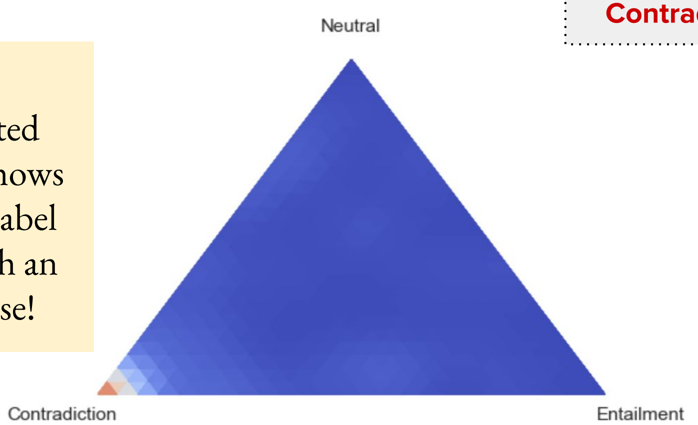
Are full-input models **sensitive to changes in non-target** components of the input (e.g. perturbations in the **premise**?)

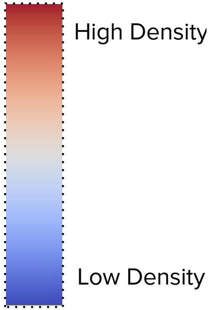
SNLI		Edited Label (*l'*)		
		e	n	c
Original Label (*l*)	e	-	**0.76**	**0.76**
	n	0.42	-	**0.78**
	c	**0.90**	**0.78**	-

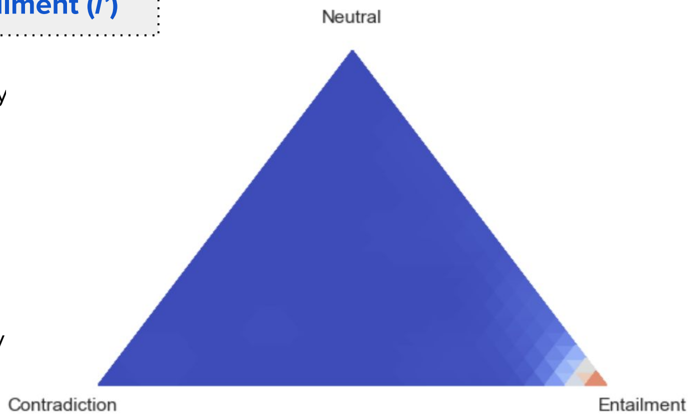
Consistent achievement of above 70% accuracy on edited examples illustrates full-input models are in fact sensitive to context modifications.

Heatmap of confidences plotted on the simplex shows shift when gold label is flipped through an edit to the premise!

Contradiction (*l*) → Entailment (*l'*)







Key Takeaways

Main Takeaway

It is hasty to conclude that models trained on artifact-ridden datasets are not capable of reasoning.

Even though high-scoring partial-input baselines show that full-input models could ignore context, our experiments show they don't: they can leverage this context quite effectively.

Key Takeaways

Main Takeaway

It is hasty to conclude that models trained on artifact-ridden datasets are not capable of reasoning.

Even though high-scoring partial-input baselines show that full-input models could ignore context, our experiments show they don't: they can leverage this context quite effectively.

Conclusion 1

Of course, artifacts can and do lead to models with exploitable heuristics, but:

Artifacts don't *necessarily* spell disaster for a model's reasoning capabilities!

Key Takeaways

Main Takeaway

It is hasty to conclude that models trained on artifact-ridden datasets are not capable of reasoning.

Even though high-scoring partial-input baselines show that full-input models could ignore context, our experiments show they don't: they can leverage this context quite effectively.

Conclusion 1

Of course, artifacts can and do lead to models with exploitable heuristics, but:

Artifacts don't *necessarily* spell disaster for a model's reasoning capabilities!

Conclusion 2

NLI models can and do meet one of the necessary conditions for reasoning: leveraging the full input.

This isn't a sufficient, but inherently necessary.

Key Takeaways

Main Takeaway

It is hasty to conclude that models trained on artifact-ridden datasets are not capable of reasoning.

Even though high-scoring partial-input baselines show that full-input models could ignore context, our experiments show they don't: they can leverage this context quite effectively.

Conclusion 1

Of course, artifacts can and do lead to models with exploitable heuristics, but:

Artifacts don't *necessarily* spell disaster for a model's reasoning capabilities!

Conclusion 2

NLI models can and do meet one of the necessary conditions for reasoning: leveraging the full input.

This isn't a sufficient, but inherently necessary.

Conclusion 3

Partial-input baselines should be understood as agnostic warning signs.

They are sufficient to conclude that full-input models might not be leveraging critical context, but insufficient to prove that they don't.

Resources + References

Read our paper at: https://arxiv.org/abs/2205.12181

Data + Annotations: https://github.com/nehasrikn/context-editing

References

[1] Annotation Artifacts in Natural Language Inference Data (Gururangan et al., NAACL 2018)
[2] Hypothesis Only Baselines in Natural Language Inference (Poliak et al., SemEval 2018)