Overcoming Spurious Correlations in NLP: Successes and Failures

He He

NEW YORK UNIVERSITY

AKBC, London

November 4, 2022

Sentence classification

label$= +1$	label$= -1$
Riveting film of the highest calibre!	Thank God I didn't go to the cinema.
Definitely worth the watch!	Boring as hell.
A true story told perfectly!	I wanted to give up in the first hour...

Two equally good hypotheses:

- Predict $+1$ if the input ends with "!"
- Predict $+1$ is the input gives a positive recommendation

Complete waste of two hours of my time! $+1/-1$?

Models may not generalize as expected in deployment domains

Real examples

- **NLI**: negation words $\rightarrow$ contradition [Poliak et al., 2018]
- **NLI**: lexical overlap $\rightarrow$ entailment [McCoy et al., 2019]
- **Paraphrase identification**: lexical overlap $\rightarrow$ paraphrase [Zhang et al., 2019]
- **QA**: lexical overlap $\rightarrow$ answer sentence [Jia and Liang, 2017]
- **Co-reference**: gender $\rightarrow$ occupation [Zhao et al., 2018]

Large performance drop when the simple heuristic fails

Real-world impact

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			

Figure: [Ribeiro et al., 2020]

Google sentiment analysis service

- Negation causes 76.4% failure rate
- Named entity causes 20.8% failure rate

Avoid learning spurious correlations

Input	Label	Quantity	Biased prediction
P: I love dogs H: I don't love dogs	con		$p(\text{con} \mid \text{don't}) = 0.8$
P: The bird is red H: The bird is not green	ent		$p(\text{ent} \mid \text{not}) = 0.1$

- Training loss does not tell the model that not $\rightarrow$ con is unreliable
- **Idea**: learn from examples where the heuristic fails
- **Assumption**: we know the spurious feature

Fitting the residual of a biased predictor

[He et al., 2019]

1. Train the biased classifier using only spurious features $\phi(x)$

$$\max \mathbb{E}_{x,y} \log p_{\text{bias}}(y \mid \phi(x))$$

2. Train the debiased classifier by fitting the residuals

$$\max \mathbb{E}_{x,y} \log \underbrace{\text{softmax}(\log p_{\text{bias}} + \log p_{\text{debias}})[y]}_{p(y \mid x) \propto p_{\text{bias}}(y \mid x) p_{\text{debias}}(y \mid x)}$$

3. Run inference using the debiased classifier

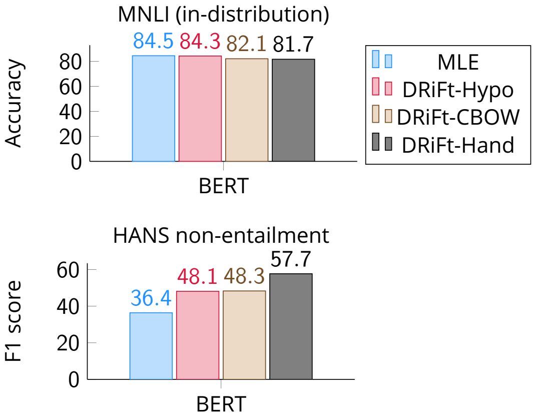
He, Zha, Wang. Unlearn Dataset Bias for Natural Language Inference by Fitting the Residual. EMNLP DeepLo workshop, 2019

Results

- **Train**: MNLI [Williams et al., 2017]
- **OOD Test**: HANS [McCoy et al., 2019]

  The doctors visited the lawyer.
  $\not\Longrightarrow$ The lawyer visited the doctors.
- **Spurious features**: hypothesis, BoW, overlapped words

Better knowledge of the spurious features leads to larger improvement



He, Zha, Wang. Unlearn Dataset Bias for Natural Language Inference by Fitting the Residual. EMNLP DeepLo workshop, 2019

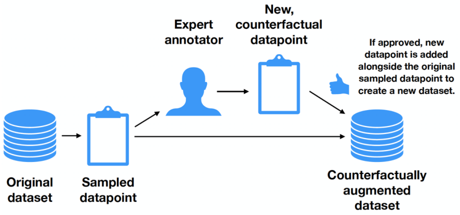
Summary

If we know the spurious features, we can “tell” the model not to use them.

If we don’t know the spurious features, is there a general way to improve robustness?

Can humans tell us which are causal vs spurious features?

Figure: Crowdsourcing counterfactually-augmented data (CAD) [Kaushik et al., 2020]



pos "Election" is a highly fascinating and thoroughly captivating thriller-drama
neg "Election" is a highly expected and thoroughly mind-numbing thriller-drama

- Assumption: edited spans are core features (that generalize to OOD)

Using CAD to improve OOD generalization

Incorporate CAD into training:

- Train on original data + CAD
- Consistency regularization on CAD pairs

Mixed results:

Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data

William Huang
New York University
`will.huang@nyu.edu`

More Bang for Your Buck:
Natural Perturbation for Robust Question Answering

Daniel Khashabi and **Tushar Khot** and **Ashish Sabharwal**
Allen Institute for AI, Seattle, WA, U.S.A.
`{danielk,tushark,ashishs}@allenai.org`

CAD reveals useful features, but why aren't they helpful?

Toy example: sentiment classification

[Joshi and He, 2022]

The book is good	pos
The book is not good	neg
The movie is boring	neg
The movie is fascinating	pos

Naive Bayes model weights:

data	book	movie	good	boring	fascinating	not
original	$+1$	-1	$+1$	-1	0	0
CAD	0	0	0	-0.5	$+0.5$	-0.5

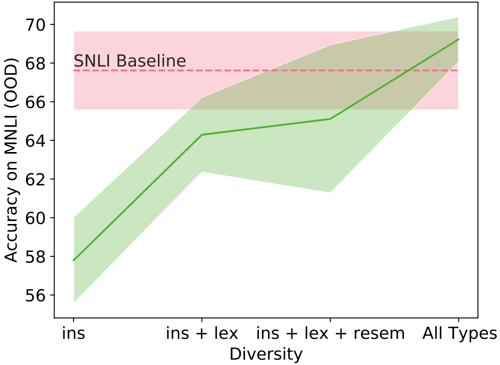
Regularization effect from CAD:

- Predictions should be invariant to unintervened features (book, movie, good)
- But, CAD may not cover all features that can be intervened to flip the label

Joshi, He. An Investigation of the (In)effectiveness of Counterfactually Augmented Data. ACL 2022.

Edit diversity vs performance

- **Train**: CAD (pairs) from SNLI [Kaushik et al., 2020]
- **OOD Test**: MNLI
- **Varying intervened features**: group edits by types, increase number of edit types while **controlling data size**



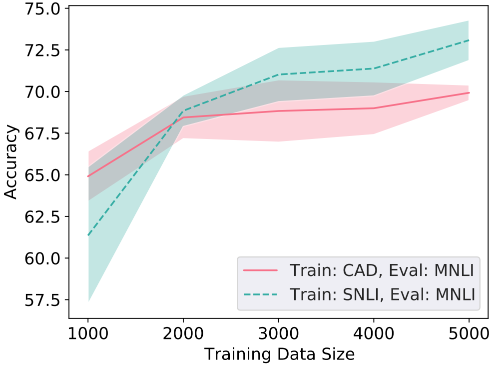
Diverse edits leads to better OOD performance

Joshi, He. An Investigation of the (In)effectiveness of Counterfactually Augmented Data. ACL 2022.

CAD data size vs performance

Does more CAD data lead to better performance?

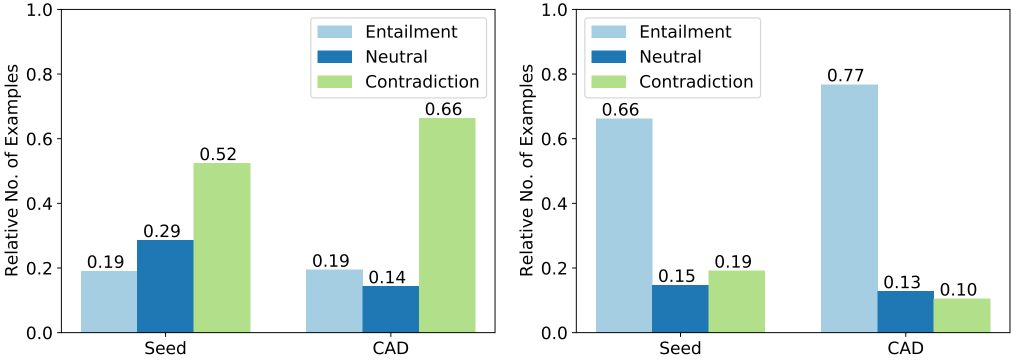
- **Train**: CAD (pairs) vs SNLI
- **OOD Test**: MNLI
- CAD is more effective in the low-data regime
- But plateaus quickly (suggesting limited edit diversity)



Joshi, He. An Investigation of the (In)effectiveness of Counterfactually Augmented Data. ACL 2022.

Does CAD reduce dataset bias?

Label distribution conditioned on spurious features:



(a) Negation word

(b) Word overlap $> 90\%$

Intervention without control may amplify existing spurious correlation

Joshi, He. An Investigation of the (In)effectiveness of Counterfactually Augmented Data. ACL 2022.

Revisit CAD

- The promise is that we don’t need to explicitly specify spurious features
- It turns out we still need a better understaning of them
- Revisit the assumption: edited spans are core features
- There are often many things we can edit to change the label

	I love dogs
con	I don’t love dogs
neu	You don’t love dogs
ent	I do love dogs
ent	I don’t fear dogs
ent	I don’t love dog-haters

Are all edited words non-spurious?

Some spurious features are irrelevant

The simple case: spurious features and core features are *disentangled*

- Changing the spurious feature doesn't affect prediction

Spielberg's new film is brilliant	positive
Zhang's new film is brilliant	positive

water $\rightarrow$ waterbird

land $\rightarrow$ waterbird

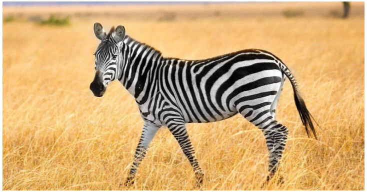
Some spurious features are necessary for prediction

The complex case: spurious features are *part of* the core features

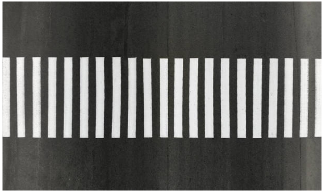
- The "spurious" feature is necessary but not sufficient for prediction

I love dogs	/	I don't love dogs	contradiction
I love dogs	/	I don't love cats	neutral

stripes $\rightarrow$ zebra

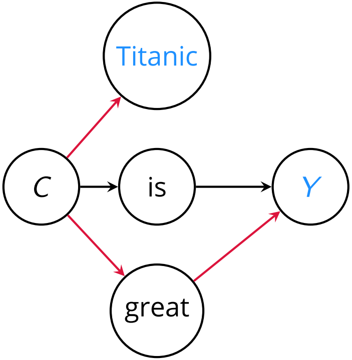


stripes $\rightarrow$ crosswalk



Two ways for a word to associate with the label

[Joshi et al., 2022]

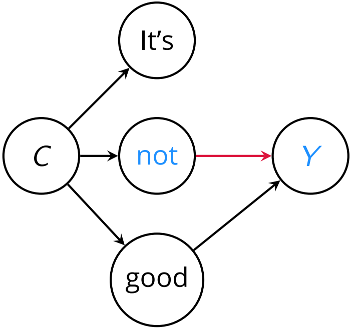


- C: the review writer
- Y: sentiment
- Titanic has no causal relation with Y
- But they may be correlated through C: famous movies tend to receive good reviews

The spurious feature is **irrelevant** to predicting the label.

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Two ways for a word to associate with the label



- C: the review writer
- Y: sentiment
- not causally affects Y

The spurious feature is **necessary** to predicting the label.

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Categorize spurious features

A feature is **spurious** if it is **not sufficient** for predicting the label.

But it may be necessary for prediction:

Irrelevant	**Necessary**
Titanic is great	I don't like the movie
Has no causal relation with the label Model should be invariant to them	Causally affect the label Model should be sensitive to them *More common in NLP (messier...)*

Next, lessons learned when dealing with necessary spurious features.

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Breaking the spurious correlation is not enough

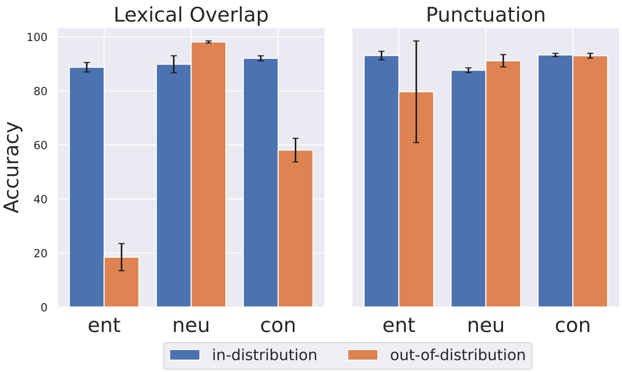
Does the model generalize well if the spurious feature is *independent* of the label on the training set?

- **Dataset**: MNLI
- **Model**: finetuned RoBERTa-Large
- **Spurious features**:
  - Punctuation: adding !! to the end of neutral examples
  - Overlap: lexical overlap and entailment [McCoy et al., 2019]
- **Train**: subsampled MNLI where spurious feature $\perp\!\!\!\perp$ label [Sagawa et al., 2020]
  - Uniform label distribution given high overlap
- **OOD Test**: examples without the spurious feature
  - Low overlap examples

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Breaking the spurious correlation is not enough

- **Train**: high overlap / has punctuation
- **ID Test**: high overlap / has punctuation
- **OOD Test**: low overlap / no punctuation



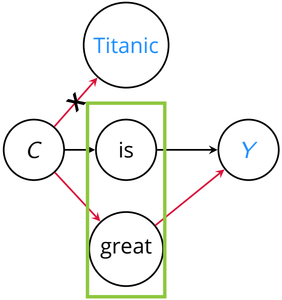
Performance is sensitive to necessary spurious feature even if they are independent to the label during training

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Effect of data balancing

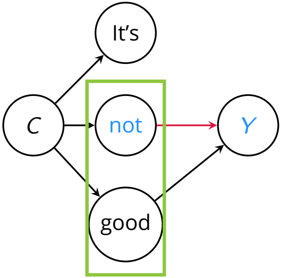
Irrelevant spurious features:

- Core features are the same with and without the spurious feature
- Breaking the correlation allows the model to learn the core features



Necessary spurious features:

- Core features vary with the spurious feature
- The model encounters new/rare features on OOD examples



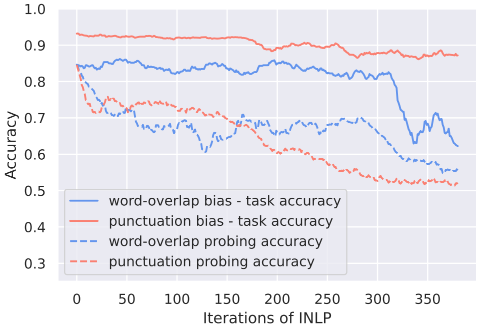
Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Removing spurious features from the representation may hurt performance

Do we want the representation to encode spurious features?

- **Train**: subsampled MNLI
- **OOD Test**: minority group (high overlap, non-entailment)
- **Debiasing**: iteratively projecting out the spurious feature [Ravfogel et al., 2020]
- **Probing accuracy**: is the feature removed?
- **Task accuracy**: is the debiased representation useful for NLI?

Figure: Overlap vs Punctuation



Removing necessary spurious features may also remove the dependent core features

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Evaluating robustness is tricky

How do we evaluate model robustness to necessary features like overlap?

Construct OOD examples with the spurious feature and different labels:

- Want entailed and non-entailed examples with high overlap
- HANS: hand-crafted
- MNLI-subsets: sampled from MNLI

Train on MNLI (biased), test on different OOD sets:

Models	HANS		MNLI subsets	
	Ent/Non-ent	Δ	Ent/Non-ent	Δ
BERT-base	99.2/12.9	86.3	96.4/82.5	13.9
RoBERTa-large	99.9/56.2	43.7	97.1/93.6	3.5

Diverging results on different challenge sets

Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Evaluating robustness to necessary spurious features

Goal: Test if the model is only relying on the spurious feature and ignoring the context

Approach: Construct challenge sets:

- Fixing the spurious feature, change the context to produce different labels

  P: The doctor believed the lawyer saw the officer.
  H: The doctor believed the lawyer

Potential problems:

- Likely to introduce new (non-spurious) features!
- Conflates performance drop due to latching on spurious features vs failing to use unseen features

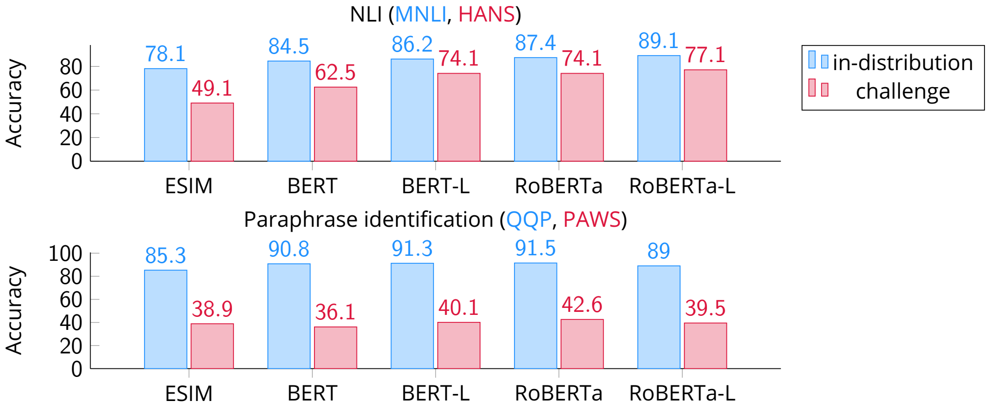
Joshi*, Xiang*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Summary so far

- The nice setting: we know the spurious feature, and it is irrelevant to prediction
  - Break the correlation (subsampling, reweighting, invariance etc.)
- The real setting: we don't know the spurious feature, there are many of them, and they may be necessary for prediction
  - Learn patterns on the long tail (data diversity, representation learning)
  - **Pre-training/scaling** could help

Pre-trained models appear to be more robust

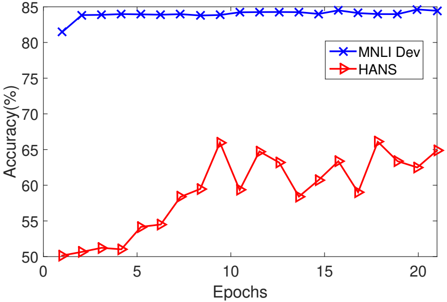
[Tu et al., 2020]



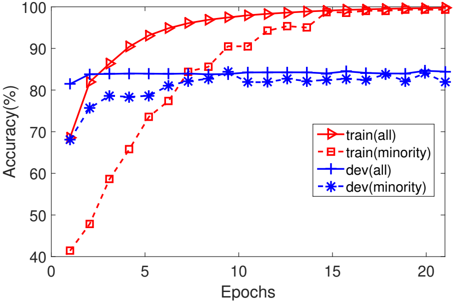
- Pre-training improves both in-distribution and challenge data performance
- Outperforming debiasing method with *longer fine-tuning*

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Minority examples take longer to learn



(a) Dev performance on MNLI and HANS



(b) Train/dev performance on MNLI

- Accuracy on HANS increases after MNLI plateaus
- Accuracy on minority examples (-∗-) correlates with accuracy on HANS (-Δ-)

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Counterexamples in the training data

Minority examples counter the spurious correlation and resemble the challenge data

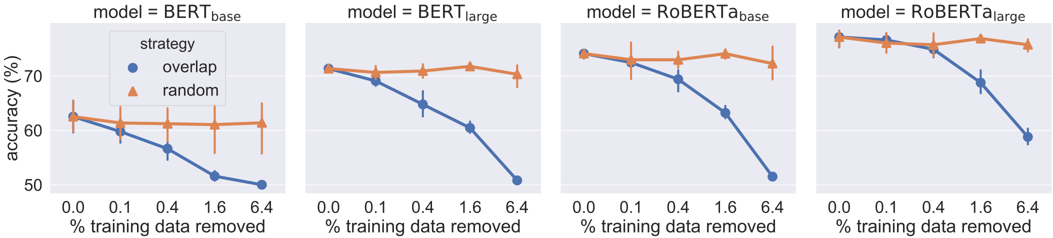
Natural language inference (HANS)		
P: The doctor mentioned the manager who ran. H: The doctor mentioned the manager.	overlap & entailment	
P: The actor was advised by the manager. H: The actor advised the manager.	overlap & non-entailment	**727 in MNLI**
Paraphrase Identification (PAWS [Zhang et al., 2019])		
S_1: Bangkok vs Shanghai? S_2: Shanghai vs Bangkok?	same BoW & paraphrase	
S_1: Are all dogs smart or can some be dumb? S_2: Are all dogs dumb or can some be smart?	same BoW & non-paraphrase	**247 in QQP**

Do pre-trained models generalize better from the minority examples?

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Ablation: removing minority examples

OOD Accuracy when removing random vs minority examples



- Pre-training improves robustness to group imbalance
- But they cannot generalize to challenge data without minority examples

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Improve generalization by multitasking

Idea: Improve generalization from minority examples by transfering knowledge from related tasks

Multitasking learning setup

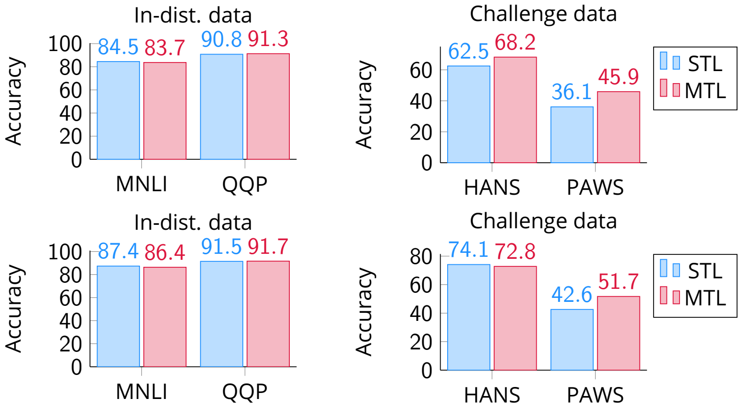
- Model: shared BERT encoder + linear task-specific classifier
- Auxiliary data:
    - Textual entailment: MNLI + SNLI, QQP, PAWS
    - Paraphrase identification: QQP + SNLI, MNLI, HANS

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Results

BERT-base

RoBERTa-base



- MTL improves robust accuracy without hurting indistribution performance
- MTL improves robustness on top of pre-training

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

How does MTL help?

Removing examples from target vs auxiliary tasks

Method	In-dist. (QQP)	Challenge (PAWS)
STL (QQP)	90.8	36.1
MTL (QQP+MNLI,SNLI,HANS)	91.3	45.9
remove random examples from MNLI	$+0.1$	-0.9
remove random examples from QQP	-0.0	-1.6
remove minority examples from MNLI	$+0.0$	-1.6
remove minority examples from QQP	$+0.0$	-7.7

- Remove *minority examples from target tasks* hurt OOD generalization

Support for examples countering spurious correlations is important

Tu, Lalwani, Gella, He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. TACL 2020.

Robustness in the era of large language models

Greg Durrett
@gregd_nlp

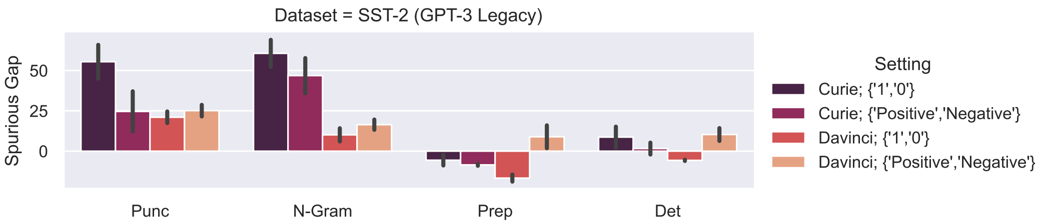
10:53 PM · Nov 3, 2022 · Twitter Web App

- Do we still need supervised learning?
- What is OOD wrt to the pretraining data?
- What's the inductive bias of LM pretraining?

Is in-context learning robust to biases in the demonstration?

[Si et al., 2022]

- **Data**: semi-synthesized spurious features (punctuation, n-grams etc.)
- **Prompt**: spurious feature is perfectly predictable of the label
- **Metric** ↓: gap between bias-support and bias-countering examples

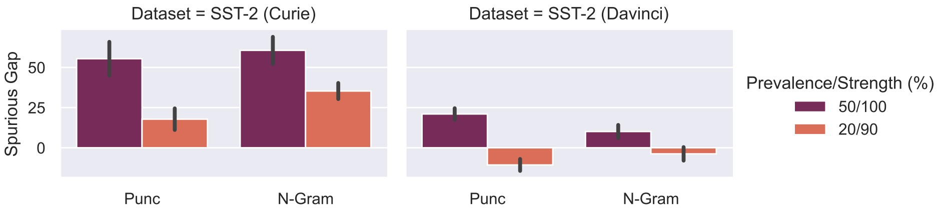


- GPT-3 suffers from (extreme) spurious correlation in the prompt
- But it can be alleviated with verbalized labels

Under submission

Is in-context learning robust to biases in the demonstration?

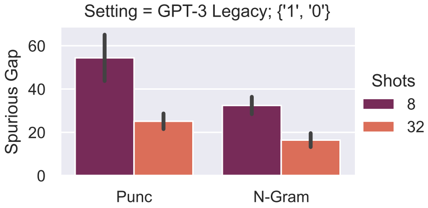
Reduced gap under weaker spurious correlation



Diverse demonstration examples are helpful

Is in-context learning robust to biases in the demonstration?

Reduced gap given more in-context examples



Behavior of in-context learning is quite different from supervised learning!

Summary

Takeaways:

- Tackling all sorts of spurious features in NLP tasks is a hard battle
- Pretraining and scaling have consistently improved model robustness so far

Open questions:

- What is OOD wrt to pretraining (rare events, human biases)?
- How does prompting or in-context learning work?
- How does human interaction / feedback help?

Collaborators


Garima Lalwani

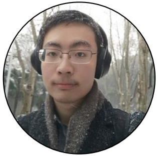

Lifu Tu


Spandana Gella


Sheng Zha


Haohan Wang

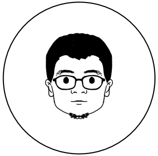

Nitish Joshi


Xiang Pan


Shi Feng


Danqi Chen


Dan Friedman


Chenglei Si

Thank you!