Hard battles and easy victories in robustifying NLP models

He He

NEW YORK UNIVERSITY

RobustSeq Workshop

December 2, 2022

A motivation example: 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...

A motivation example: 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

A motivation example: 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

A motivation example: 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

A motivation example: 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

Distribution shift due to perturbation of the spurious feature:

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

A motivation example: 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

Distribution shift due to perturbation of the spurious feature:

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

> Models may not generalize as expected in deployment domains

Real examples

Biases in NLP datasets:

- **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 on OOD data where the simple heuristic fails

Challenges

What assumption should we make about spurious features (in language data)?

- $p(y \mid x)$ should be invariant to perturbations of the spurious feature
- The label should be independent of the spurious feature
- The learned representation should not contain information about the spurious feature

Recipe: assumptions $\rightarrow$ objective $\rightarrow$ optimization

Common assumptions may not apply to certain spurious featurs in NLP data

Some spurious features are irrelevant

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

- Changing the spurious feature doesn't affect label

Some spurious features are irrelevant

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

- Changing the spurious feature doesn't affect label

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

Some spurious features are irrelevant

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

- Changing the spurious feature doesn't affect label

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

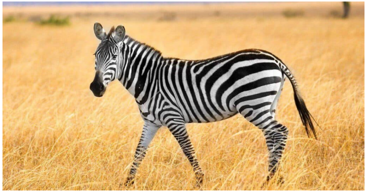
Some spurious features are necessary for prediction

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

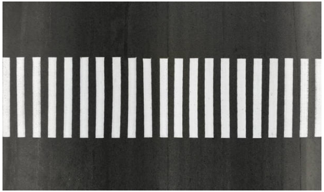
- The "spurious" feature is necessary but not sufficient for prediction

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

stripes $\rightarrow$ zebra

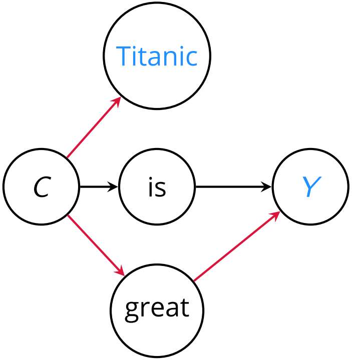


stripes $\rightarrow$ crosswalk



Two ways for a word to associate with the label

[Joshi et al., 2022]

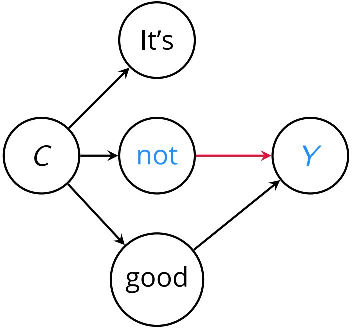


- C: the review writer
- Y: sentiment
- Titanic has no causal relation with Y
- But they may be correlated through C: famous movies tend to receive good reviews

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

Joshi*, Pan*, 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*, Pan*, 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

Joshi*, Pan*, 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...)*

How well do existing methods work on different types of spurious features?

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

Setup

- **Dataset**: MNLI
- **Model**: finetuned RoBERTa-Large

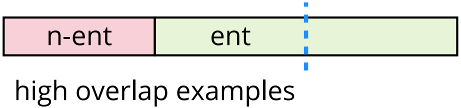
Setup

- **Dataset**: MNLI
- **Model**: finetuned RoBERTa-Large
- **Spurious features**:
  - **Irrelevant**: adding !! to the end of neutral
    The kids are playing football.
    The kids are shouting!! (neutral)
  - **Necessary**: lexical overlap and entailment [McCoy et al., 2019]
    The woman is selling sweets to the kids.
    The woman is selling sweets. (entailment)
- **Methods**:
  - Data balancing through subsampling
  - Representation debiasing

Data balancing

Assumption: the label should be independent of the spurious feature

Method: subsample the data s.t. label is independent of the feature [Sagawa et al., 2020; *i.a.*]

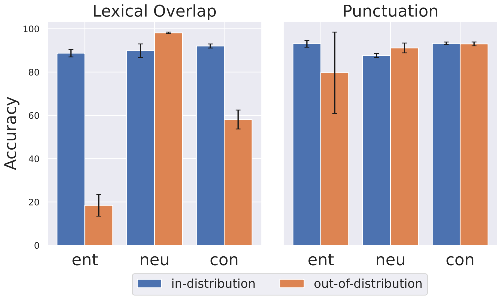


Does the model generalize well if the spurious feature is *independent* of the label?

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

Breaking the spurious correlation is not enough

ID	**OOD**
high overlap	low overlap
has punctuation	no punctuation
$y \perp\!\!\!\perp x_{\text{spurious}}$	



Takeaway: feature-label independence leads to

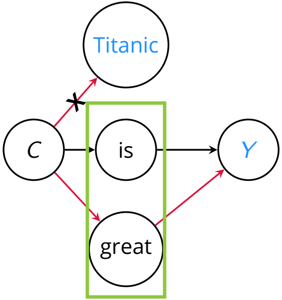
- good OOD performance for **irrelevant features**
- but we still see large ID-OOD performance gap for **necessary features**

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

Effect of data balancing

Irrelevant spurious features:

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

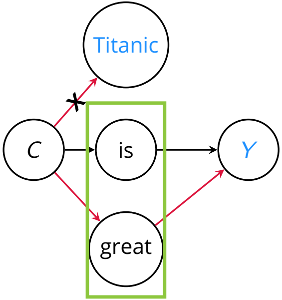


Joshi*, Pan*, He. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. EMNLP 2022.

Effect of data balancing

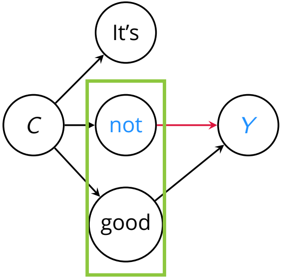
Irrelevant spurious features:

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



Necessary spurious features:

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



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

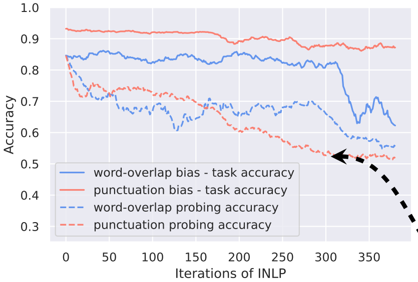
Removing necessary features from the representation hurt performance

What if we remove spurious features from the learned representation?

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

Removing necessary features from the representation hurt performance

What if we remove spurious features from the learned representation?

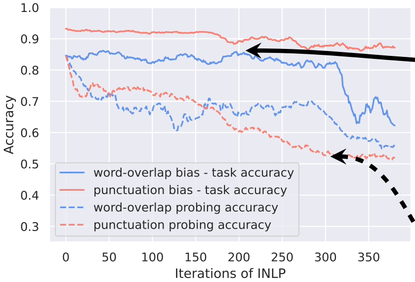


Probing accuracy: lower $\rightarrow$ the feature gets removed

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

Removing necessary features from the representation hurt performance

What if we remove spurious features from the learned representation?



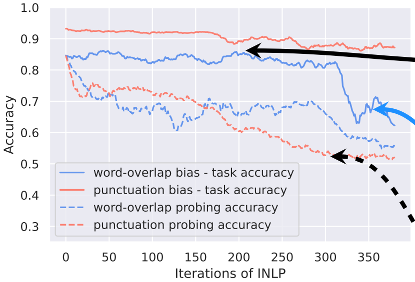
Task accuracy: higher → the representation is useful for NLI

Probing accuracy: lower → the feature gets removed

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

Removing necessary features from the representation hurt performance

What if we remove spurious features from the learned representation?



Task accuracy: higher → the representation is useful for NLI

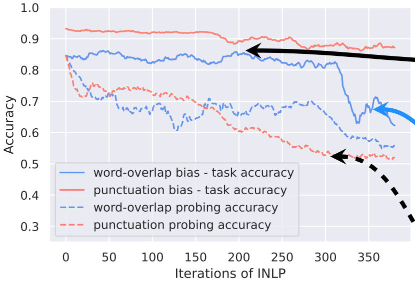
Removal of necessary features degrades task performance

Probing accuracy: lower → the feature gets removed

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

Removing necessary features from the representation hurt performance

What if we remove spurious features from the learned representation?



Task accuracy: higher $\rightarrow$ the representation is useful for NLI

Removal of necessary features degrades task performance

Probing accuracy: lower $\rightarrow$ the feature gets removed

Takeaway: removing spurious features

- does not affect task accuracy for irrelevant features
- but degrades task accuracy for necessary features

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

Summary so far

What assumption should we make about spurious features (in language data)?

- The nice setting: we know the spurious feature, and it is irrelevant to prediction
  - Break the feature-label correlation (subsampling, reweighting, invariance etc.)

Summary so far

What assumption should we make about spurious features (in language data)?

- The nice setting: we know the spurious feature, and it is irrelevant to prediction
  - Break the feature-label 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
  - Existing independence or invariance assumptions do not apply
  - Learn patterns on the long tail

Next, what's the role of pretraining in robust language understanding?

Robustness in the era of large language models

Greg Durrett
@gregd_nlp

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

Supervised learning → zero-shot / in-context learning

- What's the underlying distribution shift?
- What's the inductive bias of pretraining and prompting?

Our study: Is in-context learning robust to spurious correlations in the demonstration?

Setup

- **Prompt** with spurious correlation:
  Input: Riveting film of the highest calibre ! Label: +1
  Input: Thank God I didn't go to the cinema. Label: -1
  Input: Definitely worth the watch ! Label: +1
  Input: Boring as hell. Label: -1

Setup

- **Prompt** with spurious correlation:
  Input: Riveting film of the highest calibre ! Label: +1
  Input: Thank God I didn't go to the cinema. Label: -1
  Input: Definitely worth the watch ! Label: +1
  Input: Boring as hell. Label: -1
- **Features**: semi-synthesized spurious features (punctuation, n-grams etc.)

Setup

- **Prompt** with spurious correlation:
  Input: Riveting film of the highest calibre ! Label: +1
  Input: Thank God I didn't go to the cinema. Label: -1
  Input: Definitely worth the watch ! Label: +1
  Input: Boring as hell. Label: -1
- **Features**: semi-synthesized spurious features (punctuation, n-grams etc.)
- **Metric** $\downarrow$: gap between bias-support and bias-countering examples on test set
  Input: A story told perfectly! Label:
  Input: Complete waste of time! Label:

Setup

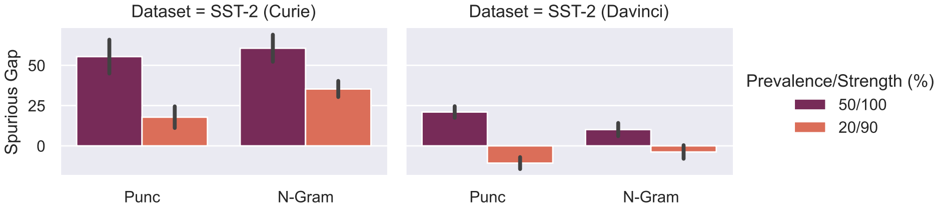
- **Prompt** with spurious correlation:
  Input: Riveting film of the highest calibre ! Label: +1
  Input: Thank God I didn't go to the cinema. Label: -1
  Input: Definitely worth the watch ! Label: +1
  Input: Boring as hell. Label: -1
- **Features**: semi-synthesized spurious features (punctuation, n-grams etc.)
- **Metric** $\downarrow$: gap between bias-support and bias-countering examples on test set
  Input: A story told perfectly! Label:
  Input: Complete waste of time! Label:
- **Models**: Curie (13B), Davinci (175B, original GPT-3)

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

[Si et al., 2022]

Controlling the strength of prompt bias:

- Prevalence: % examples with the spurious feature
- Strength: $p(y \mid x_{\text{spurious}})$



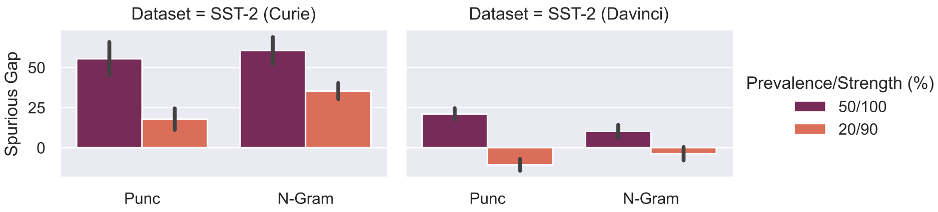
Preliminary work

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

[Si et al., 2022]

Controlling the strength of prompt bias:

- Prevalence: % examples with the spurious feature
- Strength: $p(y \mid x_{\text{spurious}})$



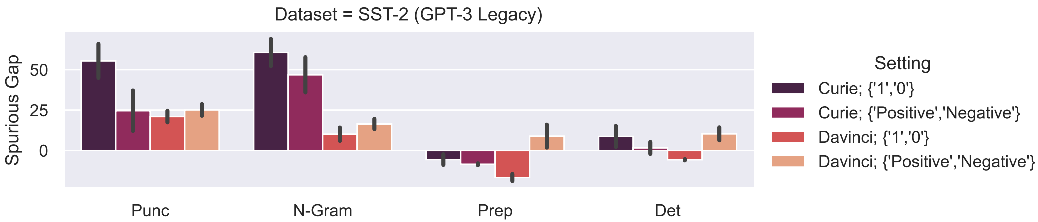
- GPT-3 suffers from **extreme bias** in the prompt, but less so under **weaker bias**.
- The larger model seems to infer the intended task even under extreme bias

Preliminary work

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

What's the effect of prompt engineering?

Input: Riveting film of the highest calibre ! Label: positive
Input: Thank God I didn't go to the cinema. Label: negative

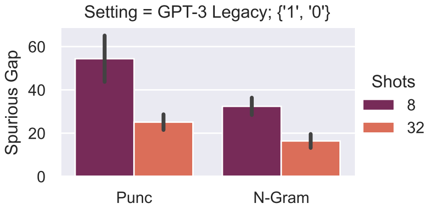


- Using verbalized labels helps the model learn the target task to some extent

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

What if we give it more in-context examples?

- Still with uninformative labels (0/1) and extreme bias (100% predictive)



- More (biased) examples help the model infer the intended task
- which is different from supervised learning

Summary so far

- Scaling up LLMs improve robustness without explicit assumptions about the spurious features
- They are still susceptible to spurious correlations in the demo examples
- But proper prompt design can mitigate the problem by informing the model of the intended task

Parting remarks

Takeaways:

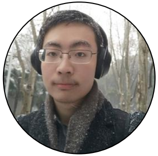
- Tackling all sorts of spurious features in NLP tasks is a hard battle
- Pretraining and scaling have consistently improved model robustness so far

Open questions:

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

Collaborators

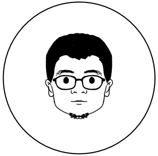

Nitish Joshi


Xiang Pan


Shi Feng


Danqi Chen


Dan Friedman


Chenglei Si

Thank you!