Finding Dataset Shortcuts with Grammar Induction

Dan Friedman, Alexander Wettig, and Danqi Chen

Department of Computer Science, Princeton University

Introduction

Many NLP datasets contain **shortcuts**, simple decision rules that achieve **high in-domain accuracy** but **fail to generalize** to the intended test distribution. **If we can identify shortcuts, we can mitigate them** by collecting more training data or using robust optimization methods.

Prior work Earlier work on finding shortcuts relied on **intuition** and **manually designed probes**. Prior work on automatically identifying shortcuts has focused on **simple features that can be explicitly enumerated**, like unigrams [6, 1], or **qualitative interpretability methods**, like saliency maps [2, 5].

Proposal: Induce **dataset-specific grammars** to **formally characterize patterns** in sentence and sentence-pair classification datasets.

Method

Given text classification dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, where $y \in \mathcal{Y}$ is a categorical label and $x \in \mathcal{X}$ is either a sentence or a pair of sentences, $x = (x^a, x^b)$.

1. Grammar induction

Induce a grammar for (unlabeled) training instances $x_1, \ldots, x_N$ and get maximum likelihood trees $t_1, \ldots, t_N$.

- Sentence datasets: **context-free grammar** (CFG)
- Sentence-pair datasets: **synchronous CFG** [7]

Given a grammar $\mathcal{G}$, find parameters θ^*:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log \sum_{t:\text{yield}(t)=x_i} p(t \mid \mathcal{G}, \theta).$$

2. Finding features

Define a boolean indicator feature $Z_s \in \{0, 1\}$ for each complete subtree s and calculate the mutual information:

$$I(Z_s; Y) = \sum_{z_s \in \{0,1\}} \sum_{y \in \mathcal{Y}} p(y, z_s) \log \frac{p(y, z_s)}{p(y)p(z_s)}.$$

Group subtrees according to root label and majority class.

Finding Shortcuts

IMDb movie review dataset

Root	Desc.	Patterns	% Maj.
⑤	Negative actors	ed wood, steven seagal, uwe boll	95.5
㉙	Negative ratings	4 / 10, 3 / 10, 1 / 10, 2 / 10	96.8
⑧	Negative durations	30 minutes, 10 minutes, five minutes	76.7
⑤	Positive actors	walter matthau, jon voight, james stewart	88.6
㉙	Positive ratings	10 / 10, 8 / 10, 7 / 10	98.8
⑧	Positive durations	many years	69.1

Subjectivity dataset

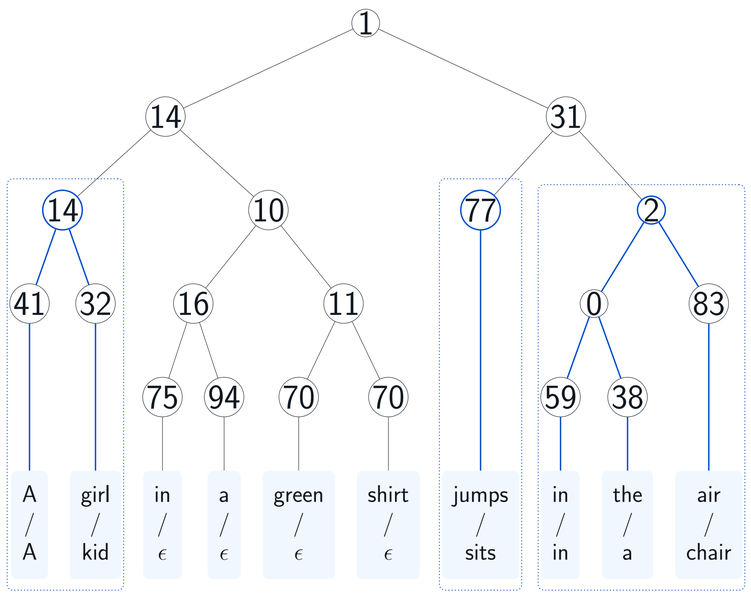
Root	Desc.	Patterns	% Maj.
㉗	Review NP	a movie, the film, the movie, this movie	86.3
③	Review VB	comes off, ' s hard, makes up, ' d expect	87.4
㉗	Summary NP	his life, his wife, his father, his mother	80.1
③	Summary VB	finds himself, finds out, falls in love	85.5

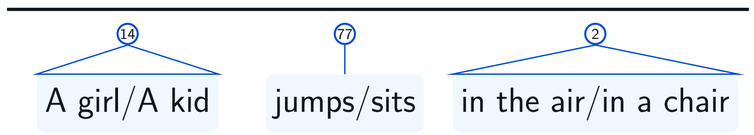
SNLI

Root	Desc.	Patterns	% Maj.
㊹	Copy verb	walking/walking, running/running	68.4
⑭	Subj. hypernym	a woman/a person, a man/a human	45.7
④	Expletive	a /there is, ϵ/there are, two /there are	63.0
⑭	Subj. antonym	a man/a woman, a boy/a girl, a dog/a cat	82.5
㊽	Verb antonym	standing/sitting, walking/sitting	92.6
㊺	Adj. antonym	black/white, red/blue, ϵ/empty, ϵ/living	76.9
㉟	Add object	ϵ/[UNK], ϵ/work, ϵ/get, ϵ/friends, ϵ/park	59.8
㊺	Add adj.	ϵ/tall, ϵ/sad, ϵ/[UNK], ϵ/new, ϵ/big	72.1
⑰	Add PP phrase	ϵ/to work, ϵ/to get, ϵ/to buy, ϵ/the park	71.4

Quora Question Pairs

Root	Desc.	Patterns	% Maj.
㊿	Additions	ϵ/[UNK], ϵ/in, ϵ/a, ϵ/-, ϵ/for	60.7
㊾	Deletions	[UNK]/ϵ, in/ϵ, a/ϵ, like/ϵ, of/ϵ	61.6
㊾	Change Q	why/how, why/what, how/why, why/can	70.8
⑭	How-to	how can/how can, how do/how can	66.2
㉕	Topics	new year/new year, world war/world war	82.9
㊾	Same Q	how/how, why/why, when/when	60.3

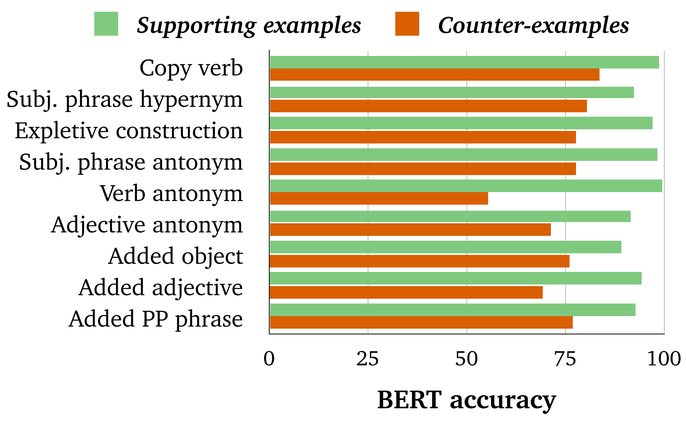




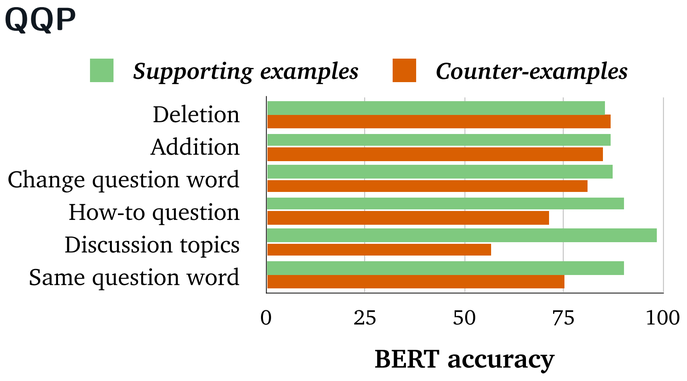
Top subtrees for *entailment*		
A man/A person A man/A human	walking/walking walk/walking	in the grass/outside down the street/outside
Top subtrees for *contradiction*		
A man/A woman A woman/A man	ϵ/sitting stand/sit	at night/during the day in the air/down
Top subtrees for *neutral*		
A man/A tall human A man/An old man	ϵ/competes running/running	down the street/home in the sand/on the beach

Do Models Use these Shortcuts?

SNLI



QQP

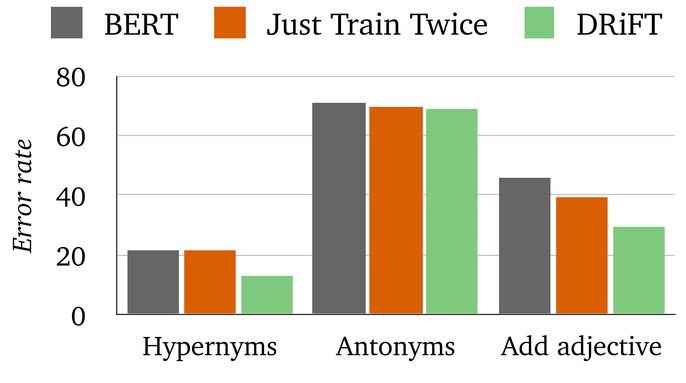


Generating counter-examples

SNLI

Edit	# Sets	Error
Hypernyms A man is smoking at sunset. A man +person smoking a cigarette.	389	$21.8_{\pm 0.8}$
Antonyms Two black dogs splash around on the beach. The dogs are playing with a +white ball.	281	$71.1_{\pm 3.8}$
Add adjective A man taking photos of nature. A +sad man is taking photos of a wedding.	1,470	$45.6_{\pm 8.4}$

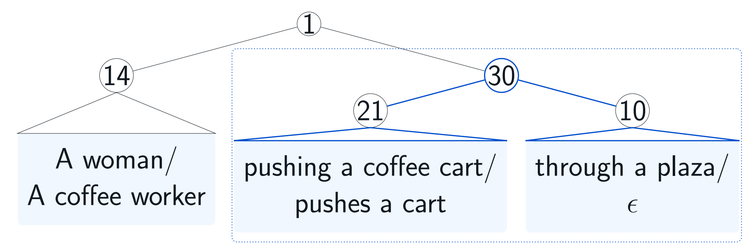
Mitigating Shortcut Learning



Robust optimization methods on SNLI. Just Train Twice [4] assumes **shortcuts** are **unknown**. DRiFT [3] incorporates information about the **known shortcuts** we discovered.

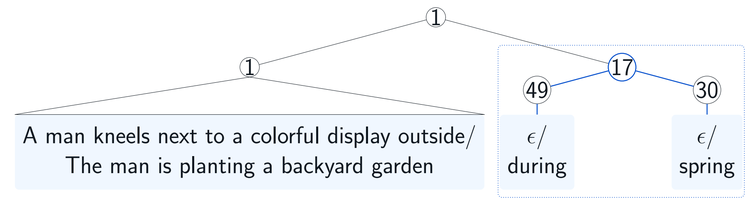
Comparing grammar features and saliency maps

A woman pushing a coffee cart through a plaza. A coffee worker pushes a cart.



Label: Neutral, *BERT prediction:* Entailment

A man kneels next to a colorful display outside. The man is planting a backyard garden during spring .



Label: Contradiction, *BERT prediction:* Neutral

Conclusions

An approach for automatically **finding dataset shortcuts** by inducing **dataset-specific grammars**. The grammar can be used to:

- **Discover** interesting shortcut features
- **Diagnose** classifier errors
- **Mitigate** shortcut learning

References

[1] Matt Gardner et al. "Competency Problems: On Finding and Removing Artifacts in Language Data". In: *EMNLP*. 2021.

[2] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. "Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions". In: *ACL*. 2020.

[3] He He, Sheng Zha, and Haohan Wang. "Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual". In: *DeepLo*. 2019.

[4] Evan Z Liu et al. "Just Train Twice: Improving group robustness without training group information". In: *ICML*. 2021.

[5] Pouya Pezeshkpour et al. "Combining Feature and Instance Attribution to Detect Artifacts". In: *ACL Findings*. 2022.

[6] Zhao Wang and Aron Culotta. "Identifying Spurious Correlations for Robust Text Classification". In: *EMNLP Findings*. 2020.

[7] Dekai Wu. "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora". In: *Computational Linguistics* (1997).