Unlearn Dataset Bias for Natural Language Inference by Fitting the Residual

aws
NYU
Carnegie Mellon University

He He, Sheng Zha, Haohan Wang

https://github.com/hhexiy/debiased

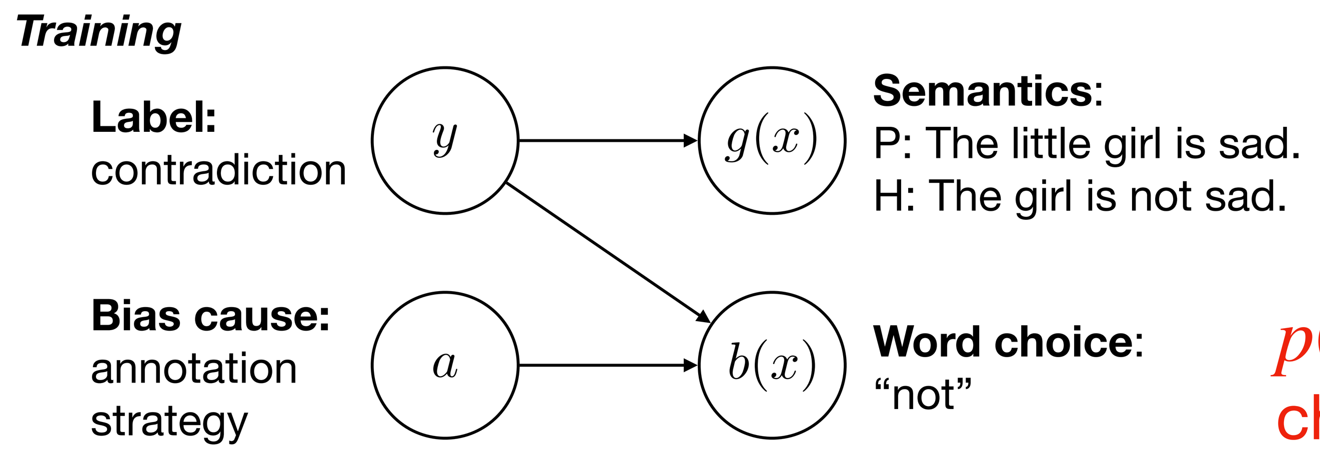
Overview

Dataset bias: spurious association between input and output

Problem: brittle model under slight distribution shift

Goal: Guard against *known* dataset bias

Key idea: don't learn from examples with strong (known) bias



$p(y, b(x))$ may change at test time

Debiasing by Residual Fitting (DRiFt)

1. Learn a biased classifier using features that should not be associated with the label

$$\theta^* = \arg\min_{\theta} \mathbb{E}_P[L(f_s(I(x); \theta), y]$$

2. Learn the debiased classifier by fitting the residuals

$$\min_{\phi} \mathbb{E}_P[L(\underbrace{f_s(I(x); \theta^*)}_{\text{Fixed}} + \underbrace{f_d(x; \phi)}_{\text{Learns what cannot be explained by } I(x)}, y)]$$

Cross-entropy loss

Biased classifer: $p_s(y \mid x) \propto \exp(f_s^y)$ ← Learned by MLE using $I(x)$

Debiased classifer: $p_d(y \mid x) \propto \exp(f_d^y)$

Objective:

$$J(\phi) = \sum_{(x,y)} \log p(y \mid x) \qquad p(y \mid x) \propto \exp(f_s^y + f_d^y) \propto p_s p_d$$

$$= C + \sum_{(x,y)} \Big[\underbrace{\log p_d(y \mid x)}_{J_{\text{MLE}}} - \underbrace{\log \sum_k p_s^*(k \mid x) p_d(k \mid x)}_{\text{Regularizer } R(x)} \Big]$$

Predictive biased classifier:

$$p_s^* \to 1 \;\Rightarrow\; \nabla_\phi R(x) = -\nabla_\phi \log p_d(y \mid x)$$

Cancels MLE gradient

Uninformative biased classifier:

$$p_s = 1/K \;\Rightarrow\; p(y \mid x) = p_d(y \mid x)$$

Reduced to MLE

Experiments

Biased models:
- Hypothesis-only
- Handcrafted
- CBOW

Debiased models:
- Finetuned BERT [Devlin+ 18]
- DA [Parikh+ 16]: ~BoW
- ESIM [Chen+ 17]: ~DA + LSTM

Learning algorithms:
- MLE
- DRiFt
- Rm

Remove biased examples

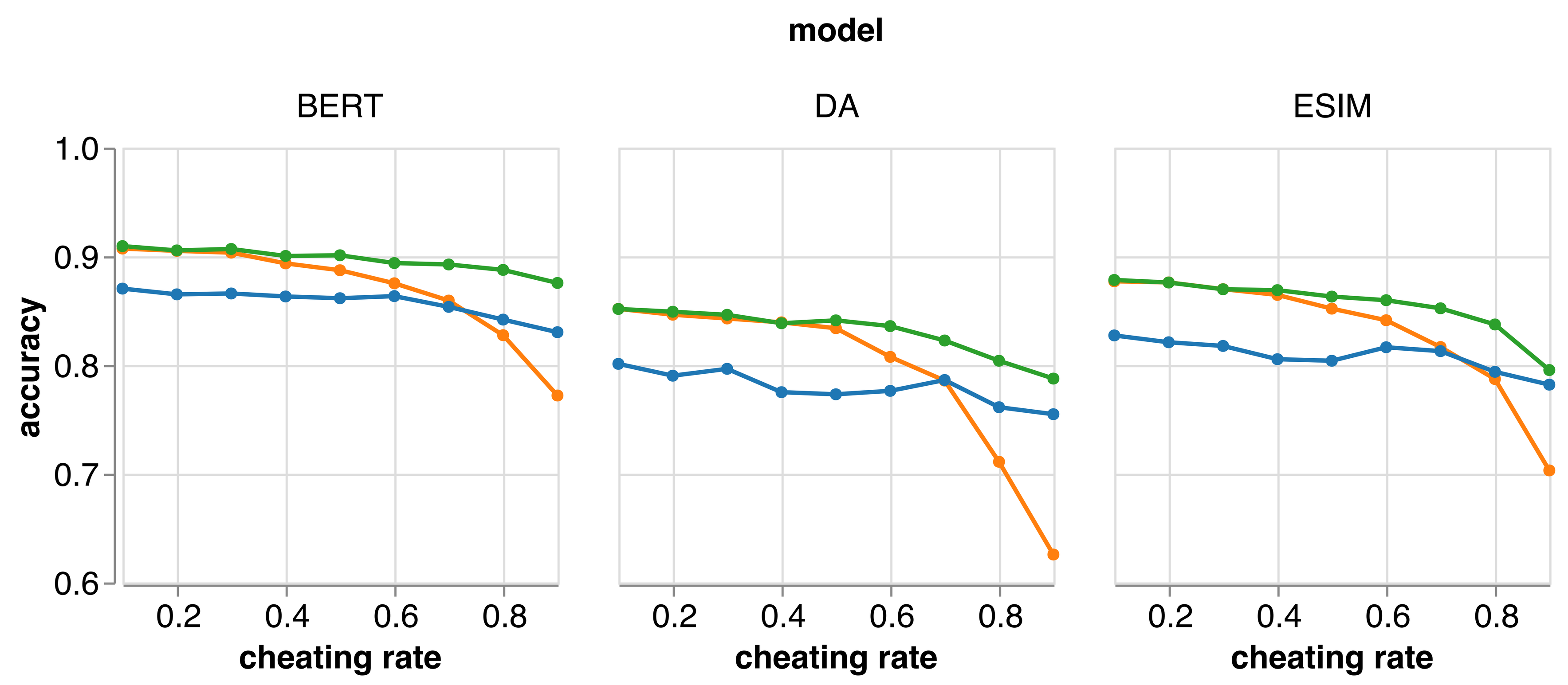
Synthetic bias:

Groundtruth with p_{cheat} at train and random at test

P: I love dogs.

H: `[contradiction]` I don't love dogs.

Results on synthetic bias



MLE: baseline

DRiFt-hypo: hypothesis-only biased classifier

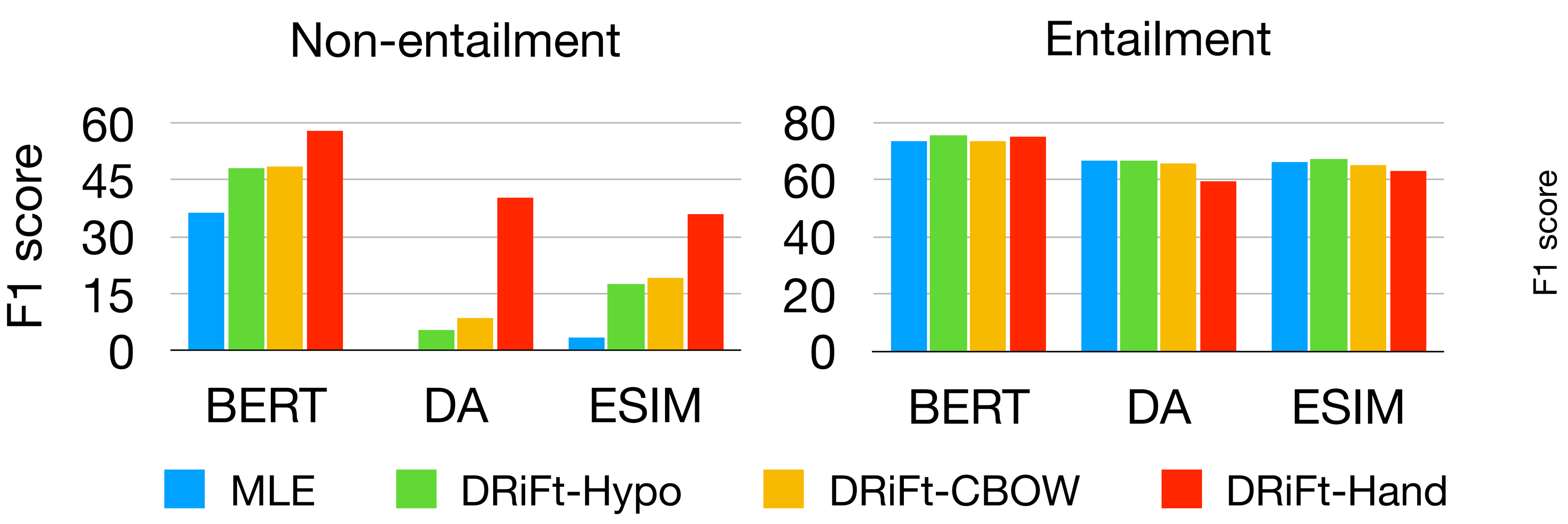
Rm-cheat: remove cheatable examples (oracle)

- Bias needs to be presented on a majority of examples
- BERT is more robust than non-pretrained models
- Importance of accurate prior knowledge on biased features (compare DRiFt-hypo and Rm-cheat)

NLI Results

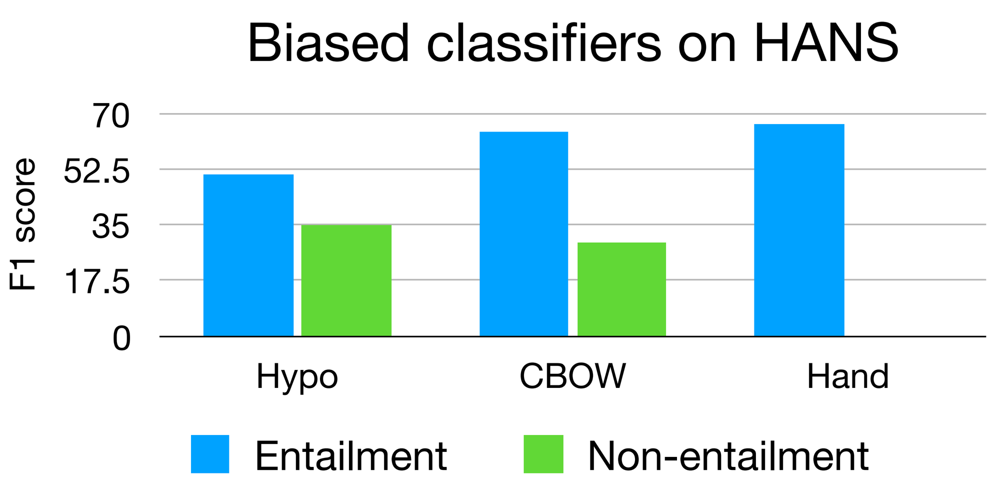
HANS [McCoy+ 19]: exploit word overlap bias, e.g., The doctor was paid by the actor ⇏ The doctor paid the actor

Train on MNLI and test on HANS



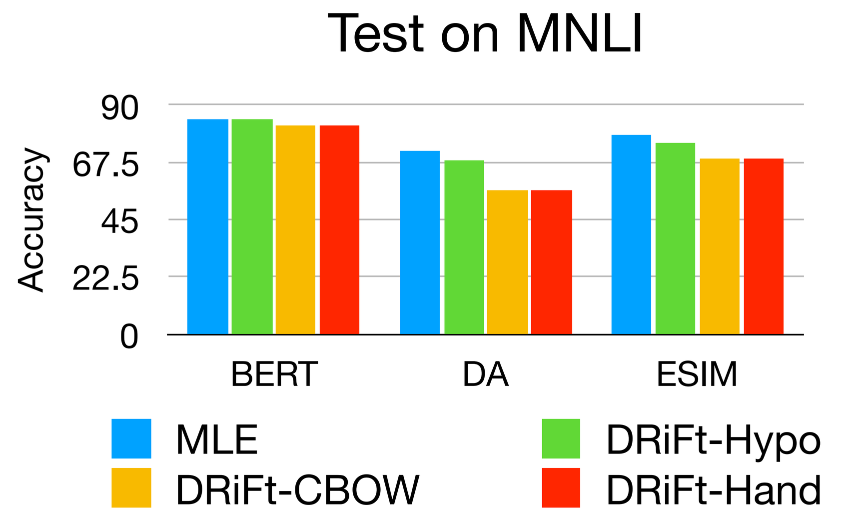
DRiFt improves performance on challenge data (non-entailment)

Why is "Hand" more effective?



"Hand" better captures bias exploited by HANS

In-distribution accuracy



Accuracy-robustness trade-off