Uncertainty Calibration for Ensemble-Based Debiasing Methods

Ruibin Xiong*, Yimeng Chen*, Liang Pang, Xueqi Cheng, Zhiming Ma,
Yanyan Lan

Contents

Contents

- Background
  - Why debiasing?
  - Ensemble-based Debiasing Methods
- Motivation
- Method and Experiments
- Conclusion and Future Work

The Impressive Performance of ML Models

Model	Accuracy	
	Train	Test
Human Performance (Estimated)	97.2%	87.7%
DR-BiLSTM (Single)	94.1%	**88.5%**
DR-BiLSTM (Single)+Process	94.1%	**88.9%**
DR-BiLSTM (Ensemble)	94.8%	**89.3%**
DR-BiLSTM (Ensem.)+Process	94.8%	**89.6%**

Natural Language Inference

Even outperform human on the SNLI dataset

Identify signs of diabetic retinopathy (糖尿病视网膜病变)

> 90% accuracy (comparable with experts),
< 10 minutes[1] v.s. 1 month (human)
(By Google Health)

Picture source: https://www.wallingfordeyes.com/eye-health/eye-diseases/107-diabetic-retinopathy

1 A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, CHI'20, April 25–30, 2020

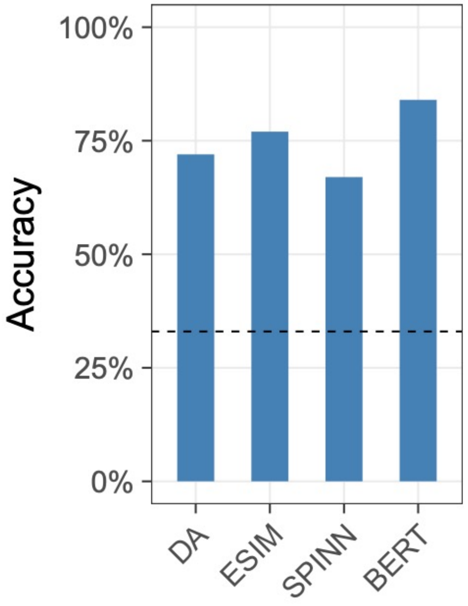
Failures in Real Applications

- When ML models went out of the training environment, significant drop in performance occurs

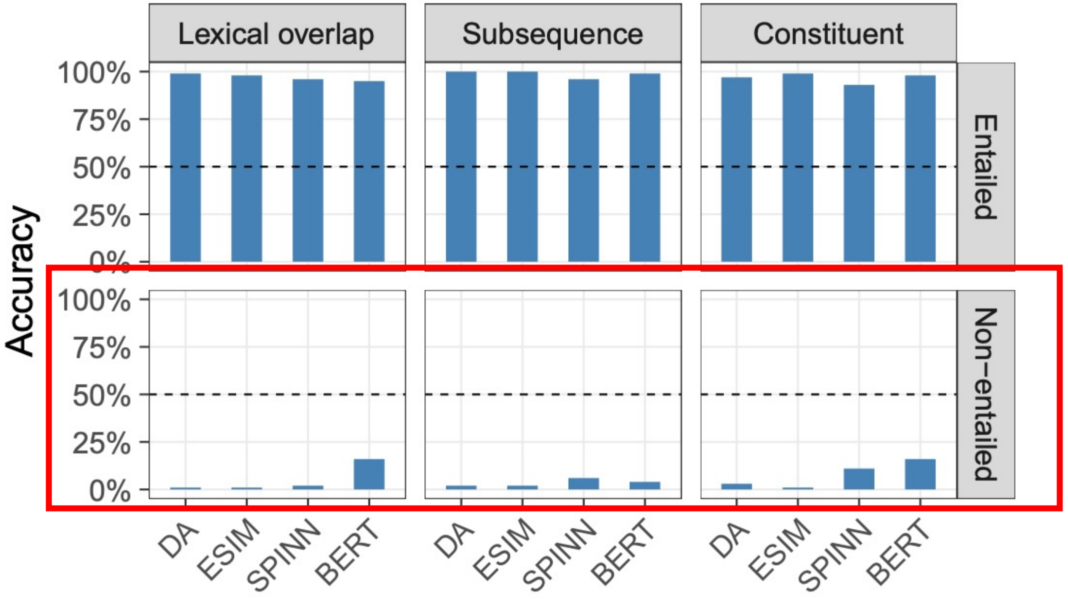
- New evidence, from Andrew Ng.
  https://spectrum.ieee.org/andrew-ng-xrays-the-ai-hype
  "It turns out [that when] you take that same model, that same AI system, to an older hospital down the street, with an older machine, and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly. In contrast, any human radiologist can walk down the street to the older hospital and do just fine. ... "
  "All of AI, not just healthcare, has a proof-of-concept-to-production gap," he says. "The full cycle of a machine learning project is not just modeling. It is finding the right data, deploying it, monitoring it, feeding data back [into the model], showing safety—doing all the things that need to be done [for a model] to be deployed. [That goes] beyond doing well on the test set, which fortunately or unfortunately is what we in machine learning are great at."

Failures in Real Applications

- When ML models went out of the training environment, significant drop in performance occurs



Performance on MNLI



Performance on HANS

Related Works: McCoy et al. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. ACL 2019

Failures in Real Applications

- When ML models went out of the training environment, significant drop in performance occurs

Identify signs of diabetic retinopathy (糖尿病视网膜病变)

> 50% images in poor lighting conditions were rejected, even no pattern of disease

Related work: A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, CHI'20, April 25–30, 2020

Failures in Real Applications

- When ML models went out of the training environment, significant drop in performance occurs

Detect COVID-19 by CXR and CT

None of 62 machine learning models is of potential clinical use

"Any machine learning algorithm is only as good as the data it's trained on."

Related work: Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nature Machine Intelligence, 2021

Challenges of ML in the Application

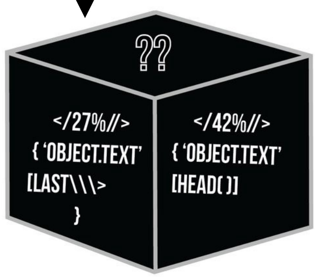
Generalizability

> 50% images in poor lighting conditions were rejected

Poor performance in OOD (out-of-distribution)

Interpretability

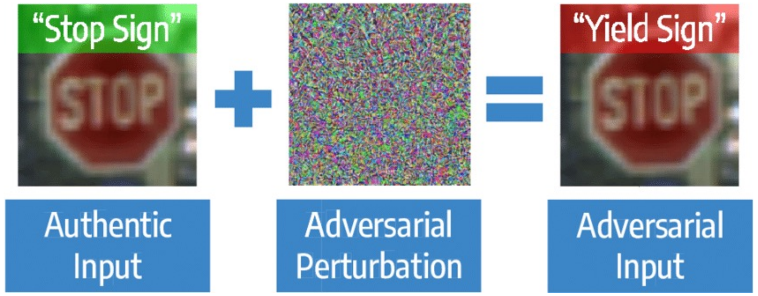
Input



output

Deep NN remains a black box to human

Robustness



Sensitive to the noise and easy to attack

Related Works: A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, CHI'20, April 25–30, 2020;
Pics source: www.topbots.com/interpretable-machine-learning/; Detecting Adversarial Inputs by Looking in the black box, 2019

Decades Efforts on These Problems

- Methods combating these problems including
  - Transfer learning
  - Data augmentation
  - Robust training
  - Causal machine learning
  - **Debiasing**
  - ....

What is Debiasing?

The Dependence on Spurious Correlations (Dataset Bias)

- Debiasing: to mitigate model's reliance on **Dataset bias**

Bias features

Heuristic	Supporting Cases	Contradicting Cases
Lexical overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

"Entailment"
"Neutral"
"Contradiction"

Training set

P： **The** little **boy is happy**.
H： **The boy is happy**.

→ **high** word-overlap → entailment ✓

The Dependence on Spurious Correlations (Dataset Bias)

- Debiasing: to mitigate model's reliance on **Dataset bias**

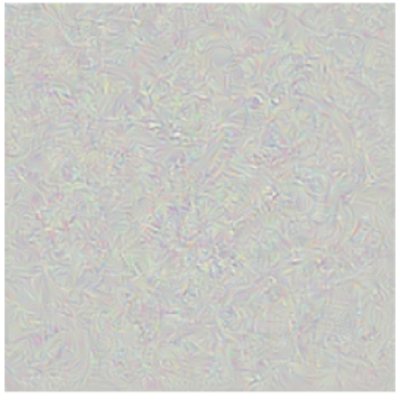
Heuristic	Supporting Cases	Contradicting Cases
Lexical overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

Test

P： **The** doctor **saw the author** and the **tourist**.
H： **The author saw the tourist**.

→ **high** word-overlap → entailment ✗

The Dependence on Spurious Correlations (Dataset Bias)



Article: Super Bowl 50

Paragraph: *"Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV."*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

Original Prediction: John Elway

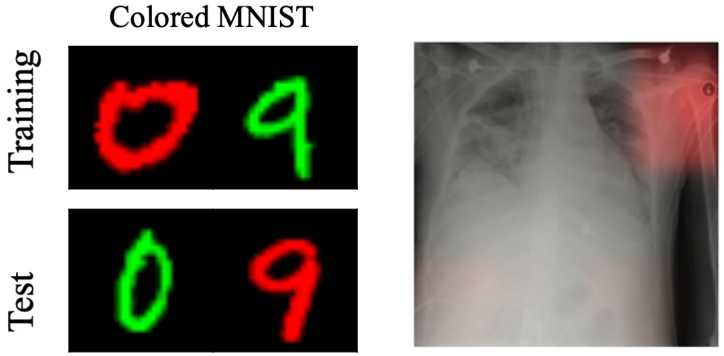
Prediction under adversary: Jeff Dean

Task	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Bias	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Picture source: Geirhos, R., Jacobsen, JH., Michaelis, C. et al. Shortcut learning in deep neural networks. Nat Mach Intell ***2****, 665–673 (2020)*

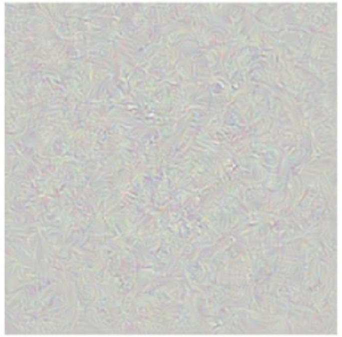
The Effects of Spurious Correlation

Generalizability



Spurious correlations are prone to change on the test set

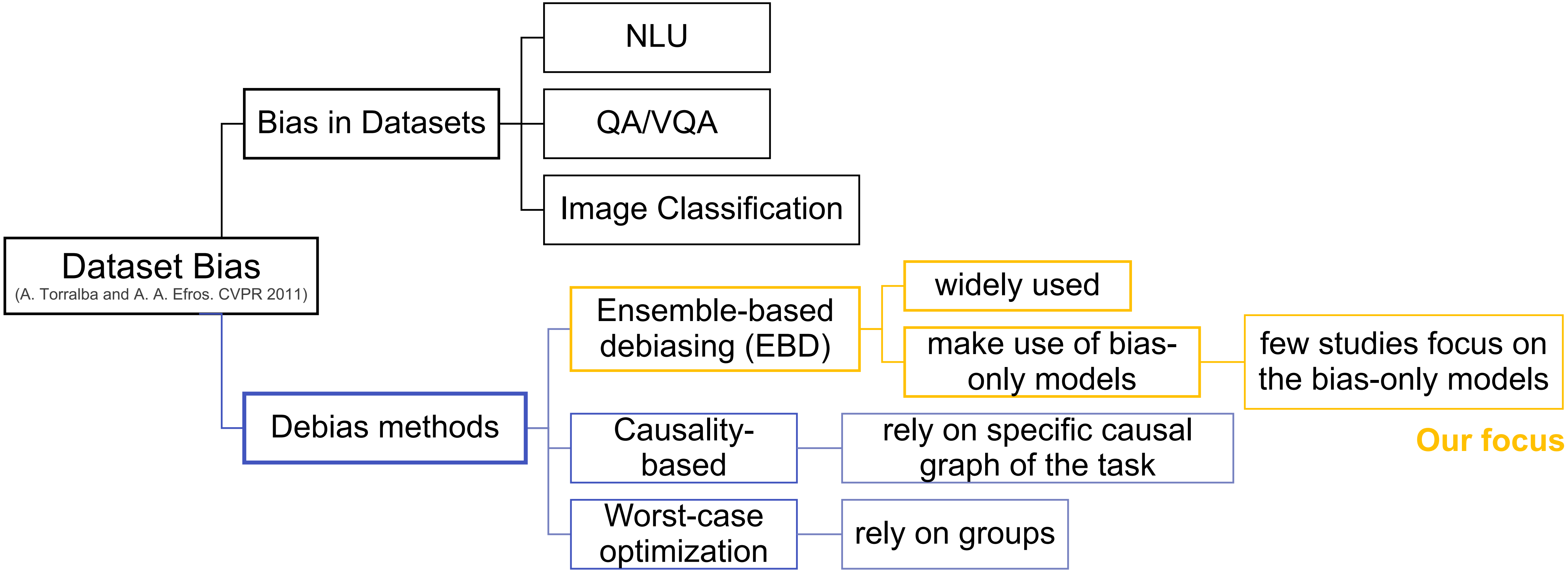
Interpretability & Robustness



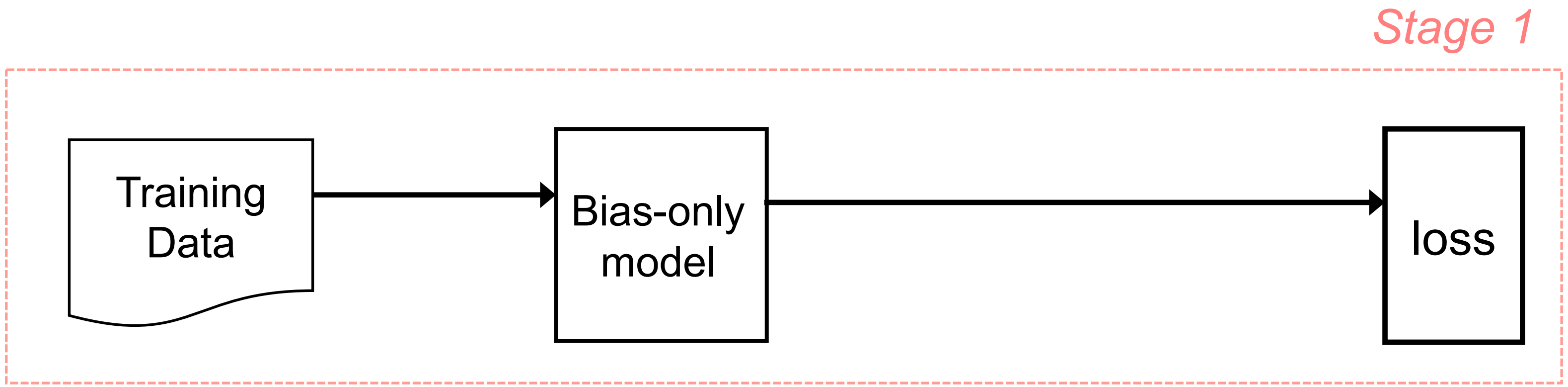
Use features unrecognizable to humans

Geirhos, R., Jacobsen, JH., Michaelis, C. et al. Shortcut learning in deep neural networks. Nat Mach Intell ***2,*** *665–673 (2020); Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019;*

Debiasing: reliance the effect of Dataset Bias



Ensemble-based Debiasing Framework



Ensemble-based Debiasing - Example

- Debiasing: to mitigate model's reliance on **Dataset bias**

Training set

P: **The** little **boy is happy**.
H: **The boy is happy**.

→ **high** word-overlap → entailment ✓

Test

P: **The** doctor **saw the author** and the **tourist**.
H: **The author saw the tourist**.

→ **high** word-overlap → entailment ✗

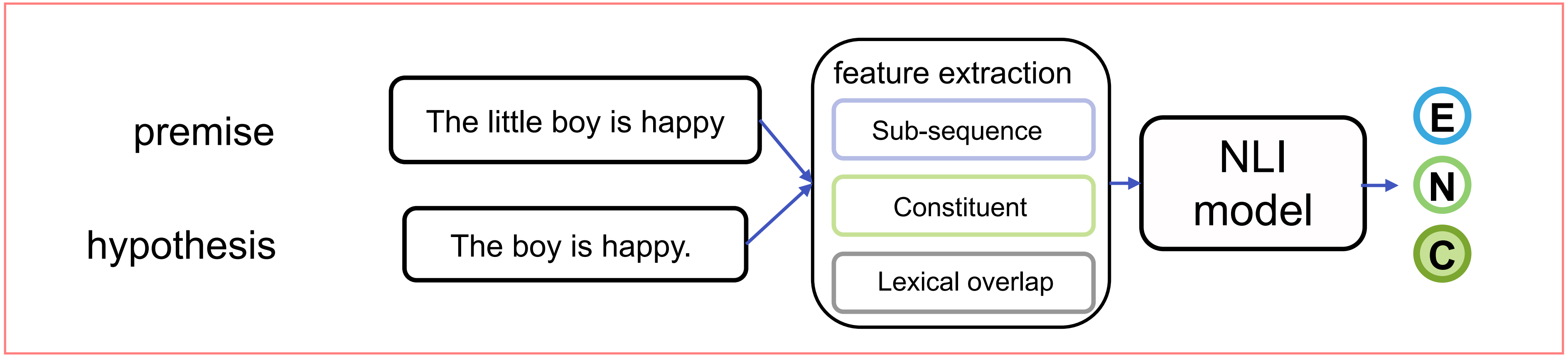
Ensemble-based Debiasing - Example

- Use a bias-only model

Training set

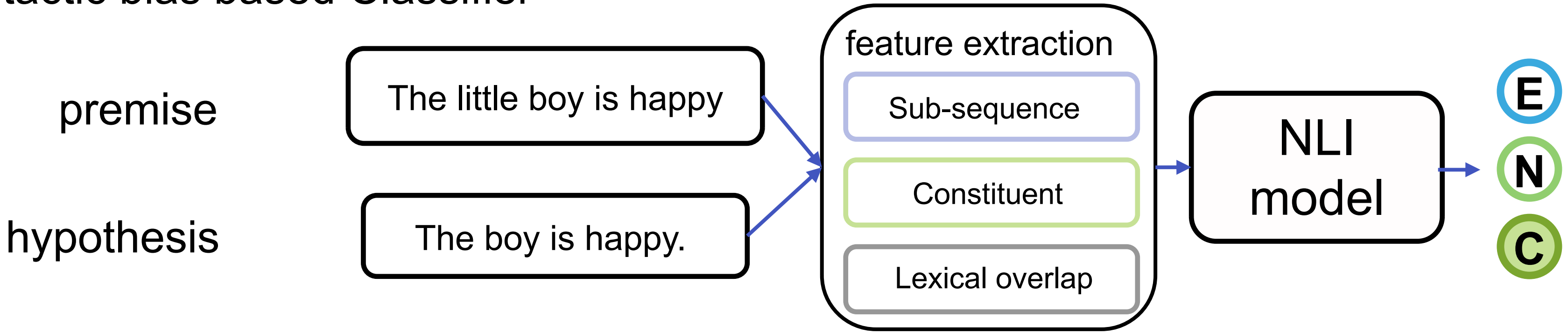
P： The little boy is happy.
H： The boy is happy.

→ **high** word-overlap → entailment ✓

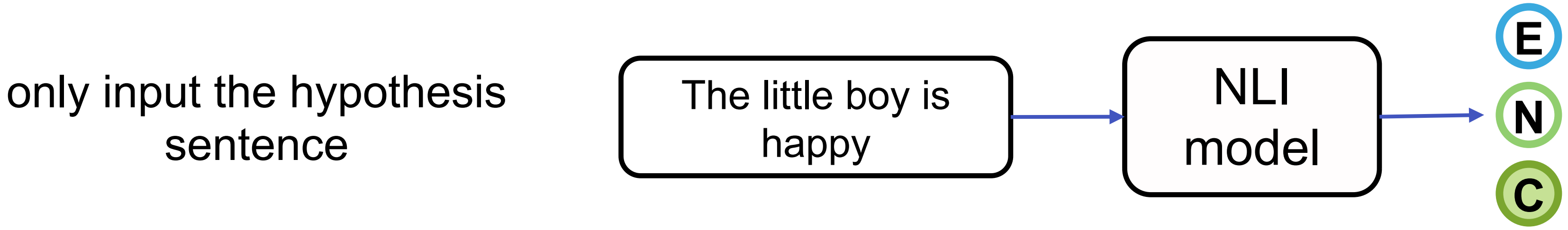


Bias-only Models

- Bias-Known: we have prior knowledge about bias features
  - Syntactic bias based Classifier



  - Hypothesis-only Classifier



Improvements on Bias-only Models

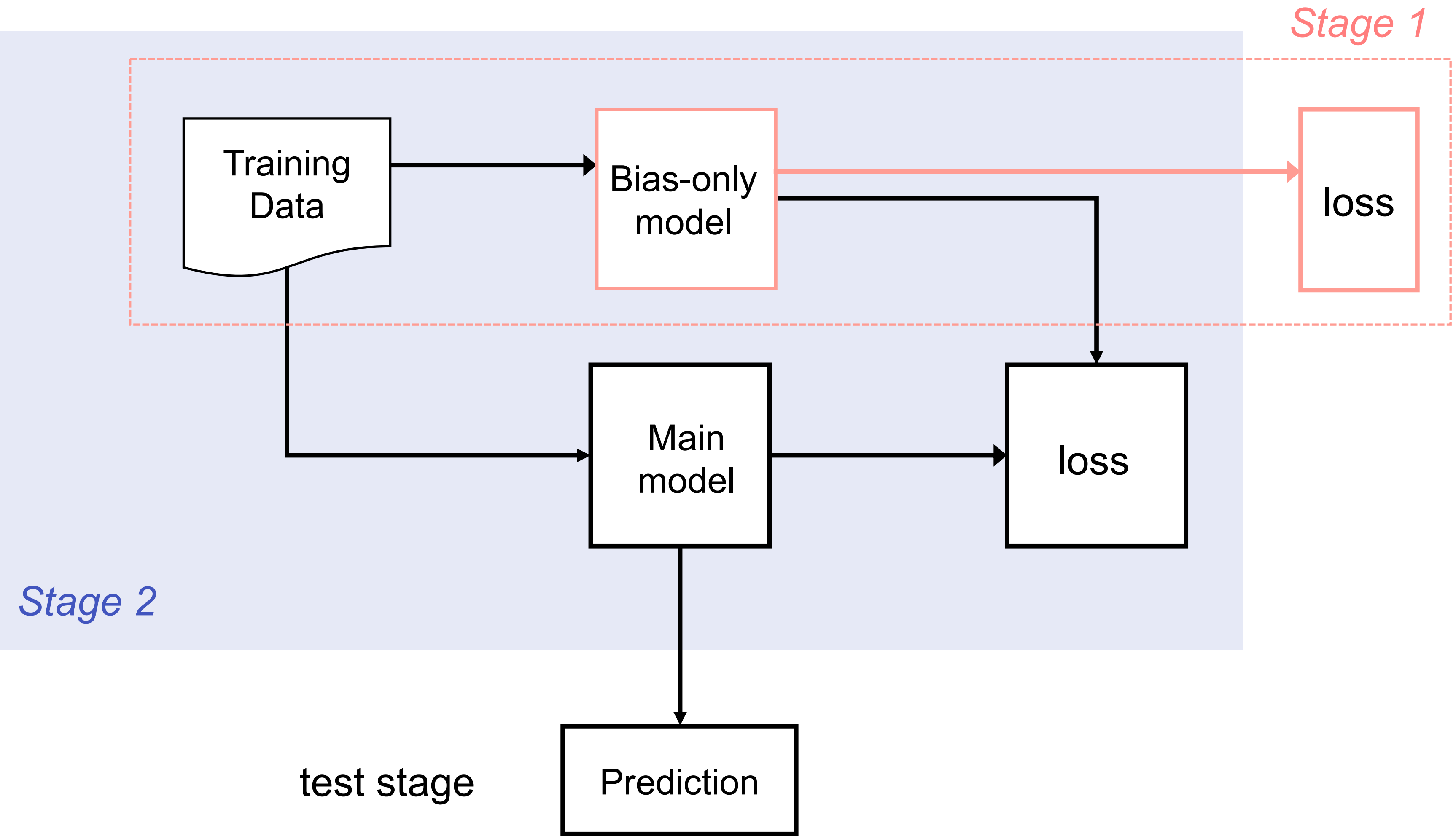
- Bias-Unknown: no identified bias features, using other assumptions
    - Low-capacity model (Clark et al. 2020)

“short-cuts”

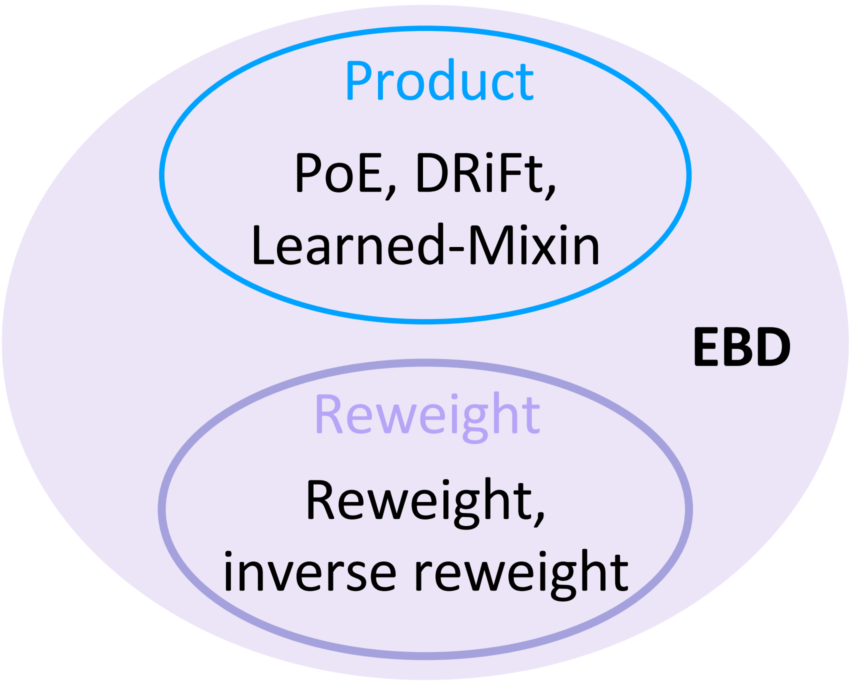
    - Early-stage model (Utama et al. 2020)

Previous work focus on dataset bias other than bias-only model itself

Ensemble-based Debiasing Framework



Ensemble Strategies



CE loss bias-only

- Product methods $\min_{f_M} \mathbb{E}_{X,Y\sim\mathbb{P}_\mathcal{D}}[\mathcal{L}_c(Y, m(\mathbf{q}^b(X) \cdot \mathbf{q}^m(X))],$
  - Product-of-Experts: probability output
  - DRiFt: exponential of the logits output

CE loss

- Re-weight methods $\min_{f_M} \mathbb{E}_{X,Y\sim\mathbb{P}_\mathcal{D}}[\frac{1}{p_Y^b(X)}\mathcal{L}_c(Y, \mathbf{p}^m(X))],$
  - Inverse reweight: probability output

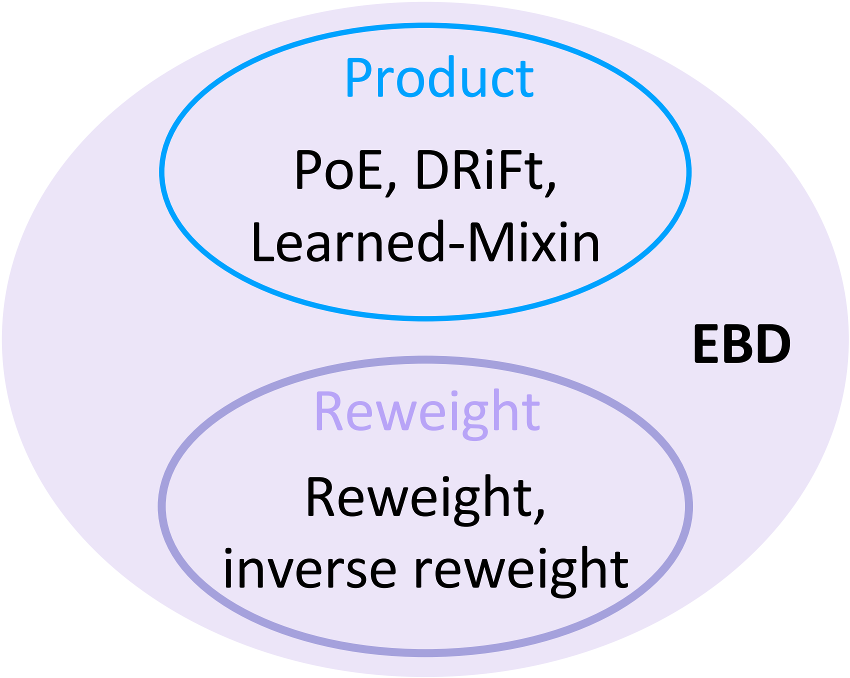
bias-only

Previous work focus on training unbiased model given bias-only model

Contents

- Background
- Motivation
  - Why bias-only models?
  - Why calibration?
- Method and Experiments
- Conclusion and Future Work

Ensemble Strategies



$$\mathbf{p}^{m*} \propto \frac{\mathbb{P}_{\mathcal{D}}(Y \mid X)}{\boxed{\mathbf{p}^{b}}}$$

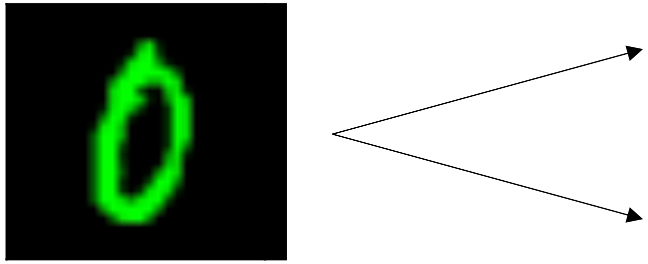
the uncertainty estimation of the bias-only model

The best main model relies on the uncertainty estimation of the bias-only model !

Theoretical Basis of EBD

- **The signal and bias**

$Y = 0$



X

X^S: the shape "0"

X^B: the color green

$$\mathbb{P}_{\mathcal{D}}(Y|X^S) = \mathbb{P}_{\mathcal{D}'}(Y|X^S), \forall \mathcal{D}, \mathcal{D}'$$

The intrinsic (invariant) principle

$\mathbb{P}_{\mathcal{D}}(Y|X^B)$ usually **changes** across different $\mathcal{D}$

Theoretical Basis of EBD

- **The decomposition**

$$\mathbb{P}_{\mathcal{D}}(Y \mid X = x) \propto \mathbb{P}_{\mathcal{D}}\big(Y \mid X^B = x^b\big)\mathbb{P}_{\mathcal{D}}\big(Y \mid X^S = x^s\big)\frac{1}{\mathbb{P}_{\mathcal{D}}(Y)}$$

E.g. Conditional independence $X^S \perp\!\!\!\perp X^B \mid Y$

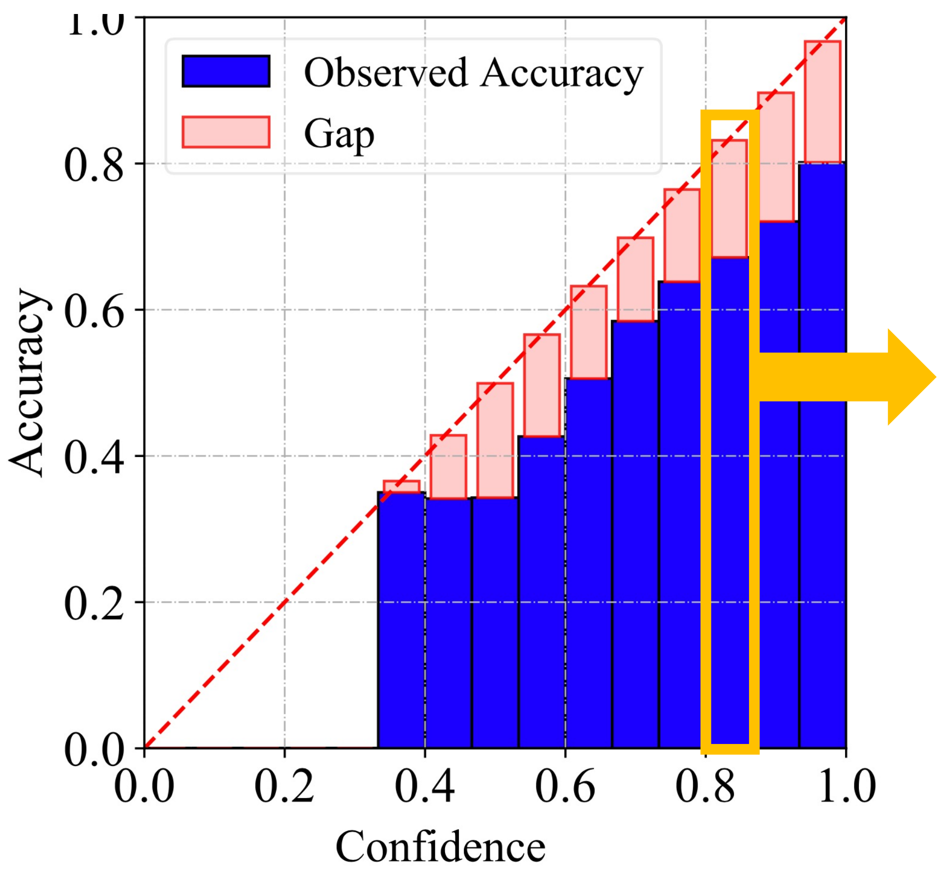
$$\mathbf{p}^{m*} \propto \frac{\mathbb{P}_{\mathcal{D}}(Y \mid X)}{\mathbf{p}^b}$$

➡

When $\mathbf{p}^b \propto \mathbb{P}_{\mathcal{D}}\big(Y \mid X^B\big)$, $\mathbf{p}^{m*} \propto \mathbb{P}_{\mathcal{D}}\big(Y \mid X^S\big)$

The Calibration Problem

Modern machine learning models are poorly calibrated, many are overconfident (Guo et al. 2019)

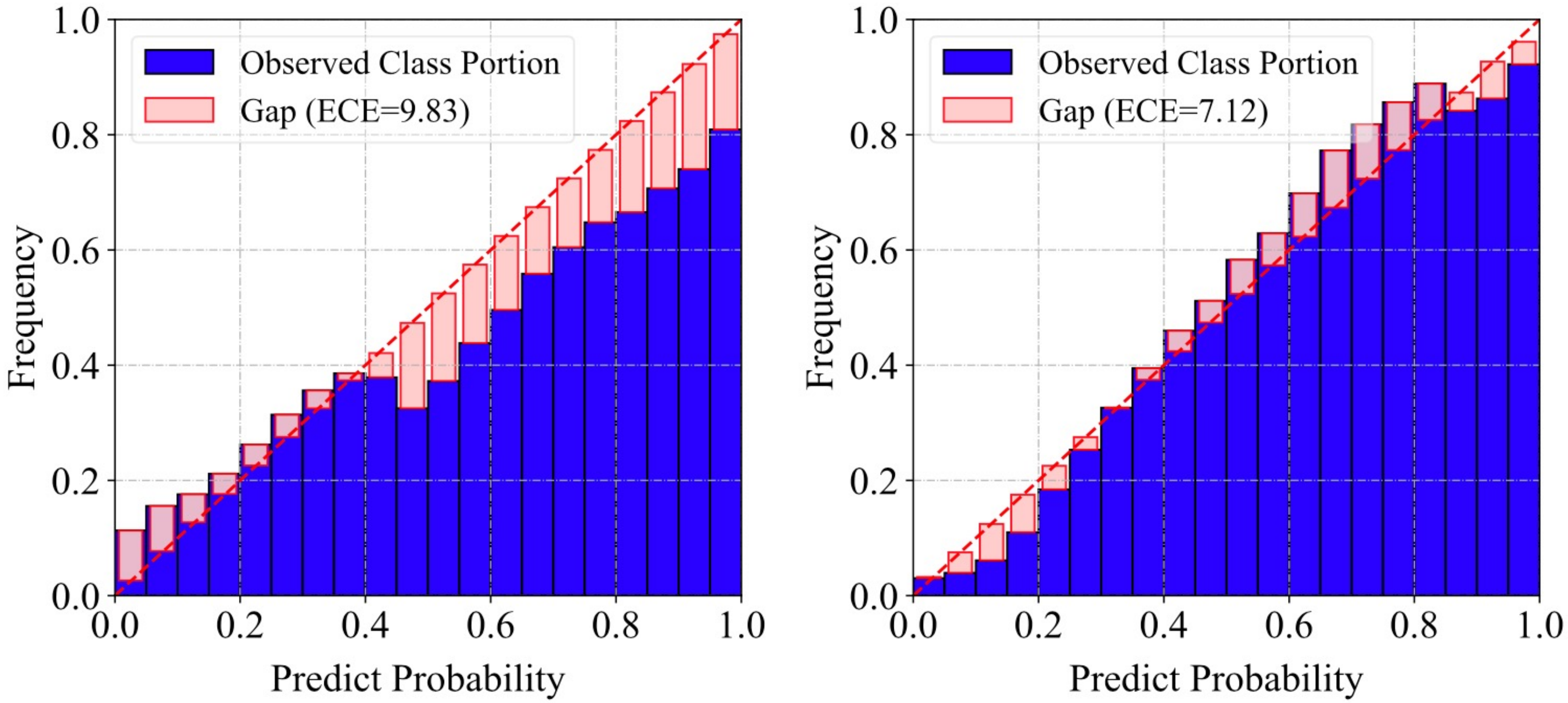


$$\mathbb{P}_{model}(label = i \,|\, x) \approx 0.85$$

$$\mathbb{P}[\text{real label} = i \,|\, \mathbb{P}_{model}(\text{label} = i \,|\, x) \approx 0.85)] \approx 0.65$$

The confidence of model is higher than its accuracy!
------“over confident”(otherwise, lower)

Evidence: Poorly Calibrated Bias-only Models



(a) MNLI (b) FEVER

Bias-only models in EBD methods are poorly calibrated

The Importance of Calibration

- **Theorem 1 (debiasing performance)**

  The out-of-distribution accuracy of the debiased model is **monotonically decreasing with the calibration error** of the bias-only model when such error exceeds a threshold

Theorem 1. *For any* $l \in [0,1]$, *assume that* $\exists l_0$ *s.t.* $\mathbb{P}_{\mathcal{D}}(Y = 0|X^B) \in (l_0 - \epsilon, l_0 + \epsilon)$ *when* X *takes values in* $\mathcal{S}_{f_B}(l)$. *If the calibration error* $|l - \mathbb{P}_{\mathcal{D}}(Y = 0|\mathcal{S}_{f_B}(l))| \geq \delta(l_0, \epsilon, \alpha) > 0$, *the debiasing performance* $\mathbb{P}_{\mathcal{D}}(\{x \in \mathcal{S}_{f_B}(l) | \tilde{Y}(x) = Y(x)\})$ *declines as* $|l - \mathbb{P}_{\mathcal{D}}(Y = 0|\mathcal{S}_{f_B}(l))|$ *increases, where* $\delta(l_0, \epsilon, \alpha)$ *is a constant dependent with* l_0, ϵ *and* α. *When* $\alpha < \frac{1}{2} + \frac{\epsilon}{2l_0(1-l_0)+2\epsilon^2}$, $0 \leq \delta(l_0, \epsilon, \alpha) < 2\epsilon$, *where* $2\epsilon \leq \frac{\epsilon}{2l_0(1-l_0)+2\epsilon^2} < \frac{1}{2}$. *Otherwise* $C < \delta(l_0, \epsilon, \alpha) < 2\epsilon + C$, *where* $0 < C := l_0 - \epsilon - \frac{l_0+\epsilon}{(l_0+\epsilon)+(1-l_0-\epsilon)\frac{\alpha}{1-\alpha}}$, *which increases as* α *increases.*

The Importance of Calibration

- **Theorem 2 (In distribution performance)**

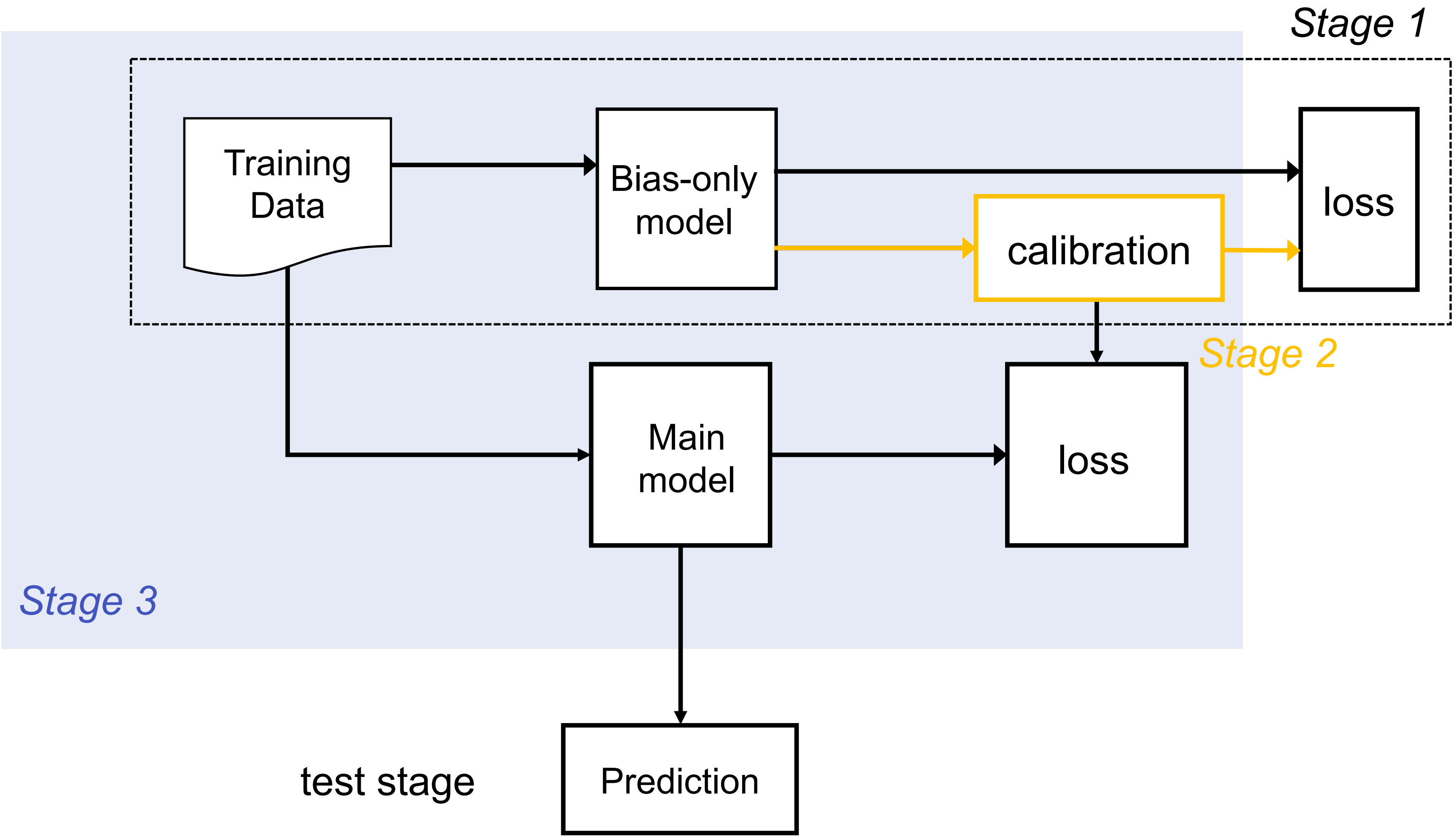
Theorem 2. *For any* X, $\tilde{Y}(X) \neq \hat{Y}(X)$ *if and only if* $p^b_{\hat{Y}(x)}(x) > \mathbb{P}_{\mathcal{D}}(Y = \hat{Y}(x)|X = x)$.

The in-distribution error is **non-decreasing** as the range of the uncertainty estimation of bias-only models increases

An important case: when the bias-only model is over-confident, decreasing its calibration error can **improve both the in-distribution and out-of-distribution** performance of the debiased model according to the two theorems

Contents

Our Framework: MoCaD



Our Framework: MoCaD

- Temperature Scaling (Guo, 2017)

$$L = \frac{1}{n} \sum_{i=1}^{n} \text{logloss}(\sigma(\mathbf{z}/T), y_i)$$

- Dirichlet (Kull, 2019) ➡ **Stronger**

$$\hat{\mu}_{\text{DirLin}}(\mathbf{q}; \mathbf{W}, \mathbf{b}) = \sigma(\mathbf{W} \ln \mathbf{q} + \mathbf{b})$$

$$L = \frac{1}{n} \sum_{i=1}^{n} \log \text{los}(\hat{\mu}_{\text{DirLin}}(\hat{\mathbf{p}}(\mathbf{x}_i); \mathbf{W}, \mathbf{b}), y_i) + \lambda \cdot \left(\frac{1}{k(k-1)} \sum_{i \neq j} w_{ij}^2 \right) + \mu \cdot \left(\frac{1}{k} \sum_{j} b_j^2 \right)$$

Experiment: Datasets

Task	Considered Bias	Train set	IID dev set	OOD test set
NLI	Syntactic	MNLI	MNLI	HANS
	Hypothesis-only			MNLI-Hard-CD MNLI-Hard-SP
	Unknown			HANS
Fact Verification	Claim-only	FEVER	FEVER	FEVER-Symm v1 FEVER-Symm v2

Experiment: Metrics for Calibration

- Class-wise Expected Calibration Error (Class-wise ECE)

$$\text{classwise-ECE} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{m}\frac{|B_{i,j}|}{n}\left|y_j(B_{i,j}) - \hat{p}_j(B_{i,j})\right|$$

Difference between average prediction of class *j* probability and the actual proportion of class *j* in the bin $B_{i,j}$

Experiment: Calibration Results

- Bias-only models after calibration …

	FEVER	HANS	MNLI	Unknown
Un-Cal	7.11	9.83	3.01	7.41
TempS	6.23	7.70	2.38	3.07
Dirichlet	1.73	4.47	0.87	1.45

Classwise-ECE used to measure the performance of calibration, the lower the better

Classwise-ECE significantly drops on all datasets illustrate the effect of TempS & Dirichlet

Experiment: Results on FEVER

	In-distribution	Test (out-of-distribution)	
Method	**ID**	**Symm. v1**	**Symm. v2**
CE	87.1 ± 0.6	56.5 ± 0.9	63.9 ± 0.9
PoE	84.0 ± 1.0	62.0 ± 1.3	65.9 ± 0.6
$\text{PoE}_{\text{TempS}}$	82.0 ± 0.9	63.3 ± 0.9	66.4 ± 0.8
$\text{PoE}_{\text{Dirichlet}}$	87.1 ± 1.0	**65.9** ± 1.1	**69.1** ± 0.8
DRiFt	84.2 ± 1.2	62.3 ± 1.5	65.9 ± 0.7
$\text{DRiFt}_{\text{TempS}}$	81.7 ± 0.9	63.5 ± 1.3	66.5 ± 0.7
$\text{DRiFt}_{\text{Dirichlet}}$	87.4 ± 1.2	**65.7** ± 1.4	**69.0** ± 1.3
InvR	84.3 ± 0.8	60.8 ± 1.2	65.2 ± 1.0
$\text{InvR}_{\text{TempS}}$	83.8 ± 0.6	61.5 ± 0.9	65.4 ± 0.7
$\text{InvR}_{\text{Dirichlet}}$	87.0 ± 0.8	**63.8** ± 2.2	**68.2** ± 1.7
LMin	84.7 ± 1.8	59.8 ± 2.7	65.3 ± 1.1
$\text{LMin}_{\text{TempS}}$	84.9 ± 1.7	60.0 ± 2.5	65.6 ± 1.5
$\text{LMin}_{\text{Dirichlet}}$	87.5 ± 1.1	**61.5** ± 2.4	**67.1** ± 1.3

Consistently better performance in OOD and Dirichlet is a better one

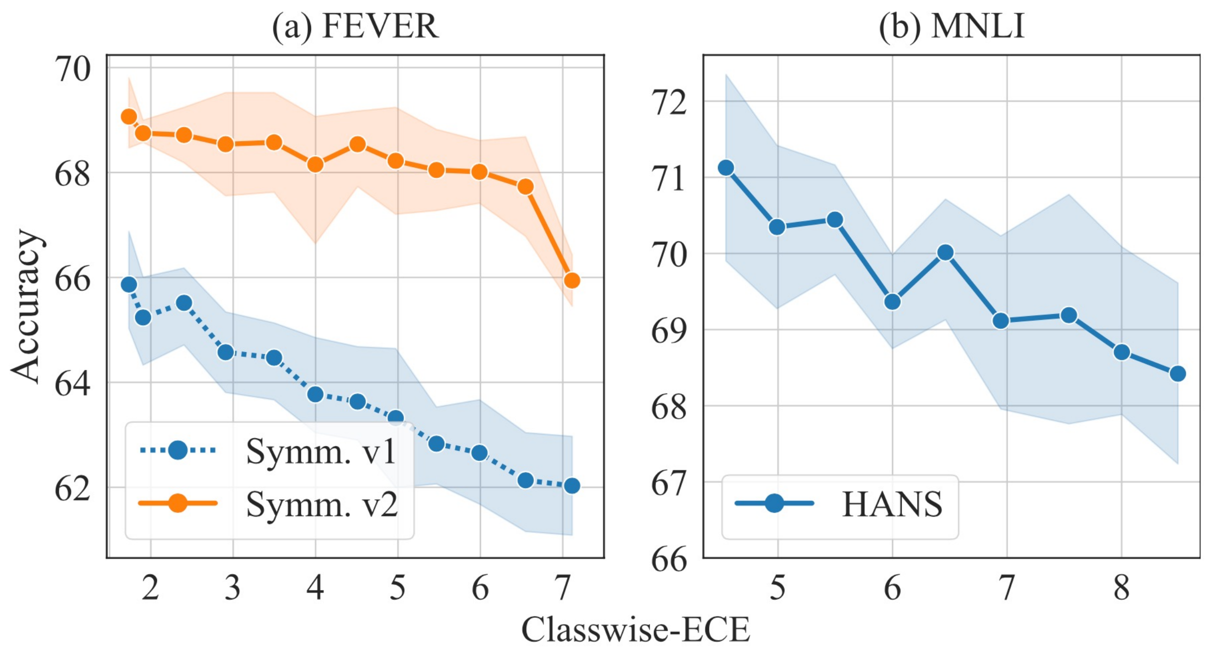
Experiment: Results on MNLI-HANS/MNLI-Hard

Test (out-of-distribution)

Method	Syntactic Bias ID	Syntactic Bias HANS	Hypothesis-only Bias ID	Hypothesis-only Bias Hard_{CD}	Hypothesis-only Bias Hard_{SP}	Unknown Bias ID	Unknown Bias HANS
CE	84.2 ± 0.2	61.2 ± 3.2	84.2 ± 0.2	76.8 ± 0.4	72.6 ± 2.0	84.2 ± 0.2	61.2 ± 3.2
PoE	82.8 ± 0.4	68.1 ± 3.4	83.2 ± 0.2	79.4 ± 0.4	76.8 ± 2.4	80.7 ± 0.2	69.0 ± 2.4
$\text{PoE}_{\text{TempS}}$	83.9 ± 0.3	69.1 ± 2.8	82.9 ± 0.3	79.6 ± 0.4	77.4 ± 2.4	82.1 ± 0.2	69.9 ± 1.6
$\text{PoE}_{\text{Dirichlet}}$	84.1 ± 0.3	**70.7** ± 1.5	82.7 ± 0.4	79.4 ± 0.2	**77.6** ± 2.1	82.3 ± 0.3	**70.7** ± 1.0
DRiFt	81.8 ± 0.4	66.5 ± 4.0	83.5 ± 0.4	79.5 ± 0.6	76.3 ± 1.6	80.2 ± 0.3	69.1 ± 1.3
$\text{DRiFt}_{\text{TempS}}$	83.0 ± 0.4	69.7 ± 1.8	83.1 ± 0.2	79.6 ± 0.2	77.4 ± 3.3	81.5 ± 0.3	**70.0** ± 0.9
$\text{DRiFt}_{\text{Dirichlet}}$	83.6 ± 0.3	**69.8** ± 1.9	82.8 ± 0.3	79.6 ± 0.2	**79.0** ± 1.6	81.9 ± 0.6	69.4 ± 1.1
InvR	82.5 ± 0.1	68.4 ± 1.2	83.1 ± 0.2	78.4 ± 0.5	77.1 ± 2.0	78.7 ± 4.8	64.7 ± 2.6
$\text{InvR}_{\text{TempS}}$	83.6 ± 0.2	69.4 ± 1.6	82.8 ± 0.2	78.6 ± 0.2	77.9 ± 1.7	81.4 ± 0.5	65.8 ± 0.9
$\text{InvR}_{\text{Dirichlet}}$	83.7 ± 0.4	**69.4** ± 1.3	82.5 ± 0.2	78.9 ± 0.4	**80.8** ± 2.0	81.5 ± 0.2	**68.2** ± 0.8
LMin	84.1 ± 0.3	**65.5** ± 3.7	80.5 ± 0.3	80.0 ± 0.4	78.2 ± 2.0	83.1 ± 0.3	**66.5** ± 1.1
$\text{LMin}_{\text{TempS}}$	84.1 ± 0.2	63.2 ± 2.7	80.5 ± 0.6	80.3 ± 0.2	80.8 ± 3.6	83.3 ± 0.2	66.2 ± 1.0
$\text{LMin}_{\text{Dirichlet}}$	84.3 ± 0.3	62.7 ± 2.6	80.1 ± 0.5	79.8 ± 0.4	**83.2** ± 2.2	82.7 ± 0.2	66.4 ± 1.2

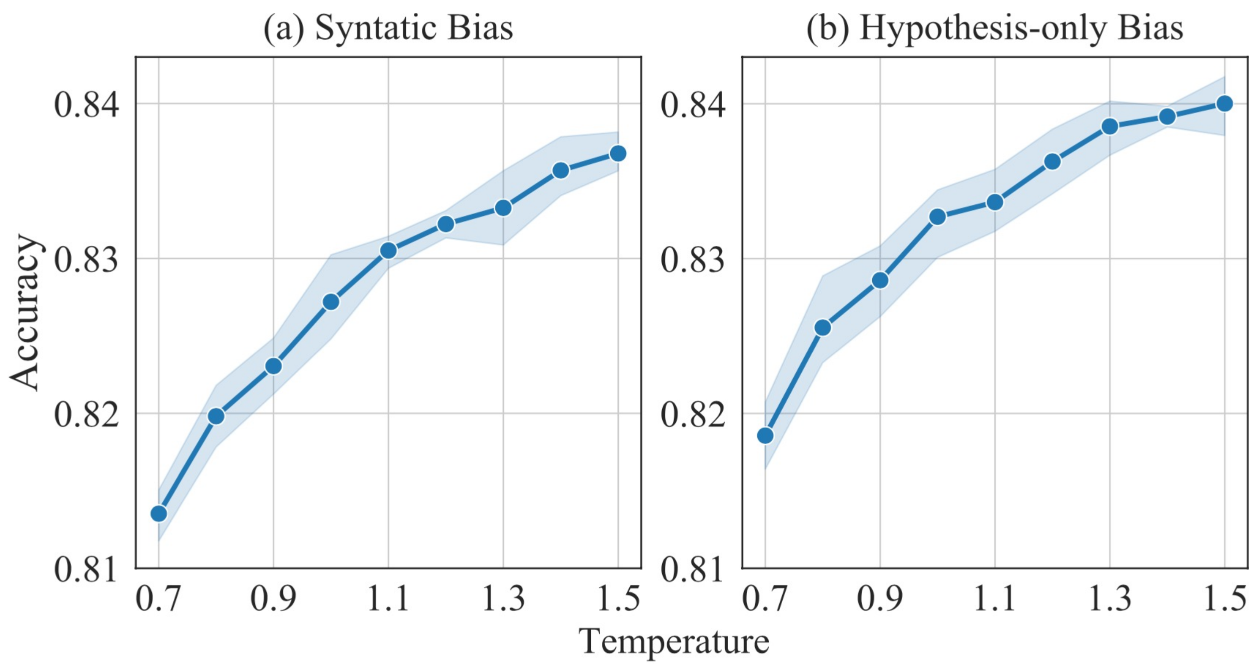
Consistently better performance in OOD and Dirichlet is a better one

Empirical Verification of Theorem 1 (On debiasing performance)



Debiasing performance of bias-only model decreases as the classwise-ECE goes up

Empirical Verification of Theorem 2 (On IID performance)



Bigger temperature -> lower confidence -> better in-distribution performance

Empirical Verification: Over/under Confident

Method	Syntactic Bias ID	Syntactic Bias HANS	Hypothesis-only Bias ID	Hypothesis-only Bias $Hard_{CD}$	Hypothesis-only Bias $Hard_{SP}$	Unknown Bias ID	Unknown Bias HANS
CE	84.2 ± 0.2	61.2 ± 3.2	84.2 ± 0.2	76.8 ± 0.4	72.6 ± 2.0	84.2 ± 0.2	61.2 ± 3.2
PoE	82.8 ± 0.4	68.1 ± 3.4	83.2 ± 0.2	79.4 ± 0.4	76.8 ± 2.4	80.7 ± 0.2	69.0 ± 2.4
PoE_{TempS}	83.9 ± 0.3	69.1 ± 2.8	82.9 ± 0.3	79.6 ± 0.4	77.4 ± 2.4	82.1 ± 0.2	69.9 ± 1.6
$PoE_{Dirichlet}$	84.1 ± 0.3	**70.7** ± 1.5	82.7 ± 0.4	79.4 ± 0.2	**77.6** ± 2.1	82.3 ± 0.3	**70.7** ± 1.0

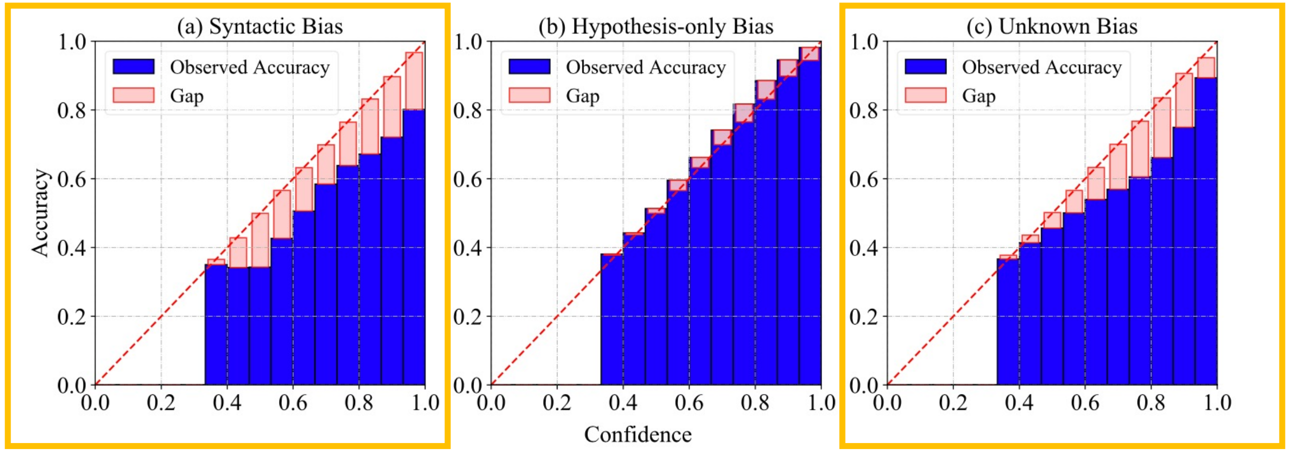


Figure 1: Reliability diagrams of the bias-only models on MNLI. On MNLI, (a) the syntactic bias-only model and (c) the unknown bias-only model are over-confident, (b) the hypothesis-only bias-only model is under-confident.

Calibration of over-confident bias-only benefits performance on both in and out of distribution

Contents

- Background
- Motivation
- Method and Experiments
- Conclusion and Future Work

NEURAL INFORMATION PROCESSING SYSTEMS

Uncertainty Calibration for Ensemble-Based Debiasing Methods

Ruibin Xiong*, Yimeng Chen*, Liang Pang, Xueqi Cheng, Zhiming Ma, Yanyan Lan

Investigating the Bias-only model → (Theory & Empirical evidence) → Calibration of Bias-only model → MoCaD framework

Experiments ✓

THE WAY FORWARD

- Stronger Calibration Methods?
- More Tasks?

Experiments on Image Classification

Table 2: Classification accuracy on image classification.

Method	ID	UnBiased	ImageNet-A
PoE	94.6 ± 0.2	94.3 ± 0.3	31.8 ± 1.9
PoE$_{\text{TempS}}$	94.7 ± 0.3	**94.5** ± 0.3	**31.9 ± 1.1**
PoE$_{\text{Dirichlet}}$	94.6 ± 0.4	94.3 ± 0.4	30.5 ± 1.2
DRiFt	94.6 ± 0.2	94.4 ± 0.3	31.9 ± 0.8
DRiFt$_{\text{TempS}}$	94.8 ± 0.4	**94.4** ± 0.4	**32.5 ± 1.2**
DRiFt$_{\text{Dirichlet}}$	94.5 ± 0.2	94.3 ± 0.2	32.4 ± 1.0
InvR	94.5 ± 0.4	94.1 ± 0.5	31.6 ± 0.3
InvR$_{\text{TempS}}$	94.3 ± 0.1	93.8 ± 0.1	**32.2** ± 1.5
InvR$_{\text{Dirichlet}}$	94.4 ± 0.4	**94.2** ± 0.2	31.8 ± 0.9
LMin	90.9 ± 0.5	90.5 ± 0.6	27.7 ± 1.6
LMin$_{\text{TempS}}$	91.1 ± 0.6	90.6 ± 0.6	**28.1** ± 1.8
LMin$_{\text{Dirichlet}}$	91.2 ± 0.2	**90.9** ± 0.2	26.1 ± 0.8

Experiments on 9-Class ImageNet dataset

MoCaD can achieve the best debiasing performance among all EBD methods, but the improvement is inconsistent.

In Progress: Invariant learning for Debiasing

- Invariant learning for debiasing:
  - Infer environments
  - Minimize the loss with an invariance penalty

$$\min_{f,\theta} \sum_{e \in \mathcal{E}} \lambda_e \mathcal{R}^e(f,\theta) + \lambda \cdot \text{penalty}\big(\{S_e(f,\theta)\}_{e\in\mathcal{E}}\big)$$

(a) **Inferred environment 1** *(mostly) landbirds on land, and waterbirds on water*

(b) **Inferred environment 2** *(mostly) landbirds on water, and waterbirds on land*

- Problem:
  - Optimal solution of Invariant learning may still rely on bias
  - Unstable performance

- Our contribution:
  - Prove necessary and sufficient conditions for the equivalence of invariant learning and debiasing
  - Propose a new method based on the theory

Thanks for Your Attention !