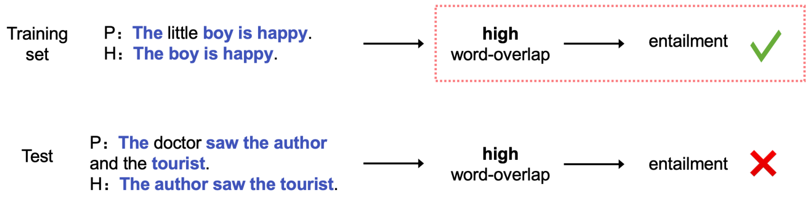
Uncertainty Calibration for Ensemble-Based Debiasing Methods

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

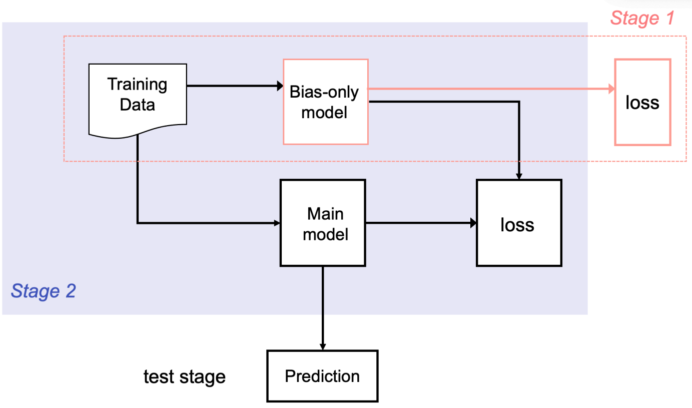
https://github.com/Beastlyprime/MoCaD
https://neurips.cc/virtual/2021/poster/27094

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 ✗

As an example of the dataset bias, suppose in the NLI task, in the training set most sentence pairs with high word overlap are labeled “entailment”. Models that capture this spurious correlation can have low accuracy on the test set with a different data distribution.

EBD Methods



Ensemble-based debasing methods (EBD), e.g. PoE, DRiFt, and Inverse-Reweight usually adopt a two-stage framework.

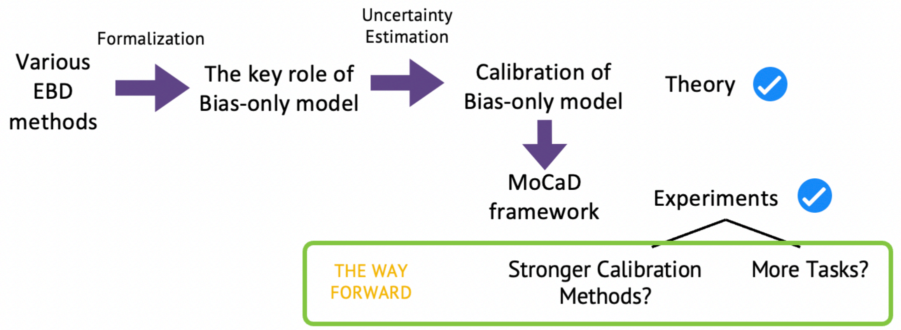
1. A biased predictor is trained based on the bias features only, namely the bias-only model.
2. The output of the bias-only model is then utilized to adjust the learning target of the main model by using different ensembling strategies.

MOTIVATION

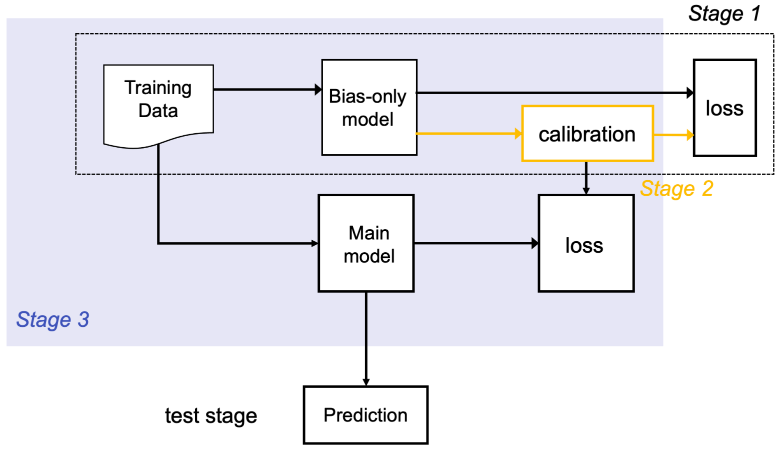
- Ensemble-based debiasing methods have been shown effective in mitigating the reliance of classifiers on specific dataset bias, by exploiting the output of a bias-only model to adjust the learning target.
- Previous works are mainly limited to designing different ensembling strategies, without considering the bias-only model, which clearly plays an essential role in the whole process.

HIGHLIGHTS

- We explore, both theoretically and empirically, the effect of the bias-only model in the EBD methods. A critical problem is revealed: existing bias-only models are poorly calibrated, which will hurt the debiasing performance.
- We propose a model-agnostic three-stage EBD framework to tackle the above problem.
- Experimental results show the superiority of our proposed framework as against the traditional two-stage one.

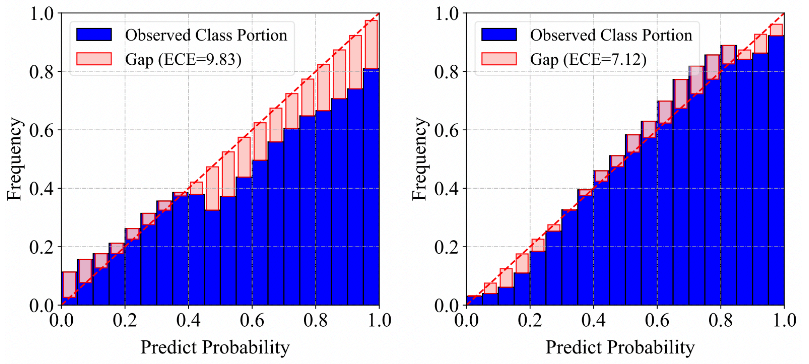


The New Framework



Analysis

- Theoretically, the out-of-distribution accuracy of the debiased main model is monotonically decreasing with the calibration error of the bias-only model when such error exceeds a threshold.
- Especially, when bias-only models are over-confident, decreasing its calibration error can improve both the in-distribution and out-of-distribution performance of the debiased model.
- Empirically, we show the existence of calibration error in existing bias-only models: red bars represent the calibration error.



(a) MNLI (b) FEVER

RESULTS

We experiment with two off-the-shelf calibrators: Temperature Scaling and the Dirichlet calibrator; four challenging benchmarks for NLI and fact verification tasks. The following table shows the 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
PoE$_{\text{TempS}}$	82.0 ± 0.9	63.3 ± 0.9	66.4 ± 0.8
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
DRiFt$_{\text{TempS}}$	81.7 ± 0.9	63.5 ± 1.3	66.5 ± 0.7
DRiFt$_{\text{Dirichle}}$	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
InvR$_{\text{TempS}}$	83.8 ± 0.6	61.5 ± 0.9	65.4 ± 0.7
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
LMin$_{\text{TempS}}$	84.9 ± 1.7	60.0 ± 2.5	65.6 ± 1.5
LMin$_{\text{Dirichlet}}$	87.5 ± 1.1	**61.5** ± 2.4	**67.1** ± 1.3

- We also conduct other detailed experiment to verify our theoretical analysis, See section 6.2.1 and section 6.2.2.