IRM-when it works and when it doesn't: A test case of natural language inference

Yana Dranker[1] He He[2] Yonatan Belinkov[1]

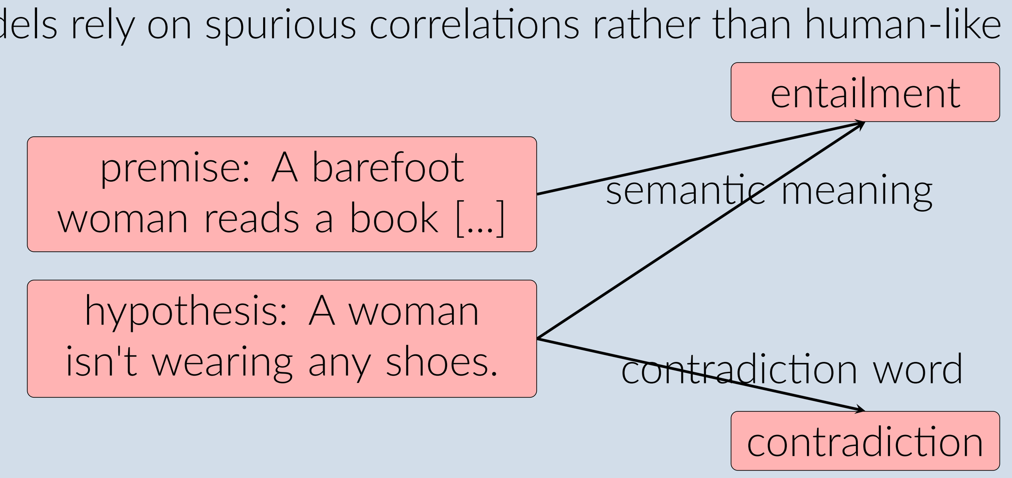
yanadr@campus.technion.ac.il hhe@nyu.edu belinkov@technion.ac.il

[1]Technion - Israel Institute of Technology [2]New York University

Out of Distribution Generalization and Bias

- Good performance on similar distribution to the training distribution.
- Degraded performance on different distribution from the training distribution.
- Lack of out-of-distribution generalization —no performance guarantees in real world scenarios.

Bias is a specific case of out-of-distribution generalization, where models rely on spurious correlations rather than human-like reasoning.



An example of biased and unbiased prediction from the natural language inference (NLI) task, in which we need to infer the relationship between two text fragments. The biased prediction is done based on a ``give-away'' word in the hypothesis.

Invariant Risk Minimization (IRM)

	ERM	IRM
features	predictive	causal
data	shuffled: represents consistent distribution	unshuffled: represents different distributions
o.o.d generalization	✗	✓

Empirical risk minimization (ERM) approach vs. with invariant risk minimization (IRM).

IRM searches for data representation such that the optimal classifier on top of it is optimal for all training environments:

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{tr}} R^e(w\circ\Phi)$$

$$\text{s.t: } w \in \arg\min_{w':\mathcal{H}\to\mathcal{Y}} R^e(w'\circ\Phi)\ \forall e\in\mathcal{E}_{tr} \quad (1)$$

where $\mathcal{E}_{tr}$ are the training environments, R^e is the risk for environment e, w is the classifier and Φ is the data representation. This optimization problem is relaxed into a regularized objective function to yield the practical version of IRM:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda\cdot\|\nabla_{w|_{w=1.0}} R^e(w\cdot\Phi)\|^2 \quad (2)$$

Assuming existence of different environments **IRM suggests a training scheme that uses the different environments to recognize stable rather than environment specific correlations for the classification** process.

Methodology

Previous work exploring IRM either focused on theoretical analysis or on experimentation in simple, synthetic settings. We plan our work based on the following guidelines:

- **Focus on bias** - a specific case of out-of-distribution.
- **NLI task** as a test case—a widely accepted task, with available large datasets.
- **Target two known dataset biases** —overlap bias (correlation between label and word overlap) and hypothesis bias (correlation between label and patterns in the hypothesis).
- **Flexible environment characterization** —analyze effect on performance.

Experiments

We propose 3 steps towards applying IRM to debias NLI models:

	data	bias
toy example	synthetic	synthetic
synthetic bias	natural	synthetic
natural bias	natural	natural

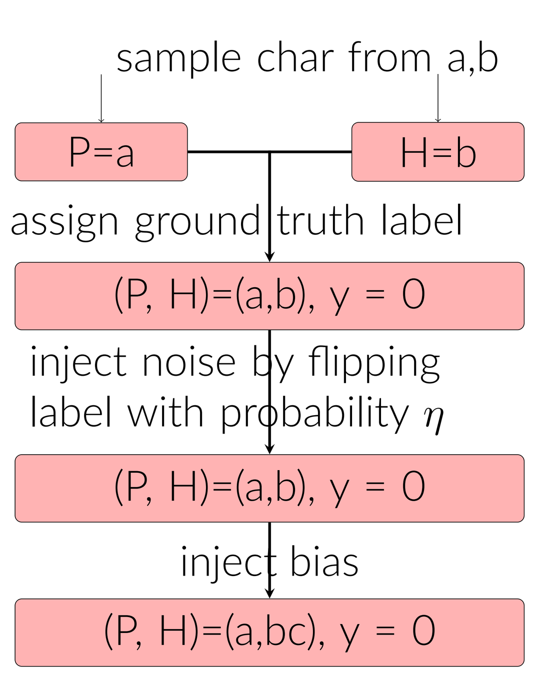
Analysis

We explore the following environment characteristics' effect on model performance:

- *Bias strength* p —how strong is the correlation between a label and a biased feature.
- *Bias prevalence* α —how many examples are biased.
- *Data size*

In all experiments we compare performance of Empirical Risk Minimization (ERM) and IRM.

Toy example



env.	p	α	η	feature used	acc.
Train	0.8	1.0	0.25	true/ biased	85%/ 75%
	0.9	1.0	0.25		
Test	0.0	1.0	0.0	true/ biased	0.0%/ 100%

Environments characteristics and expected accuracy.

	Train	Test
ERM	85.304 ±0.4	0.0 ±0.0
IRM	75.384 ±0.69	100.0 ±0.0

Results on synthetic dataset example

- Model —MLP on top of sum of character embeddings.
- ERM conforms to a classifier relying on the bias signal.
- **IRM relies on the invariant signal to predict**.

Synthetic bias

- Inject synthetic hypothesis bias into SNLI by prepending the hypothesis with a bias token.
- Each label is correlated with a different bias token.
- The model used is bert-base-uncased (Devlin et al. 2018).

env.	p	α
Train	0.7	1.0
	0.9	1.0
Test	0.8/0.33/0.2	1.0

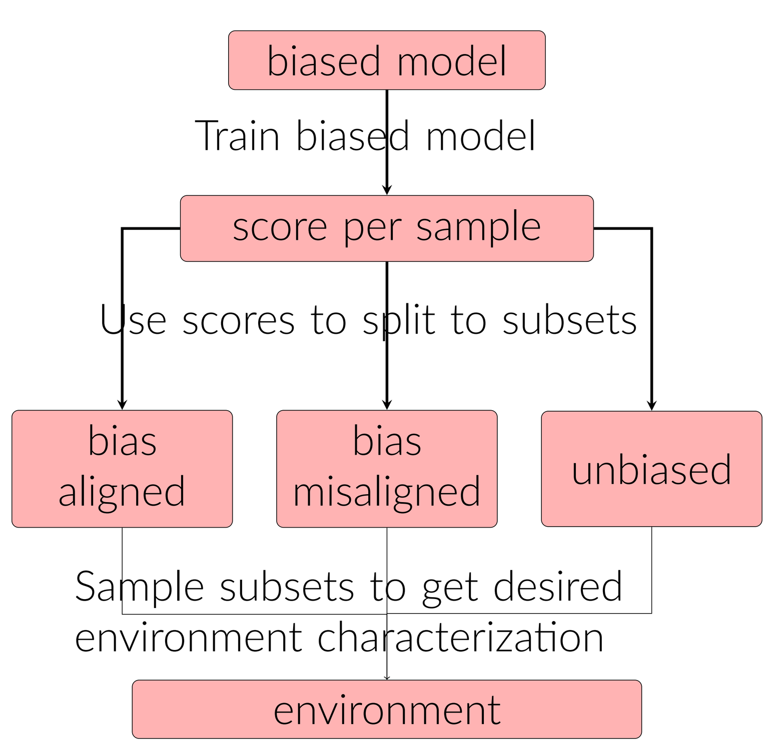
Environments' characteristics.

	$p=0.8$	$p=0.33$	$p=0.2$
ERM	93.49 ± 0.28	85.16 ± 0.9	79.16 ± 1.48
IRM	92.32 ± 0.3	87.22 ± 0.45	83.5 ± 0.71

results in synthetic bias setting.

- As p decreases, **both ERM's and IRM's performance decreases**.
- ERM shows moderate degradation in performance.

Natural bias



env.	p	α
Hypothesis bias		
Train	0.7 0.9	0.82
Overlap bias		
Train	0.7 0.9	0.52
Test —bias aligned	1.0	1.0
Test —bias misaligned	0.0	1.0
Test —unbiased	-	0.0

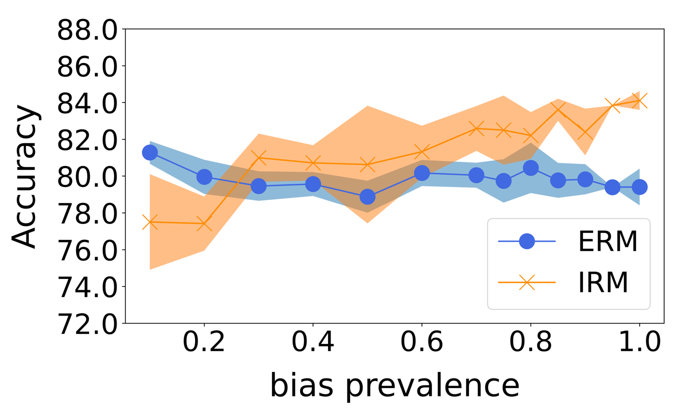
Environments' characteristics.

	unbiased	bias aligned	bias misaligned
hypothesis bias			
ERM	84.46 ± 0.64	97.4 ± 0.26	62.63 ± 1.19
IRM	82.64 ± 1.33	91.4 ± 2.92	65.12 ± 2.28
overlap bias			
ERM	85.23 ± 0.69	96.97 ± 0.28	62.66 ± 1.95
IRM	83.75 ± 0.46	95.44 ± 1.1	64.12 ± 3.86

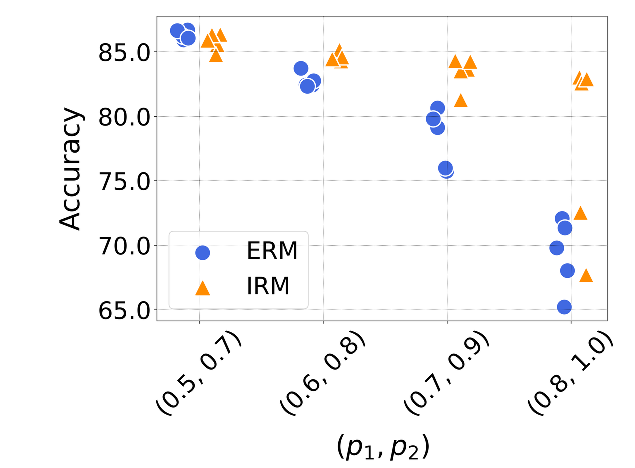
results in natural bias setting.

- Significant performance discrepancies across splits.
- Best performance on bias aligned and worst on bias misaligned, as expected.
- IRM is generally unstable across initialization.
- **IRM improves out-of-distribution performance at the expense of in-distribution performance**.
- Performance degradation of IRM on unbiased split are moderate, as expected.

Analysis



(a) Bias prevalence effect on performance in synthetic bias setting.



(b) Bias strength effect on performance in synthetic bias setting.

- Analyze 3 factors: bias strength, bias prevalence, and data size.
- Vary one factor while keeping the other two fixed.
- Report performance trends for **synthetic bias**.
- similar trends observed for natural bias setting.

factor	ERM	IRM
bias prevalence	—	↑
bias strength ↑	↓	—
data size	↑	↑

Performance improves (↑), degrades (↓), or stays roughly the same (—) in the synthetic bias setting.

Conclusions

- IRM works in natural setting.
- ERM does not solely rely on bias and IRM is not able to fully discard it, thus improvement is rather small.
- Environment characteristics have significant impact on performance.

We hope that our work will encourage research to explore performance in realistic scenarios and flexible settings.