Distributionally Robust Neural Networks for Group Shifts:
On the Importance of Regularization for Worst-Case Generalization

Shiori Sagawa*  Pang Wei Koh*  Tatsunori B. Hashimoto  Percy Liang
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input $x$: bird image

ML model

label: bird type

waterbird vs landbird

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input $x$: bird image

spurious correlation: water background

ML model

prediction $\hat{y}$: waterbird

true label $y$: waterbird

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input $x$: bird image

spurious correlation: land background

ML model

prediction $\hat{y}$: landbird

true label $y$: waterbird

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

input $x$: face image

ML model

label: hair color

blonde hair vs dark hair

Liu et al. (2015)
Models can latch onto spurious correlations

input $x$: face image

spurious correlation: gender

ML model

prediction $\hat{y}$: dark hair

true label $y$: blonde hair

Liu et al. (2015)
Models can latch onto spurious correlations

<table>
<thead>
<tr>
<th>label: object</th>
<th>waterbird</th>
<th>landbird</th>
</tr>
</thead>
<tbody>
<tr>
<td>spurious attribute: background</td>
<td></td>
<td></td>
</tr>
<tr>
<td>water background</td>
<td>majority</td>
<td>minority</td>
</tr>
<tr>
<td>land background</td>
<td>minority</td>
<td>majority</td>
</tr>
</tbody>
</table>
Models perform well on average

average error: 0.03
But models can have high worst-group error

<table>
<thead>
<tr>
<th>label: object</th>
<th>waterbird</th>
<th>landbird</th>
</tr>
</thead>
<tbody>
<tr>
<td>water background</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>land background</td>
<td>![image]</td>
<td>![image]</td>
</tr>
</tbody>
</table>

worst-group error: 0.40
Goal: low worst-group error

- Relies on spurious correlation
- High worst-group error

- More robust to spurious correlation
- Low worst-group error
Our approach: minimize the worst-group loss

Standard (ERM): average loss

\[ R_{\text{ERM}}(w) = \hat{E}_{(x,y,g)} [\ell(w; (x,y))] \]

Group DRO: worst-group loss

\[ R_{\text{gDRO}}(w) = \max_{g' \in G} \hat{E}_{(x,y,g)} [\ell(w; (x,y)) | g = g'] \]

Train: known groups for each example
Test: unknown groups

Optimization algorithm for Group DRO

- Optimizer
- Model
- Model parameters: Update on weighted loss
- Group weights: which are worst-case?

✓ Scalable
✓ Theoretical guarantees
✓ Similar # iterations to convergence as ERM
Attempt 1: ERM $\rightarrow$ high worst-group test error
Attempt 1: ERM \(\rightarrow\) high worst-group test error
Attempt 1: ERM $\Rightarrow$ high worst-group test error

worst-group error is high because of poor generalization
Attempt 1: zero training error $\Rightarrow$ ERM $\approx$ group DRO

worst-group error is high because of poor generalization but group DRO only controls \textit{training} error!
Attempt 1: poor generalization $\Rightarrow$ group DRO fails

worst-group error is high because of poor generalization but group DRO only controls *training* error!
New challenge: train error $\neq$ test error on worst group

Prior work: train error $\approx$ test error for worst-case group
• Small convex or generative models

Our setting: high worst-group test error despite zero train error
• State-of-the-art neural networks

Approach: regularization + group DRO

Problem: zero \textit{train} error, but high worst-group \textit{test} error

Solution: regularization

Counterintuitive with respect to recent trends:
More complex models with zero training error $\rightarrow$ better average error

Hoffer, Hubara, Soudry (2017), Belkin et al. (2019), Nakkiran et al. (2020)
Attempt 2: regularization + group DRO works

ERM with L2 penalty

Group DRO with L2 penalty

✓ worst-group test error
Group DRO + regularization mitigates the spurious correlation problem

Goal: low worst-group error
Thanks!

Thank you to Shyamal Buch, Yair Carmon, Zhenghao Chen, John Duchi, Jean Feng, Christina Heinze-Deml, Robin Jia, Daphne Koller, Ananya Kumar, Tengyu Ma, Jesse Mu, Hongseok Namkoong, Emma Pierson, and Fanny Yang

Funded by Toyota Research Institute (TRI), Open Philanthropy Project Award, Stanford Graduate Fellowship, and Facebook Fellowship Program.