Towards Last-layer Retraining for Group Robustness with Fewer Annotations

Tyler LaBonte[1], Vidya Muthukumar[1], Abhishek Kumar[2]
[1]Georgia Tech [2]Google DeepMind

NEURAL INFORMATION PROCESSING SYSTEMS

Problem: Spurious correlations reduce generalization on minority groups

- Datasets often suffer from ***spurious correlations*** which are predictive but irrelevant for the classification task
- ERM neural networks overfit to spurious correlations and hence perform poorly on ***minority groups*** [1]
- ***Goal:*** Improve robustness by maximizing ***worst-group test accuracy*** rather than average performance

Landbird on land (73%) | Landbird on water (4%) | Waterbird on water (22%) | Waterbird on land (1%)

Prior Work: Last-layer retraining on group-balanced data upweights core features

- ERM models learn ***core features*** of the data, but the spurious features are overweighted in the last layer
- If group annotations are available, ***last-layer retraining*** on a group balanced held-out set can boost WGA [DFR, 2]
- However, group annotations are often ***sensitive to obtain,*** unknown ahead of time, or expensive to annotate

Train full model | Annotate the spurious features on held-out set | Retrain last layer

Baseline: ERM worst-group accuracy depends on data composition and class balance

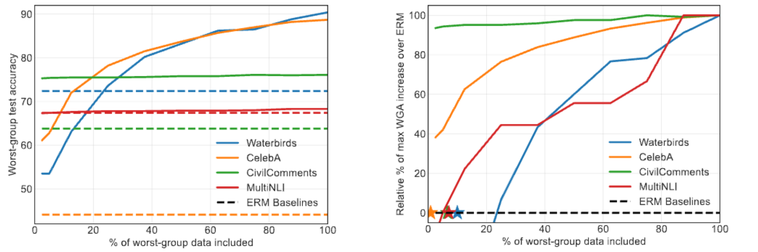
Method	Held-out dataset included	Worst-group test accuracy			
		Waterbirds	CelebA	CivilComments	MultiNLI
CU ERM	✗	$72.4_{\pm 1.0}$	$44.1_{\pm 0.9}$	$63.8_{\pm 6.2}$	$67.4_{\pm 2.4}$
CB ERM	✗	$72.6_{\pm 3.2}$	$66.3_{\pm 3.2}$	$60.2_{\pm 2.7}$	$67.4_{\pm 2.4}$
CU ERM	✓	$81.6_{\pm 1.5}$	$44.5_{\pm 3.4}$	$59.1_{\pm 2.2}$	$69.1_{\pm 1.3}$
CB ERM	✓	$81.9_{\pm 3.4}$	$67.2_{\pm 5.6}$	$61.4_{\pm 0.7}$	$69.2_{\pm 1.6}$

Our contributions

- Performance of last-layer retraining is largely due to ***class-balancing***; perfect group balance is not necessary for improvement
- Retraining the last layer on a group-unbalanced held-out subset of the ***training distribution*** can greatly improve WGA over ERM
- Our SELF algorithm uses ***model disagreements*** to match DFR performance without using group annotations for training

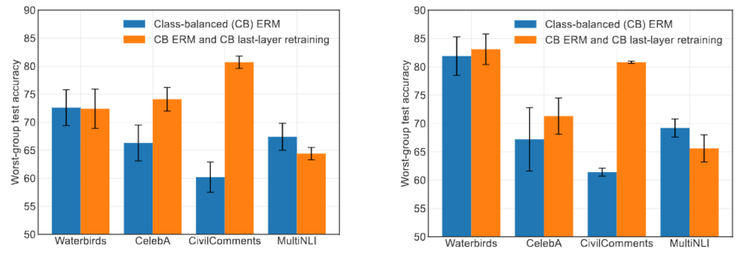
Finding: Last-layer retraining is a free lunch for robustness, no group annotations needed

- Group-balancing implies ***class-balancing***, so how much of the performance of last-layer retraining is due to class-balancing?
- Out of the performance solely due to group-balancing, how does it scale with more worst-group data? ***Is group balance necessary?***



(a) Worst-group test accuracy (b) Relative % of max WGA increase over ERM

- While group balancing is still important for best results, ***class balancing achieves an average of 94%*** of DFR performance
- On average, first ~33% of group balance achieves ~67% of robustness
- A free lunch in group robustness: ***holding out 5% of the data*** for class-balanced last-layer retraining improves WGA 17% over ERM



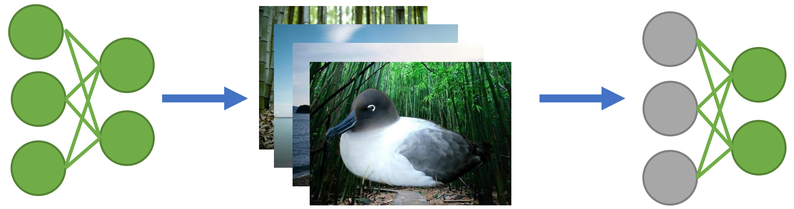
(a) Training dataset only (b) Combined training and held-out datasets

- Works best if ERM is ***saturated***, i.e., if ERM performance is not much improved with more data (like MultiNLI)
- Surprising and unexplained result given that the two splits have ***equally drastic*** group imbalance!

Proposal: Selective last-layer finetuning (SELF) uses disagreement to match DFR performance

Method	Annotations		Worst-group test accuracy			
	Group	Class	Waterbirds	CelebA	CivilComments	MultiNLI
Class-balanced ERM	✗	✓	$81.9_{\pm 3.4}$	$67.2_{\pm 5.6}$	$61.4_{\pm 0.7}$	$69.2_{\pm 1.6}$
CB last-layer retraining	✗	✓	$92.6_{\pm 0.8}$	$73.7_{\pm 2.8}$	$\mathbf{80.4_{\pm 0.8}}$	$64.7_{\pm 1.1}$
ES disagreement SELF	✗	✗	$\mathbf{93.0_{\pm 0.3}}$	$\mathbf{83.9_{\pm 0.9}}$	$79.1_{\pm 2.1}$	$\mathbf{70.7_{\pm 2.5}}$
DFR (our impl.)	✓	✓	$92.4_{\pm 0.9}$	$87.0_{\pm 1.1}$	$81.8_{\pm 1.6}$	$70.8_{\pm 0.8}$
DFR	✓	✓	$91.1_{\pm 0.8}$	$89.4_{\pm 0.9}$	$78.8_{\pm 0.5}$	$72.6_{\pm 0.3}$

- Balanced last-layer retraining is effective but ***still underperforms*** on highly group-imbalanced datasets like CelebA and MultiNLI
- Can we use ***latent model knowledge*** to sample the held-out dataset?
- Yes! Finetune on ***disagreements*** of early-stop and convergent model

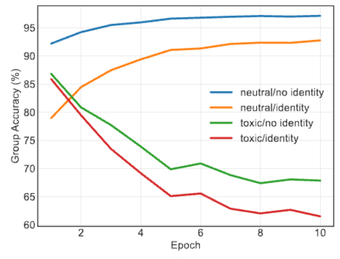


Train full model | ***Early-stop disagreements*** computed on held-out set | ***Finetune*** last layer

Analysis: Why does disagreement SELF greatly outperform misclassification on CivilComments?

Method	Held-out annotations		Worst-group test accuracy			
	Group	Class	Waterbirds	CelebA	CivilComments	MultiNLI
Misclassification SELF	✗	✓	$92.6_{\pm 0.8}$	$83.0_{\pm 6.1}$	$62.7_{\pm 4.6}$	$72.2_{\pm 2.2}$
ES misclassification SELF	✗	✓	$92.2_{\pm 0.7}$	$80.4_{\pm 3.9}$	$65.8_{\pm 7.6}$	$\mathbf{73.3_{\pm 1.2}}$
Dropout disagreement SELF	✗	✗	$92.3_{\pm 0.5}$	$\mathbf{85.7_{\pm 1.6}}$	$69.9_{\pm 5.2}$	$68.7_{\pm 3.4}$
ES disagreement SELF	✗	✗	$\mathbf{93.0_{\pm 0.3}}$	$83.9_{\pm 0.9}$	$\mathbf{79.1_{\pm 2.1}}$	$70.7_{\pm 2.5}$

- Contrary to assumptions of JTT [3] and early-stop misclassification, CivilComments WGA ***decreases over training***
- Training accuracy low for misclassification: ***misclassification captures difficulty, disagreement captures uncertainty***



References

[1] Geirhos et al. "Shortcut learning in deep neural networks". Nature Machine Intelligence, 2:665-673, 2020.
[2] Kirichenko et al. "Last Layer Re-training is Sufficient for Robustness to Spurious Correlations". ICLR, 2023.
[3] Liu et al. "Just Train Twice: Improving Group Robustness Without Training Group Information". ICML, 2021.

Paper Link