The Group Robustness is in the Details: Revisiting Finetuning Under Spurious Correlations

Tyler LaBonte, John C. Hill, Xinchen Zhang, Vidya Muthukumar, Abhishek Kumar

Georgia Institute of Technology

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*** (WGA) rather than average performance


Landbird on land (73%)


Landbird on water (4%)


Waterbird on water (22%)


Waterbird on land (1%)

Prior Work: Class-balancing can improve WGA without any group annotations

- Best way to improve WGA is ***group-balancing***, but this requires expensive group annotations or pseudo-labeling model
- On the other hand, ***class-balancing*** was found to be a simple yet effective method for improving robustness [2]
- We study 3 popular class-balancing techniques and show despite theoretical equivalence, they have ***different empirical behavior***

Landbirds class | Waterbirds class

- ***Subsetting:*** set all classes to the same size as the smallest class by removing data from larger classes uniformly

- ***Upsampling:*** use the entire dataset for training but adjust class sampling probabilities so that SGD mini-batches are class-balanced in expectation

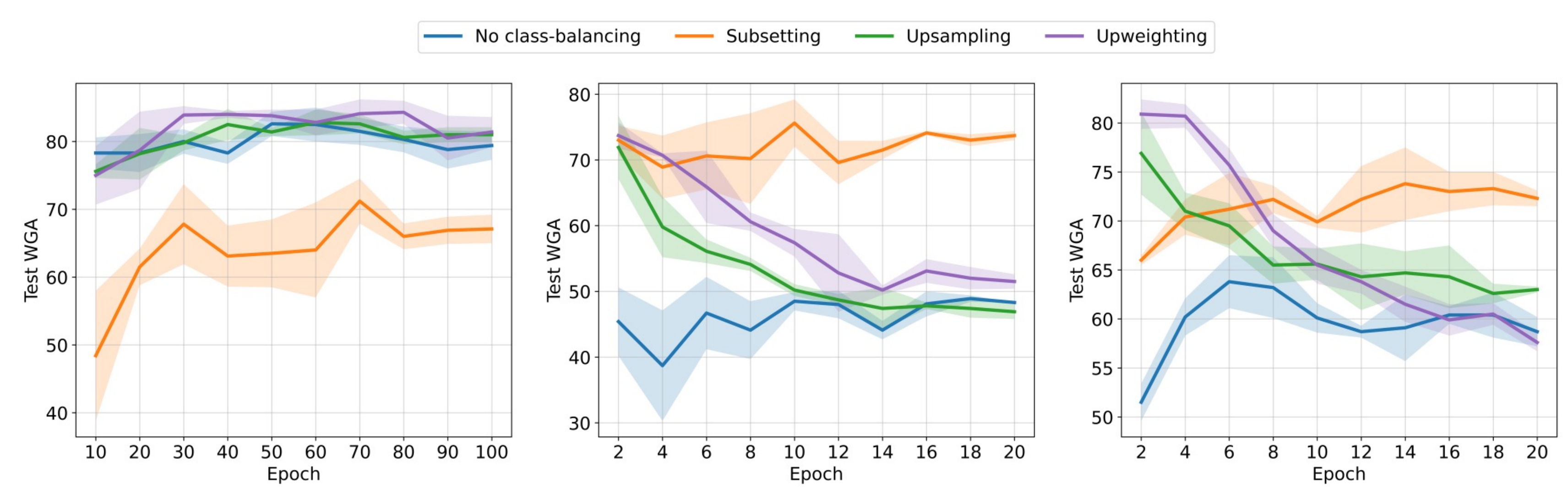
- ***Upweighting:*** use the entire dataset for training but upweight minority class samples in the loss function by the class-imbalance ratio

Our contributions

- We identify new ***failure modes*** of class-balancing: upsampling and upweighting experience catastrophic collapse without extensive tuning
- We show model scaling is beneficial for WGA ***only in conjunction*** with appropriate class-balancing—and scaling can even harm robustness
- Even when classes are balanced, we uncover a ***spectral imbalance*** in the group covariance matrices which may modulate WGA

Finding: Class-balanced upsampling and upweighting overfit minority group over training

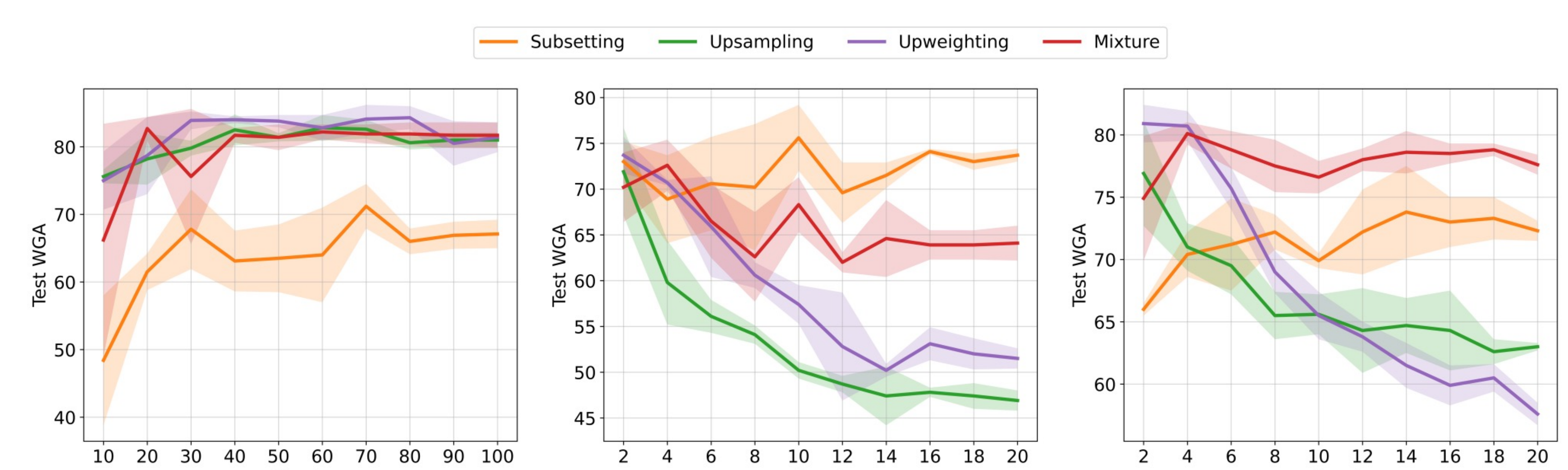
- Upsampling and upweighting experience ***catastrophic collapse*** over long training runs; convergent WGA is no better than ERM
- We also show a ***new disadvantage*** of subsetting: can greatly harm group accuracy on minority groups within majority class (Waterbirds)
- Behavior is caused by overfitting to highly-weighted minority group samples; ***contrary to theoretical equivalence*** in population setting



Waterbirds | CelebA | CivilComments

Proposal: Mixture balancing rectifies collapse by interpolating subsetting and upsampling

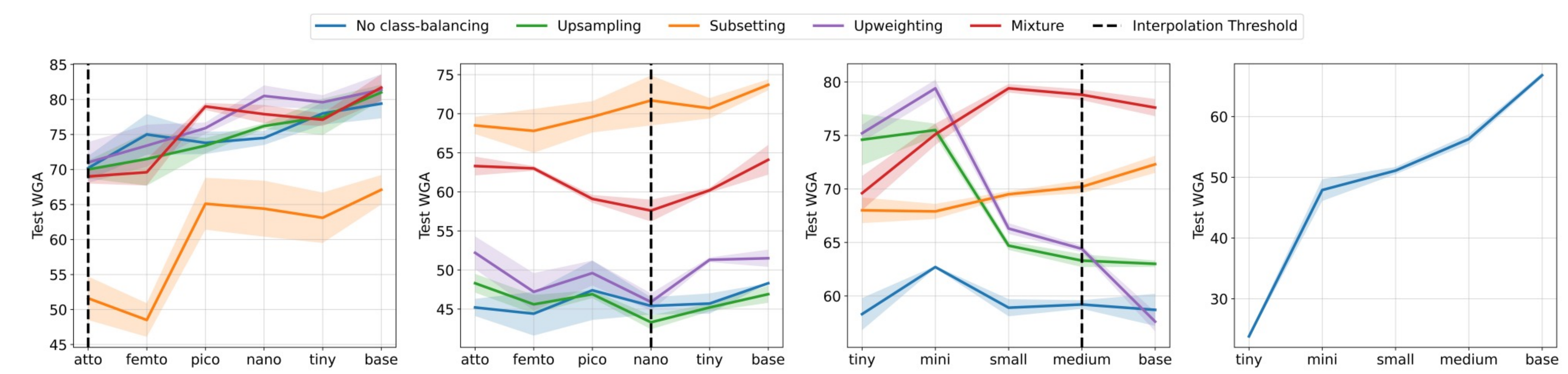
- Our goal is to increase exposure to majority class data ***without oversampling*** the minority class
- We propose ***mixture balancing***, which first takes an imbalanced subset of the original dataset, then runs upsampling on the subset
- Essentially interpolates subsetting and upsampling: achieves the ***best-of-both-worlds*** robustness without group annotations



Waterbirds | CelebA | CivilComments

Finding: Model scaling benefits WGA only in conjunction with appropriate class-balancing

- Previous work showed that model scaling typically does not hurt robustness: we argue their conclusions are ***overly pessimistic*** [3]
- We show that using the right class-balancing technique can greatly ***improve robustness during scaling*** from 3M to 100M+ parameters
- On the other hand, using the wrong class-balancing technique can ***catastrophically collapse*** WGA in large models (CivilComments)
- Takeaway for practitioners: realistic language datasets are not interpolated at any scale (MultiNLI) so ***scaling is key for robustness***

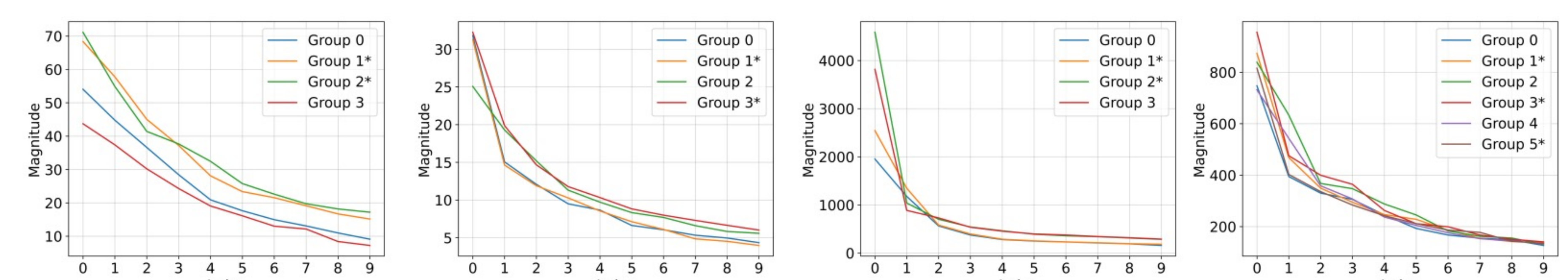


Waterbirds | CelebA | CivilComments | MultiNLI

Analysis: Limits of class-balancing explained by spectral imbalance in the group covariances

- Class-balancing does not improve WGA as much as more targeted methods, but ***isolates contribution*** of group imbalance alone
- Can we analyze ***sources of group disparities*** after class-balancing?
- We show group disparities exist in class-balanced covariance matrices: ***minority groups have larger eigenvalues*** conditioned on class

Group ***g*** covariance for features ***z***: $$\Sigma_g = \frac{1}{|\Omega_g|}\sum_{i\in\Omega_g}(z_i - \bar{z}_g)(z_i - \bar{z}_g)^\top.$$



Waterbirds | CelebA | CivilComments | MultiNLI

References

[1] Geirhos et al. "Shortcut learning in deep neural networks". Nature Machine Intelligence, 2:665-673, 2020.
[2] Idrissi et al. "Simple data balancing achieves competitive worst-group accuracy". CLeaR, 2022.
[3] Pham et al. "The effect of model size on worst-group Generalization". NeurIPS DistShift Workshop, 2021.
[4] Kaushik et al. "Balanced data, imbalanced spectra: Unveiling class disparities with spectral imbalance." ICML, 2024.

Paper Link