Dropout Disagreement: A Recipe for Group Robustness with Fewer Annotations

Tyler LaBonte[1] **Vidya Muthukumar**[1] **Abhishek Kumar**[2]
[1]Georgia Institute of Technology [2]Google Research, Brain Team
{tlabonte, vmuthukumar8}@gatech.edu abhishk@google.com

Abstract

Empirical risk minimization (ERM) of neural networks can cause over-reliance on spurious correlations and poor generalization on minority groups. Deep feature reweighting (DFR) [8] improves group robustness via last-layer retraining, but it requires full group and class annotations for the reweighting dataset. To eliminate this impractical requirement, we propose a one-shot active learning method which constructs the reweighting dataset with the disagreement points between the ERM model with and without dropout activated. Our experiments show our approach achieves 94% of DFR performance on the Waterbirds and CelebA datasets despite using no group annotations and up to $21\times$ fewer class annotations.

1 Introduction

Classification datasets in machine learning often suffer from *spurious correlations*: patterns which are predictive of the target label in the training dataset but irrelevant to the true labeling function. Neural networks trained via the standard procedure of empirical risk minimization (ERM) tend to overfit to spurious correlations and generalize poorly on *minority groups* [3, 18, 5]. This problem is exacerbated under distribution shift, when minority groups are underrepresented in the training distribution but well-represented in the test distribution [15, 9]. In such cases, often a desirable objective is to maximize the model's worst-group test accuracy instead of its mean accuracy over all groups with respect to the training distribution.

Deep feature reweighting (DFR) [8] is a recent state-of-the-art technique for efficiently improving worst-group accuracy. The key insight of DFR is that ERM models which overfit to spurious correlations still learn *core features* of the data, but they perform poorly because they overweight the spurious features in the last layer. Hence, retraining the last layer on a group-balanced *reweighting dataset* can upweight the core features and significantly improve group robustness.

DFR is an efficient and effective algorithm, but its best version requires a held-out dataset with full group and class annotations to achieve maximal performance. This strict requirement prevents its practical application, as the groups are often unknown ahead of time or are difficult to annotate, and holding out data with class annotations means less available data for training and validation.

We propose a simple modification to the DFR procedure which improves group robustness without requiring any group annotations[1] and fewer class annotations. Our key contribution is an approach which requests class annotations for the disagreements between the original model and a *resource-constrained model* to construct a nearly-group-balanced reweighting dataset. Our technique leverages the disproportionate disagreement on minority group points; this phenomenon is intuitive and justified for many common resource-constrained models. For example, early-stopped models tend to fit simple patterns first [1, 11] including the majority group, and dropout models approximate a theoretically justified uncertainty metric [4] which is likely to be higher for minority points. We study dropout

[1] Except on a small validation set for model selection, a common assumption in this setting [15, 11, 13, 8].

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

in this work for its computational efficiency and high performance, but our method is extensible to disagreements with any resource-constrained model (e.g., model size, memory, compute, etc.). Our approach can also be viewed as a *one-shot* active learning method that selects samples for labeling only once and then does last layer retraining with these samples.

Many recent approaches attempt to improve group robustness without group annotations by identifying or oversampling minority groups, and we continue this line of work. Especially relevant to our setting are JTT [11], which is a two-stage training method that upweights the training points misclassified by an early-stopped model; SSA [13], which pseudo-labels the spurious attributes of held-out data; and DivDis [10], which maximizes the disagreement between multiple prediction heads. Compared to these approaches, our method uses significantly less computation (only inference on a dropout model followed by last-layer retraining of the already trained model) and requires fewer class annotations to achieve the same level of minority group generalization performance. The phenomenon that model agreement is closely linked to OOD generalization was first studied by [2]; they focused on disagreement between model classes (*e.g.*, CNNs vs Transformers), while we leverage unique properties of the disagreement between an original and resource-constrained model.

2 Dropout Disagreement Deep Feature Reweighting

We propose dropout disagreement deep feature reweighting (DD-DFR) which achieves 94% of the performance of vanilla DFR despite using no group annotations and up to 21× fewer class annotations. The key insight of our method is that the original and dropout models disagree disproportionately on minority group datapoints, which allows us to construct a nearly-group-balanced reweighting dataset without knowing the groups *a priori*. Our DD-DFR technique is detailed in Algorithm 1.

Algorithm 1 DD-DFR

```
Require: Trained model $f$, dropout model $f'$, held-out dataset $X$, and $\gamma \geq 0$.
Let $A = \{x \in X : f(x) \neq f'(x)\}$.
Sample $B$ from $X \setminus A$ with $|B| = \gamma|A|$.
Let $D = A \cup B$.
Request class annotations for $D$.
Perform DFR with $f$ on $D$ sampling uniformly over classes.
```

We first compute the disagreements between the original and dropout models and augment the disagreement points with a γ proportion of agreement points to form a reweighting set D. We then request class annotations for D, which is typically much smaller than X. Recent work has shown that class balancing is effective for improving worst-group accuracy [7]; hence, while doing last-layer retraining with f on D, we sample uniformly over the classes to construct the SGD minibatches.

To illustrate the benefits of our method, let us compare it with benchmark DFR techniques:

- *Vanilla DFR* [8]: Uses group and class annotations for every datapoint to construct a group-balanced reweighting dataset, which can be difficult to obtain in many practical settings.
- *Misclassification DFR (M-DFR)*: A simple modification to vanilla DFR which constructs the reweighting dataset with the misclassified datapoints in the held-out dataset augmented by a γ proportion of correctly classified points. This technique uses the property that the original model does worse on the held-out samples in the minority group. M-DFR still requires class annotations for *all* datapoints in the held-out dataset, but it does not require group annotations.
- *Dropout disagreement DFR (DD-DFR)*: Our proposed technique constructs the reweighting dataset with the datapoints on which the original and dropout model disagree. Hence, DD-DFR does not require group annotations, and it only requires class annotations for the disagreement points. We observe that DD-DFR outperforms the baseline M-DFR with fewer class annotations.

3 Experiments

3.1 Group Robustness Results

We evaluate our method on the Waterbirds [17, 16, 15] and CelebA [12] benchmark datasets. In Waterbirds, the classification task is between waterbirds and landbirds, and the spurious feature is the background. In CelebA, the classification task is between non-blond and blond celebrities, and the spurious feature is gender. We detail the distributions of each dataset in Table 1.

We train a ResNet50 [6] via ERM for 100 epochs with batch size 32 on the Waterbirds dataset and 50 epochs with batch size 128 on the CelebA dataset. For both datasets, we utilize SGD with learning

Table 1: Train, validation, and test distributions of the Waterbirds and CelebA datasets. For Waterbirds, class label 0 is landbird and 1 is waterbird, while spurious feature 0 is land and 1 is water. For CelebA, class label 0 is non-blond and 1 is blond, while spurious feature 0 is female and 1 is male. Groups G_2 and G_3 are minority groups in Waterbirds, while only group G_4 is a minority group in CelebA.

Group Label	Class Label	Spurious Feature	Waterbirds			CelebA		
			Train	Val	Test	Train	Val	Test
G_1	0	0	3498	467	2225	71629	8535	9767
G_2	0	1	184	466	2225	66874	8726	7535
G_3	1	0	56	133	642	22880	2874	2480
G_4	1	1	1057	133	642	1387	182	180

Table 2: Mean and worst-group accuracy of our DD-DFR method and baselines on the benchmark datasets. DFR† is our implementation for a fair comparison. We report the mean±std over five random seeds and bold the best among methods not using group annotations.

(a) Results on the Waterbirds dataset.

Method	Minimum Additional Annotations		Test Set Accuracy (%)		
	Group	Class	Worst-group	Train Dist. Mean	Test Dist. Mean
ERM	0	0	$71.3_{\pm 0.8}$	$\mathbf{97.8}_{\pm \mathbf{0.1}}$	$89.5_{\pm 0.7}$
SSA [13]	0	599	$89.0_{\pm 0.6}$	$92.2_{\pm 0.9}$	–
DFR [8]	599	599	$92.9_{\pm 0.2}$	$94.2_{\pm 0.4}$	–
DFR†	599	599	$91.8_{\pm 0.3}$	$95.0_{\pm 0.2}$	$94.4_{\pm 0.3}$
M-DFR	0	599	$89.7_{\pm 1.3}$	$92.6_{\pm 1.6}$	$93.7_{\pm 0.8}$
DD-DFR	0	48	$\mathbf{91.6}_{\pm \mathbf{1.3}}$	$94.5_{\pm 0.7}$	$\mathbf{93.8}_{\pm \mathbf{0.6}}$

(b) Results on the CelebA dataset.

Method	Minimum Additional Annotations		Test Set Accuracy (%)		
	Group	Class	Worst-group	Train Dist. Mean	Test Dist. Mean
ERM	0	0	$43.9_{\pm 1.1}$	$\mathbf{95.9}_{\pm \mathbf{0.0}}$	$\mathbf{95.9}_{\pm \mathbf{0.0}}$
SSA [13]	0	9933	$\mathbf{89.8}_{\pm \mathbf{1.3}}$	$92.8_{\pm 0.1}$	–
DFR [8]	9933	9933	$88.3_{\pm 1.1}$	$91.3_{\pm 0.3}$	–
DFR†	9933	9933	$85.1_{\pm 1.4}$	$92.7_{\pm 0.1}$	$92.7_{\pm 0.1}$
M-DFR	0	9933	$82.1_{\pm 1.4}$	$88.8_{\pm 1.6}$	$88.4_{\pm 1.6}$
DD-DFR	0	472	$83.0_{\pm 1.0}$	$89.4_{\pm 1.8}$	$89.4_{\pm 1.8}$

rate 10^{-3}, weight decay 10^{-4}, and momentum 0.9, with standard flip and crop data augmentation. During DFR, we freeze the model up to the last layer, re-initialize the parameters, and train for 100 epochs with the same hyperparameters. Following [8], we use half the validation set for DFR and half for model selection using group annotations. We search over $\gamma \in (0, 0.5, 1, 2, 4)$ and dropout probability $p \in (0.1, 0.3, 0.5, 0.7, 0.9)$, and for CelebA we additionally search over class weights $\in [1, 2, 5]$ similarly to [8]. We report the test metrics of the model with the greatest worst-group validation accuracy in Table 2. Our implementation is in PyTorch [14].

For the sake of simplicity and a fair comparison, our DFR implementation is slightly different than the original [8]. Specifically, we use ℓ_2 regularization instead of ℓ_1, we do class balanced sampling instead of model averaging over repeated samples of majority group points, and we simply retrain the last layer with minibatch SGD instead of full-batch logistic regression on the normalized embeddings. We report results of both implementations and denote ours as DFR†.

Despite using *no additional group annotations*, our DD-DFR method matches DFR† performance with $12\times$ fewer class annotations on Waterbirds and achieves 94% of DFR performance with $21\times$ fewer class annotations on CelebA. Under distribution shift, such as in the Waterbirds dataset, DD-DFR is also effective at improving mean accuracy on the test distribution.

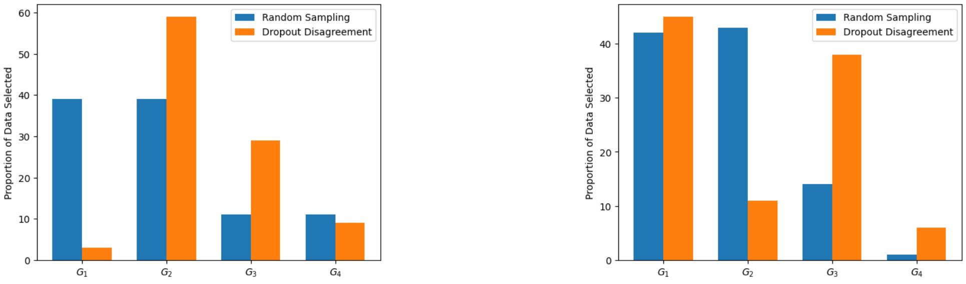
3.2 Ablation Study

We probe our DD-DFR method to quantify its dependencies and variations. In Table 3, we study how the performance of DD-DFR decreases upon ablation of its major components. In Figure 1, we show that dropout disagreement is an effective method for oversampling minority group points without group annotations. Finally, in Figure 2, we study the effect of hyperparameter tuning on worst-group accuracy for the Waterbirds dataset.

Table 3: Ablation study of DD-DFR. We report the mean±std over five random seeds.

Method	Worst-group Accuracy (%)	
	Waterbirds	CelebA
DD-DFR	$91.6_{\pm 1.3}$	$83.0_{\pm 1.0}$
No Class Balancing	$81.6_{\pm 1.8}$	$68.3_{\pm 3.2}$
No Dropout	$86.0_{\pm 8.3}$	$81.2_{\pm 3.6}$
$\gamma = 0$	$53.5_{\pm 21.8}$	$44.3_{\pm 18.3}$

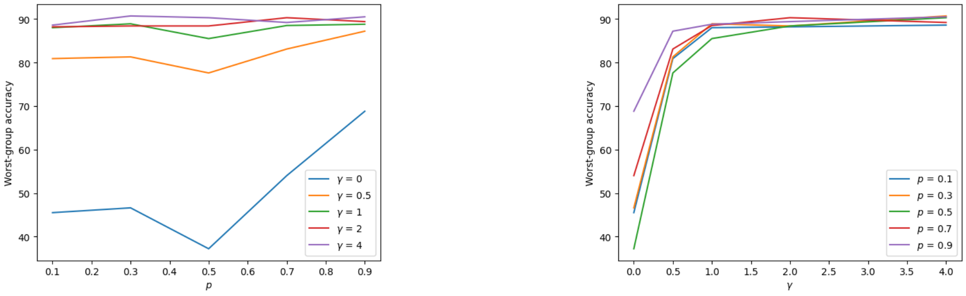
Figure 1 provides insight to why the performance of DD-DFR is slightly lower on the CelebA dataset; while it oversamples the minority group G_4 by a factor of 5, it still takes most points from the majority group G_1 which is over $40\times$ larger. Future work will attempt to further rectify this imbalance.



(a) Waterbirds dataset.

(b) CelebA dataset.

Figure 1: Group proportions in the reweighting dataset for random sampling vs. our DD-DFR technique with $\gamma = 0$. Dropout disagreement enables oversampling of minority group points (G_2 and G_3 for Waterbirds and G_4 for CelebA) without access to group annotations and efficiently constructs a nearly-group-balanced reweighting dataset. We report the mean over five random seeds.



(a) Tuning dropout probability p.

(b) Tuning agreement proportion γ.

Figure 2: Effect of tuning the agreement proportion γ and the dropout probability p on the Waterbirds dataset. Larger values of both γ and p have better performance, with diminishing returns on γ. We plot the mean over five random seeds and drop the worst seed from each configuration (since at low γ it is possible to sample no majority group points in the reweighting dataset).

4 Conclusion

We propose a one-shot active learning method for improving group robustness which utilizes the disagreement between the original and dropout models to construct the DFR reweighting dataset with no group annotations and fewer class annotations. Future work may include disagreement strategies in other resource-constrained settings such as perturbed, quantized, early-stopped, or distilled models.

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[2] Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Sara Beery, Grant van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

[5] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

[8] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Machine Learning (ICML)*, 2022. Workshop on Spurious Correlations, Invariance, and Stability.

[9] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[10] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. In *International Conference on Machine Learning (ICML)*, 2022. Workshop on Spurious Correlations, Invariance, and Stability.

[11] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

[13] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations (ICLR)*, 2022.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[15] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

[17] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Technical report, California Institute of Technology, 2011.

[18] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15:e1002683.