Identifying Spurious Correlations Early in Training through the Lens of Simplicity Bias

Yu Yang YUYANG@CS.UCLA.EDU
Department of Computer Science
University of California, Los Angeles

Eric Gan EGAN8@UCLA.EDU
Department of Computer Science
University of California, Los Angeles

Gintare Karolina Dziugaite GKDZ@GOOGLE.COM
Google Deepmind

Baharan Mirzasoleiman BAHARAN@CS.UCLA.EDU
Department of Computer Science
University of California, Los Angeles

Reviewed on OpenReview: *https://openreview.net/forum?id=t3LBgI4Uw2*

Keywords: spurious correlations, simplicity bias

Extended Abstract

The *simplicity bias* of gradient-based training algorithms towards learning simpler solutions has been suggested as a key factor for the superior generalization performance of overparameterized neural networks (Hermann and Lampinen, 2020; Hu et al., 2020; Nakkiran et al., 2019; Neyshabur et al., 2014; Pezeshki et al., 2021; Shah et al., 2020). At the same time, it is conjectured to make neural networks vulnerable to learning *spurious correlations* frequently found in real-world datasets (Sagawa et al., 2019; Sohoni et al., 2020). Neural networks trained with gradient-based methods can exclusively rely on simple *spurious features* that are highly correlated with a class in the training data but are not predictive of the class in general, and remain invariant to the predictive but more complex *core features* (Shah et al., 2020). This results in a poor *worst-group test accuracy* on groups of examples where the spurious correlations do not hold (Shah et al., 2020; Teney et al., 2022).

An effective way to mitigate a spurious correlation and improve the worst-group test accuracy is to upweight examples that do not contain the spurious feature during training (Sagawa et al., 2019). However, inspecting all training examples to find such examples becomes prohibitive in real-world datasets. This has motivated a growing body of work on group inference: separating majority groups exhibiting spurious correlation with a class, from minority groups without the spurious correlation (Liu et al., 2021; Sohoni et al., 2020; Ahmed et al., 2020; Creager et al., 2021). Despite their success on simple benchmark datasets, we show that such methods suffer from several issues: (1) they often misidentify minority examples as majority and mistakenly downweight them; then, (2) to counteract the spurious correlation they need to heavily upweight their small inferred minority group. This magnifies milder spurious correlations that may exist in the minority group (Li et al.,

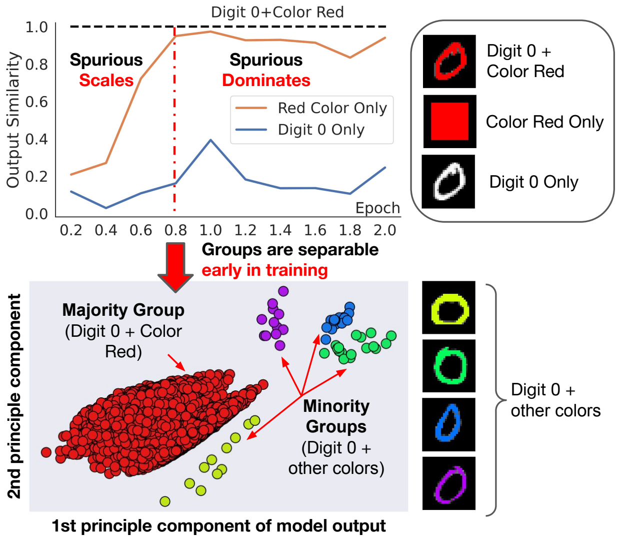

Figure 1: Training LeNet-5 on Colored MNIST. **Top**: Up to epoch 2, the network output is almost exclusively indicated by color red (spurious feature in majority group). **Bottom**: Majority and minority groups are separable based on the network output, e.g. via clustering. Minority groups that have a spurious feature in majority groups of other classes (yellow, purple, blue, green) are also separable from each other.

	Worst-group	Average
CMNIST (1 Spurious × **5 Classes**)		
ERM	$0.0_{\pm 0.0}$	$20.1_{\pm 0.2}$
Spare	$\mathbf{83.0}_{\pm 1.7}$	$91.8_{\pm 0.7}$
$(-2^{nd}$ best)	(**+5.6**)	
Waterbirds (1 Spurious × 2 Classes)		
ERM	$62.6_{\pm 0.3}$	$97.3_{\pm 1.0}$
Spare	$\mathbf{91.6}_{\pm 0.8}$	$96.2_{\pm 0.6}$
$(-2^{nd}$ best)	(**+3.1**)	
UrbanCars (**2 Spurious** × 2 Classes)		
ERM	28.4	97.6
Spare	$\mathbf{76.9}_{\pm 1.8}$	$96.6_{\pm 0.5}$
$(-2^{nd}$ best)	(**+21.1**)	

Table 1: Spare consistently outperforms ERM and the best baselines, especially on UrbanCars, where baselines have been shown to amplify one spurious when trying to mitigate the other (Li et al., 2023).

2023) and harms the performance; as (3) there is no theoretical guideline for finding the time of group inference and group weights, such that methods rely on extensive hyperparameter tuning. This limits their applicability and scalability.

In this work, we make several theoretical and empirical contributions towards addressing the above issues. First, we prove that the simplicity bias of gradient descent can be leveraged to identify spurious correlations. We analyze a two-layer fully connected neural network trained with SGD, and leverage recent results showing its early-time learning dynamics can be mimicked by training a linear model on the inputs (Hu et al., 2020). We show that the contribution of a spurious feature to the network output in the initial training phase increases linearly with the amount of spurious correlation. Thus, minority and majority groups can be *provably* separated based on the model's output, *early in training* (Figure 1). This enables more accurate identification of minorities, and limits the range of group inference to the first few training epochs, without extensive hyperparameter tuning.

Next, we show that once the initial linear model converges, if the noise-to-signal ratio of a spurious feature is lower than that of the core feature in a class, the network will not learn the core features of the majority groups. This explains prior empirical observations (Shah et al., 2020), by revealing *when and why* neural networks trained with gradient almost exclusively rely on spurious features and remain invariant to the predictive but more complex core features. To the best of our knowledge, this is the first analysis of the effect of SGD's simplicity bias on learning spurious vs core features.

Finally, we propose an efficient and lightweight method, Spare (SePArate early and REsample), that clusters model's output early in training, and leverage importance sampling based on inverse cluster sizes to mitigate spurious correlations. This results in a superior worst-group accuracy on more challenging tasks, without increasing the training time, or requiring extensive hyperparameter tuning. Unlike existing methods, Spare can operate without a group-labeled validation data, which allows it to *discover unknown spurious correlations.*

Our extensive experiments (Table 1) confirm that Spare achieves up to 42.9% higher worst-group accuracy over state-of-the-art on most commonly used benchmarks (Alain et al., 2015; Sagawa et al., 2019; Liu et al., 2015; Li et al., 2023) while being up to 12× faster. Applied to Restricted ImageNet, a dataset without known spurious correlations or group-labeled validation set available for hyperparameter tuning, identifies the spurious correlation much more effectively than the state-of-the-art group inference methods and improves the model's accuracy on minority groups by up to 23.2% higher than them after robust training.

Broader Impact Statement

This research addresses the challenge of spurious correlations in machine learning models, particularly those trained with gradient-based algorithms. The positive societal impacts of our work include enhancing the fairness and reliability of AI systems by reducing their reliance on misleading features. This is particularly crucial in high-stakes applications such as healthcare, criminal justice, and hiring, where biases can have profound consequences. Considering the nature and goals of our research, we anticipate no direct negative consequences stemming from our work. Our methodologies are designed to proactively identify and address spurious correlations without compromising the performance of the models.

Acknowledgments and Disclosure of Funding

This research was partially supported by the National Science Foundation CAREER Award 2146492, Amazon PhD fellowship, and Google Cloud research credits.

References

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2020.

Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. *Advances in Neural Information Processing Systems*, 33:17116–17128, 2020.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023.

Evan Z 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*, pages 6781–6792. PMLR, 2021.

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

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L Edelman, Fred Zhang, and Boaz Barak. Sgd on neural networks learns functions of increasing complexity. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3496–3506, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022.