Weakly Supervised Shortcut Learning Mitigation Using Sparse Autoencoders

Muhammad Ahsan[1], Despina Tawadros[1], Sari Sadiya[1],
Phuong Quynh Le[2], Jorg Schlotterer[2], Christin Seifert[2], and Gemma Roig[1]

1- Goethe Universität, Frankfurt 2- Philipps-Universität Marburg

Abstract. Reliance on spurious features that coincidentally correlate with task labels (i.e., shortcut learning) remains a major barrier to the reliable deployment of machine learning models, particularly in high-stakes domains like medical diagnostics. Moreover, in such settings retraining models or collecting and labeling additional data is often impractical, limiting the applicability of many existing shortcut mitigation methods. In this paper we propose a lightweight framework that leverages sparse autoencoders to disentangle spurious from core features to mitigate shortcut learning. Our approach requires no model retraining and works even when group annotations are scarce or unavailable for certain classes. Results on standard benchmarks demonstrate that, even with as few as 50 labeled examples, reliance on spurious features can be significantly reduced.

1 Introduction

Despite strong performance on many diagnostic classification tasks, deep neural networks (DNNs) often struggle when deployed in real diagnostic settings. For instance, a model for detecting bone fracture in CT scans used the presence of safety blankets, which were distributed to patients that were staying in the hospital, as evidence of a fracture. This example is emblematic of shortcut learning, a phenomenon in which models learn to rely on spurious features that coincidentally correlate with the classification labels in the training data [1].

Recently, there has been considerable focus on shortcut learning mitigation [1, 2, 3, 4, 5, 6, 7]. However, many existing methods are impractical in diagnostic settings, as they often require: i) large quantities of group annotations (knowing both the label of an image and if the spurious feature is present) which demand expensive expert annotation [4, 5] ii) data editing, which is challenging for medical images and risks introducing novel artifacts [3] iii) model retraining, often requiring new labels or new stimuli; or iv) features suppression based solely on statistical criteria (such as low complexity [6]), which can be problematic in medical tasks, where truthfulness and interpretability is critical.

In this paper we propose a simple intuitive shortcut mitigation framework that avoids model retraining and does not rely on extensive group annotation. Our approach builds on two observations from previous research. First, DNN neurons are polysemantic, encoding both spurious and core features [3]. Second, sparse autoencoders (SAEs) can disentangle embeddings into more interpretable, feature-specific neurons [8]. Leveraging these ideas, we use a SAE to project model embeddings into a space where spurious and core signals are more easily

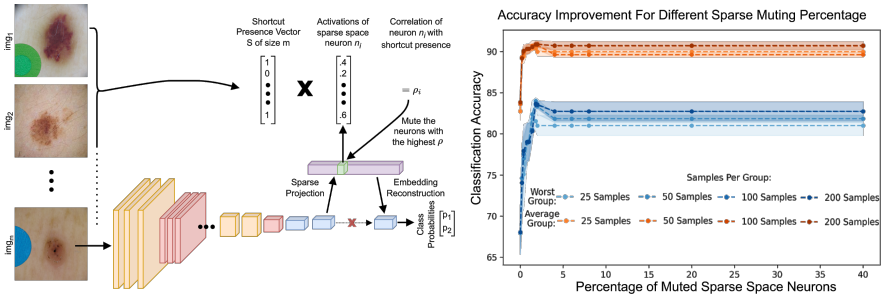

Fig. 1: Left: Overview of our approach. First, the activations of the finetuned model are used to train a SAE. For each neuron in the SAE embedding, we record its activation for each image, resulting in a vector of size M. We then calculate the correlation between this vector and S the presence of the shortcut features in the images. The neurons with the highest correlations are muted before projection back into the embedding space. The classifier head and weights of the original model are not changed. Right: The impact of muting different percentages of neurons on worst and average group accuracy for ResNet50 on WaterBirds.

separated. Correlation analysis over this space allows us to identify and mute "shortcut neurons" before the SAE decoder reconstructs the modified embedding for the classifier. Our experiments show that muting a small subset of SAE neurons reduces shortcut reliance, substantially improving worst and average group accuracy. Also, we found that performance is robust to the exact amount of muting, indicating stable separation between shortcut and core features.

Crucially, our method remains effective even when access to group labels is highly limited. We observe consistent reduction in shortcut dependence even with as little as 50 labeled exemplars in total and when only partial annotation is available, *i.e.*, when there are no exemplars from a specific class. This need for only weak supervision, combined with the minimal overhead of training a simple two-layer sparse autoencoder, makes the framework highly practical. Moreover, these properties are especially useful in real-world diagnostic settings, where finding balanced classes and subgroups are particularly challenging. All code and data necessary to reproduce our results are publicly available[1].

2 Proposed Approach

Given a classifier DNN, our approach eliminates model reliance on shortcuts by muting neurons that encode spurious features. DNN neurons are usually polysemantic and can encode both spurious and core features [9]. Therefore, muting model neurons directly would decrease performance catastrophically. To overcome this, we use the training data activations of the penultimate layer to train a sparse autoencoder, which effectively decorrelates model features into interpretable semantic concepts [8]. Then, to identify which of the N SAE embedding neurons encode spurious features, we extract the activations of every

[1] `github.com/Arsu-Lab/Weakly-Supervised-Shortcut-Mitigation-Sparse-AutoEncoders`

neuron for the M images in the validation set $\{img_j\}_1^M$, resulting in N activation vectors of size $1 \times M$. We then construct a shortcut presence vector of length $s \in \mathbb{R}^M$ which encodes whether the image contains the spurious feature or not.

$$s_j = \begin{cases} 1, & \text{if } img_j \text{ contains the spurious feature} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We compute the Pearson correlation between s and each of the N activation vectors. The resulting correlations $\{\rho_i\}_1^N$ are then used to rank the neurons, identifying those that primarily encode spurious features (highest correlation) versus task relevant features. The neurons most correlated with the spurious features are muted (set to zero) before the SAE decoder reconstructs the original embeddings, enabling the – unchanged – classification head to generate new predictions that do not rely on shortcut presence.

3 Methodology

Models: Following standard practice for diagnostic tasks, we employed popular computer vision models –AlexNet and ResNet50– pretrained on ImageNet-1K and adjusted to each dataset through end-to-end finetuning on the target task.

Sparse Autoencoders: Sparse Autoencoders (SAEs) reconstruct activations in a higher-dimensional space to disentangle and separate overlapping neural representations into distinct and interpretable concepts [8]. We utilize a simple two-layer sparse autoencoder setup. The weights of the auto-encoder were trained with the loss function combining the Mean Squared Error for input reconstructing and an L1 sparsity penalty applied to the hidden layer activations:

$$L_{total} = \frac{1}{M} \sum_{i=1}^{M} ||x_i - \hat{x}_i||^2 + \lambda \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{H} |h_{ij}| \tag{2}$$

Where M is the number of images, x_i and $\hat{x}_i$ are the original and reconstructed input vector, H is the number of neurons in the hidden layer (we always use the penultimate layer), $h_{i,j}$ is the activation of the jth hidden neuron for the ith sample, and λ is a sparsity regularization parameter that controls the strength of $L1$ penalty (in our experiments $\lambda = 0.7$). To ensure our approach is robust we repeat the experiments, including SAE training, with three random seeds.

Datasets: Following previous work [1, 7], we used two common shortcut mitigation benchmarks: The International Skin Imaging Collaboration (ISIC) and the WaterBirds dataset. We report all results when using only 25, 50, 100, or 200 annotated images from each of the four groups in the datasets for shortcut

	Num of Labels per Group:	25 Labels		50 Labels		100 Labels		200 Labels	
		Worst(%)	**Mean**(%)	**Worst**(%)	**Mean**(%)	**Worst**(%)	**Mean**(%)	**Worst**(%)	**Mean**(%)
					ISIC				
AlexNet	Baseline	38.6 ± 1.6	74.2 ± 0.5	38.6 ± 1.6	74.2 ± 0.5	38.6 ± 1.6	74.2 ± 0.5	38.6 ± 1.6	74.2 ± 0.5
	DFR	47.0 ± 0.1	71.8 ± 1.2	52.2 ± 0.5	77.3 ± 1.2	**58.4 ± 0.2**	81.1 ± 0.4	**66.1 ± 1.5**	81.6 ± 0.9
	SUBG	45.2 ± 0.2	70.3 ± 1.4	53.1 ± 0.3	77.0 ± 0.7	59.7 ± 1.6	81.0 ± 1.1	66.5 ± 0.8	81.8 ± 0.2
	Ours (partial)	51.6 ± 0.4	76.3 ± 0.1	53.5 ± 0.4	80.0 ± 0.2	55.1 ± 0.3	80.6 ± 0.2	57.1 ± 0.3	81.2 ± 0.4
	Ours (all)	**52.8 ± 0.6**	**78.5 ± 1.0**	**54.1 ± 0.6**	**80.2± 0.9**	56.5 ± 1.0	**81.7± 0.2**	58.9 ± 1.1	**82.3 ± 0.7**
ResNet50	Baseline	33.5 ± 1.5	71.6 ± 2.0	33.5 ± 1.5	71.6 ± 2.0	33.5 ± 1.5	71.6 ± 2.0	33.5 ± 1.5	71.6 ± 2.0
	DFR	20.6 ± 0.1	65.1 ± 0.1	30.3 ± 0.2	71.3 ± 0.5	58.4 ± 1.6	79.3 ± 0.2	73.0 ± 0.6	82.1 ± 0.6
	SUBG	27.7 ± 1.2	67.1 ± 0.3	38.4 ± 0.8	76.7 ± 0.2	61.6 ± 1.6	80.3 ± 1.0	58.1 ± 0.8	81.0 ± 0.9
	Ours (partial)	74.3 ± 0.1	82.1 ± 0.2	74.2 ± 0.5	82.1 ± 0.2	74.3 ± 0.3	82.9 ± 0.1	74.6 ± 0.3	82.9 ± 1.3
	Ours (all)	**74.9 ± 0.3**	**83.0 ± 0.6**	**74.8 ± 0.2**	**83.0 ± 0.1**	**74.6 ± 0.2**	**83.0 ± 0.1**	**74.8 ± 0.2**	**83.2 ± 0.1**
					WaterBirds				
AlexNet	Baseline	20.2 ± 1.1	66.8 ± 0.2	20.2 ± 1.1	66.8 ± 0.2	20.2 ± 1.1	66.8 ± 0.2	20.2 ± 1.1	66.8 ± 0.2
	DFR	**58.4 ± 0.1**	68.7 ± 0.5	56.0 ± 0.5	70.0 ± 0.3	58.2 ± 0.2	73.1 ± 0.6	57.0 ± .5	75.3 ± 0.2
	SUBG	53.1 ± 0.4	63.7 ± 0.5	**67.1 ± 0.4**	69.0 ± 0.4	**71.6 ± 0.2**	**74.1 ± 0.3**	**72.1 ± 0.5**	**76.2 ± 0.6**
	Ours (partial)	45.6 ± 0.4	69.2 ± 0.5	45.5 ± 0.1	70.9± 0.0	49.5 ± 0.4	71.5 ± 0.0	47.4 ± 0.0	71.7 ± 0.0
	Ours (all)	41.2 ± 0.2	**70.5 ± 0.0**	48.1 ± 0.1	**71.4 ± 0.2**	50.7 ± 0.3	71.6 ± 0.1	50.3 ± 0.1	72.4 ± 0.2
ResNet50	Baseline	67.9 ± 2.5	82.7± 1.0	67.9 ± 2.5	82.7± 1.0	67.9 ± 2.5	82.7± 1.0	67.9 ± 2.5	82.7± 1.0
	DFR	78.4 ± 0.1	88.1 ± 0.0	80.6 ± 0.3	89.0 ± 0.1	83.3 ± 0.2	89.2 ± 0.6	83.4 ± 0.1	88.3 ± 00
	SUBG	59.0 ± 1.2	74.0 ± 0.4	61.6 ± 0.3	72.8 ± 1.3	56.1 ± 1.3	67.0 ± 0.5	76.2.0 ± 1.3	84.2 ± 0.4
	Ours (partial)	81.8 ± 1.2	90.3 ± 0.1	**83.9 ± 0.4**	90.8 ± 0.6	**83.5 ± 0.4**	**90.9 ± 0.3**	83.6 ± 0.6	90.9 ± 0.3
	Ours (all)	**83.3 ± 0.7**	**90.8 ± 0.4**	83.5 ± 0.6	**90.8 ± 0.4**	82.9 ± 1.1	90.8 ± 0.4	**83.6 ± 0.6**	**91.0 ± 0.5**

Table 1: Worst and Average Group accuracy (mean ± std of three seeds) when only 25, 50, 100, and 200 annotated exemplars are available per group and muting 1% of sparse space neurons. Overall, our method outperforms other, more computationally expensive, methods when annotations are scarce. We report performance for baseline AlexNet and ResNet50 models on ISIC and WaterBirds, and compare our method against Deep Feature Reweighting (DFR) and Subgroup Sampling (SUBG). We report performance when using all 4 groups (all) or only two groups (partial). Specifically, in the partial setting we use the two benign groups in ISIC and waterbird groups in WaterBirds. When specific neurons clearly encode the shortcut signals (like in the case of ResNet50 on ISIC) we observe consistent results regardless of the number of labels available.

mitigation (in our method, picking which neurons to mute). We also explore a weak-supervision scenario, when only partial access to two groups is available (data from only a single label). Specifically, we report performance when all annotated exemplars are benign tumors (with/out bandages) for ISIC, and waterbirds (on land/water backgrounds) for WaterBirds.

Both ISIC and WaterBirds are widely used in shortcut learning research [7, 1]. ISIC includes images of malignant and benign tumors, where 50% of the benign examples contain colored bandages (the spurious feature) compared to only 0.5% of the malignant samples. Therefore, in ISIC worst-group accuracy refers to the classification accuracy on malignant samples with bandages. The Waterbirds datasets comprises of images of water and land birds overlaid on water and land backgrounds. A spurious correlation is introduced through the co-occurrence of bird and background types - in the training set 95% of waterbirds are shown on water backgrounds, and 95% of landbirds on land backgrounds - making background a confounding factor. For both datasets we used a held-out balanced *annotated validation set* with 200 exemplars from each of the 4 groups. Finally, for testing we used another held out set with 100 exemplars from each group. For further details regarding the datasets we refer the reader to [1].

4 Experiments and Results

We benchmark our framework against popular shortcut mitigation methods that require group annotations: subsampling groups (SUBG) [4] and Deep Feature Reweighting [5]. All experiments were repeated with three sequential random seeds. Following standard practice, we report worst and average group accuracy.

The results of our experiments demonstrate that, in low data-annotation regimes, our framework significantly improves both worst and average group classification performance compared to the baseline and other supervised shortcut mitigation approaches, (Table 1). Results are largely consistent across models and datasets, indicating that – even with as few as 25 examples from only two groups – it is possible to identify which neurons in the sparse space encode the shortcuts, and muting them does not degrade the task's relevant information.

Results with Partially Labeled Data We also observe significant improvement in worst and average group performance even when annotation from only two groups was available (benign tumors with/out bandages for ISIC, and landbird on land/water backgrounds for WaterBirds, see the "ours (partial)" results in Table 1). Partial annotations could be especially useful in the context of medical diagnostics, as it greatly simplifies annotation collection. To our knowledge, no prior work tackled shortcut mitigation in a partial supervision scenario.

Robustness to Muting Percentage To identify the optimal number of sparse space neurons to be muted to achieve the highest accuracy gain we plotted the worst and average group accuracy as a function of percentages of muted neurons. Overall, accuracy values show a sharp initial increase before reaching a

plateau that lasts until over 70% of the neurons are muted (Fig 1 right and on Appendix). This further simplifies the proposed method as we can eliminate the need for a separate validation set for identifying the optimal muting percentage.

5 Discussion and Conclusion

Shortcut learning remains a pervasive challenge with implications for domain adaptation, fairness, and bias mitigation. Although these issues often surface most critically during deployment, many existing mitigation approaches are unsuited for real-world settings, where annotation, model retraining, and data editing are costly. We introduced a lightweight shortcut mitigation method that improves both worst and average group accuracy without modifying model weights and while requiring only weak supervision. Our results demonstrate that sparse autoencoders offer an effective mechanism for attenuating shortcut reliance, even when annotations are limited or available for only a subset of groups. Overall, our framework highlights the potential of simple, targeted interventions to enhance model robustness under real-world constraints.

References

[1] Lukas Kuhn, Sari Sadiya, Jorg Schlotterer, Florian Buettner, Christin Seifert, and Gemma Roig. Efficient unsupervised shortcut learning detection and mitigation in transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[2] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

[3] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1), 2022.

[4] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[5] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *Proceedings of the International Conference on Learning Representations*, 2023.

[6] Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *Proceedings of the International Conference on Machine Learning*, 2023.

[7] Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Out of spuriousity: Improving robustness to spurious correlations without group annotations. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[8] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations*, 2024.

[9] Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Is last layer re-training truly sufficient for robustness to spurious correlations? In *International Joint Conference on Artificial Intelligence Workshop on XAI*, 2023.

A Appendix: Muting Percentage Analysis

Generally, we observe that increasing the percentage of muted neurons rapidly improves performance until around 1% and then we see a plateau until 70% to 95% (depending on the dataset and model) when performance starts to degrade. Below is the performance of our method with ResNet 50 (with annotations from all groups) on WaterBirds.

Muting Percentage	Worst Group Acc	Average Group Acc
0	65.21	84.59
0.3	76.95	90.54
0.6	79.13	91.01
1.0	84.48	91.44
1.5	84.48	91.44
2	83.07	91.53
5	83.07	91.65
10	83.07	91.65
40	83.07	91.65
70	83.07	91.65
90	83.07	91.65
95	83.07	91.65
97	83.07	91.65
98	2.28	68.33
99	0.82	66.66

Table 2: Performance of ResNet50 with various muting percentages

The consistency of this pattern eliminates the need for hyper-parameter optimization for identifying the optimal muting percentage, as any value between 5% and 50% is like to result in similar performance.