The Presence of Spurious Correlations in Contrastive Learning

Jade Xu
Department of Computer Science
UCLA, Los Angeles, CA, USA
jade1st@ucla.edu

Jenny Wang
Department of Computer Science
UCLA, Los Angeles, CA, USA
jianingwang99@g.ucla.edu

Abstract

Spurious correlations has been a long term factor that influences the machine learning model's prediction performance, where the model tends to learn the unrelated feature instead of crucial ones. While this problem could be alleviated in the supervised learning through upweighting the minority sample or upsampling their number, it is still difficult to solve in the setting of contrastive learning setting. Given the scarcity of labeled examples in the real life, exploring how to alleviate this problem in the self-supervised setting is meaningful. In this paper, we analyze the causes and impacts of spurious correlations in contrastive learning and explore methods to mitigate this problem effectively. We focus on enhancing the SPARE framework by identifying and leveraging key parameters that contribute to correct predictions, and then we incorporate network pruning to improve model robustness and efficiency to mitigate spurious correlations effectively without the need for labeled data. Our approach aims to make a balance between simplicity and performance, providing an valuable insight for an robust, explainable and efficient path to address spurious correlations in diverse and complex datasets.

1 Introduction

The problem we want to address in this research is the presence of spurious correlations in contrastive learning. In contrastive learning, we teach the model to differentiate between similar and dissimilar data points by minimizing the distance between positive examples and maximizing the distance between negative examples. Specifically, as discussed in Lecture #11 of CS 260D Mirzasoleiman [2024b], contrastive learning learns an encoder by pulling augmentations from the same example closer together while pushing augmentations from different examples further apart. The contrastive loss $\ell_{i,j}$ is defined as

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

However, models often conclude undesirable correlations between two variables due to unseen confounding variables, and this phenomenon is called spurious correlations.

A well-known example is when a model is shown images of waterbirds in water backgrounds and landbirds in land backgrounds. The model may falsely conclude that there is a strong correlation between "waterbird" and "water background" or "landbird" and "land background". As a result, if the model encounters a waterbird in a land background, its accuracy drops significantly because it relies on the spurious correlation rather than learning the actual features that define a waterbird or a landbird. As a result, addressing spurious correlations in contrastive learning is essential for improving the model's robustness and generalization. By ensuring the model focuses on task-relevant features rather than confounding variables, we can enhance the model performance on diverse datasets and mitigate the risks of low accuracy when encountering novel scenarios.

This topic is of great significance given the high frequency of correlated features in the machine learning and the rarity of labeled data in the real life. In real-world datasets, there are many potential false biases that can mislead machine learning models. Understanding and addressing these spurious correlations is crucial to ensuring that models rely on meaningful and accurate features rather than misleading biases, which not only improves their robustness but also enhances their ability to generalize effectively across new, diverse data. On the other hand, most data in real life are unlabeled and collecting labels would consume a great amount of time. Hence, exploring methods to deal with this problem under self-supervised setting is both practical and necessary. Thus, this research showcases significant values for advancing machine learning.

2 Related Work

Addressing spurious correlations in contrastive learning is difficult because each dataset is unique, making it hard to develop a generalized method that enables the model to identify confounding variables in each dataset. Naive approaches such as simply increasing the size of the dataset or adding random noise have been explored to mitigate the negative impacts of spurious correlations Yang et al. [2022]. While they can mitigate the spurious correlations to a certain degree, these methods fail to address the root cause, which is to eliminate the false correlations entirely. To overcome this issue, supervised learning methods have been proposed, where models are guided by labeled data to identify and mitigate spurious patterns. While these methods are effective, they can not perform well in unsupervised settings: for example, contrastive learning does not require explicit labels when being trained, and thus when being supervised, it is challenging to find out the spurious patterns Lin et al. [2022].

More recent research has explored advanced frameworks for mitigating challenging spurious correlations. For example, the SpuCo framework conducts a systematic evaluation of state-of-the-art algorithms that focus on balancing group distributions and leveraging group information to improve performance in the presence of spurious correlations. By evaluating methods such as Group Balancing (GB) and GroupDRO (GDRO), SpuCo highlights the effectiveness of mitigating disparities across underrepresented groups and obtaining higher worst-group accuracy compared to prior methods Joshi et al. [2024]. Additionally, the Correct-N-Contrast method is a two-stage contrastive approach aiming to improve robustness against spurious correlations by pushing representations for samples with the same class but different spurious attributes together, while pulling those with different classes and the same spurious attribute apart Zhang et al. [2022]. Both works illustrate the ongoing progress in achieving effective generalization in the presence of spurious features. However, they also admit the limitations of current methods in fully addressing the spurious features in diverse, complex datasets.

By building on these advanced works, our research aims to explore further improvement by addressing the limitations of existing techniques in handling underrepresented groups and achieving robust performance across heterogeneous datasets.

3 Problem Formulation

The problem we identified is that oftentimes there are undesirable correlations between class-irrelevant features and labels in datasets, which results in poor worst-group accuracy or even test accuracy. Moreover, given the unlabeled nature, existing contrastive learning frameworks often fail to identify and mitigate spurious correlations effectively, especially in complex, diverse datasets. Therefore, we want to identify and mitigate spurious correlations in contrastive learning frameworks, especially in unsupervised settings, where explicit labels for spurious features are not available.

Our research aims to make an improvement in three aspects: robustness, generalization, and explainability. The core target is to ensure the model focuses on the meaningful patterns in the data instead of relying on irrelevant features. Besides, as computation cost becomes a practical problem in real world application, we aim to achieve better generalization without significantly increasing computational complexity. Additionally, it would benefit further research to understand how the model learns and how specific parameters contribute to predictions in the presence of spurious correlations. By addressing these challenges, we hope to develop a framework that can not only mitigate spurious correlations but also strengthen the overall reliability and interpretability of contrastive learning

models. Specifically, we identify and leverage the effect of the parameters in the state-of-the-art SPARE framework that play an important role in making the right prediction. By incorporating techniques such as network pruning, our goal is to identify and mitigate spurious correlations effectively without relying on labeled data. Key research questions include investigating the impact of spurious correlations on generalization and exploring parameter-based methods to enhance robustness and model reliability.

4 Proposed Method

While there are a lot of existing methods to mitigate the spurious correlations, they often relies heavily on labeled data, which lacks robustness and fails to generalize effectively across complex and diverse datasets. To address these issues, we propose an enhanced approach that incorporates network pruning into the SPARE framework to boost the model performance and interpretablity.

Network pruning is a technique that removes network connections that are considered unimportant to keep the network performance unchanged Mirzasoleiman [2024a]. By reducing the number of parameters, network pruning can improve model efficiency and effectiveness. Types of pruning include weight pruning, group pruning, kernel pruning, and filter pruning Mirzasoleiman [2024a]. Studies show that many machine learning model replies on last layer to make prediction. Thus, we assume the parameters in the last layer hold the same importance for wrongly learning the spurious correlations. To better explore which part of parameters play an important role in studying correlated features, we applies network pruning technique. To improve the model's performance and gain the method's explainability, we apply network pruning strategy with three versions in two scenarios.

Three Versions of Last Layer Network Pruning: To systematically study which part of weights matters in the training stage under the unsupervised setting, we added a customized mask during the training and evaluation stage. Given the weights are are to interpret, we applied absolute rank to evaluate the weights and create three masks: low ranked weights, middle ranked weights, and high ranked weights. All masks set the masked value to 1.

Two Scenarios of Last Layer Network Pruning: We applied masks to the last layer of the backbone model at two stages: training and evaluation. During training, three types of masks were used to study the effect of pruning on learning spurious correlations. During evaluation, masks were applied to identify which weights contribute most to learning spurious correlations.

This comparison aims to identify the most effective pruning strategy for improving worst-group accuracy and overall model performance across varying levels of spurious correlations.

5 Experiments

To conduct the experiment, we use the dataset **Spuco_Minst**, which includes digits with colored backgrounds. The spurious feature is the colored background. This dataset contains six level of difficulties in two dimensions of difficulty. In magnitude level, there are MAGNITUDE_LARGE, MAGNITUDE_MEDIUM, and MAGNITUDE_SMALL, each represents the area of colored background. In the variance level, there are VARIANCE_LOW, VARIANCE_MEDIUM, VARIANCE_HIGH, each presents the categories of features. By exploring our methods on these settings, we could easily identify the inefficiency of current methods and figure out the effect of our method. We chose eight baseline models and conduct experiments with MAGNITUDE_LARGE difficulty using their default parameters.

5.1 Results and Findings

5.1.1 Baseline Comparison

Even though most methods achieved high average accuracy (ranging from 85.42% for EIIL to 99.72% for Group Balancing), their performance on the worse group accuracy was relatively poor: The unsupervised learning methods **EIIL**, **Cluster**, **CNC**, **JTT**, and **SPARE** perform a bad performance in worst group accuracy when the dataset is very difficult, which aligns with our assumption; The methods **Group Balancing (GB)**, **GroupDRO** and **SSA** leverages the power of labels achieves much better performance.

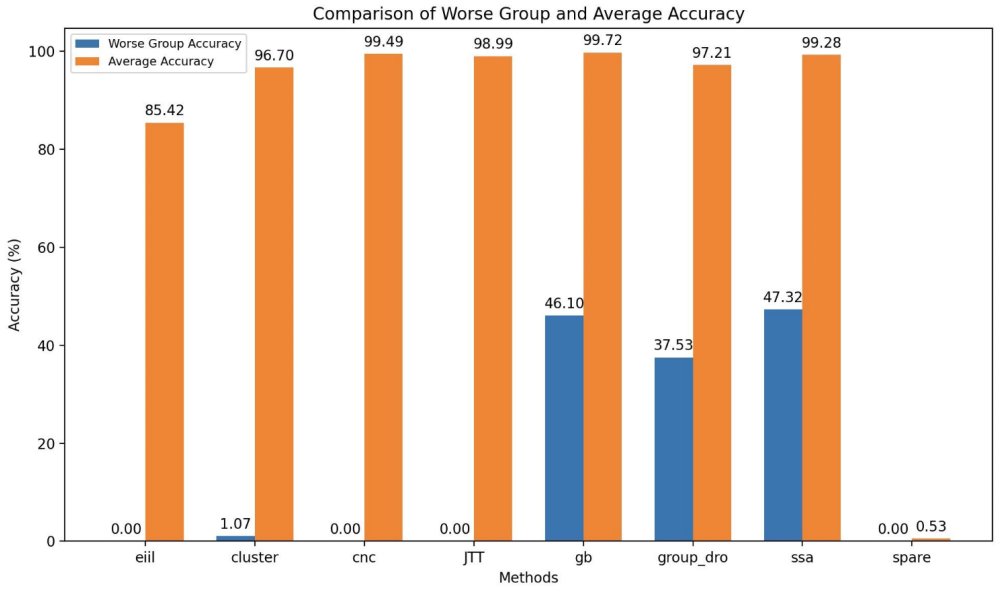

Figure 1: Comparison of worse group accuracy and average accuracy across different methods: EIIL, Cluster, CNC, JTT, Group Balancing (GB), GroupDRO, SSA, and SPARE. The blue bars represent the worse group accuracy, while the orange bars represent the average accuracy

These results align with our assumption and prove that the spurious correlation is even more severe in unsupervised learning methods. The poor performance of SPARE under the default parameter settings also raises our interest to explore its performance for other difficulties of the dataset.

5.1.2 Ablation Study of SPARE across Different Difficulty Levels of Spuco_Minst

To get a fair comparison, we fine-tuned model parameters before conducting ablation studies. Besides, we also take early stop metrics, which will stop training when performance getting downgraded. As shown in Figure 2, SPARE performs robust in most datasets except when magnitude is large. There are two observations worth our attention. The first one is the earliest stop worst group accuracy is always higher than normal worst group accuracy, which means there may exist the problem of overfitting. Another observation is that **SPARE** performs relatively robust in different variance levels but performs bad when magnitude is large.

5.1.3 Method Comparison with Fine-tuned SPARE

In this section, we formally compare our network pruning strategy with six variants with the fine-tuned SPARE in two difficulty levels. We want to explore the effect of our methods on easy and hard datasets, based on the experiment in the last section, for SPARE, the most difficult dataset is when magnitude equal to large and the easiest one is when variance is low.

Before conducting comparisons, we further analyze the parameters for better performance. Figure 3 shows the influence of batch size for SPARE, we choose batch_size = 128 for the following experiments. For fair comparison in the last section, we adopt batch_size = 64 since the performance is closer to default settings and most parts are good enough.

MAGNITUDE_LARGE

The result when difficulty is MAGNITUDE_LARGE is shown in Figure 4. For the hardest dataset, overfitting problem arises since the phenomenon that early stop performance outperforms overall performance still exist. As shown in the left part of figure 4, networking pruning in the training stage brings us some supersizing findings. The weights in middle and high weights contribute together for overall accuracy and worst group accuracy. However, the low ranking of weights contribute to overall accuracy but may lower worst group accuracy. Network pruning in the evaluation stage

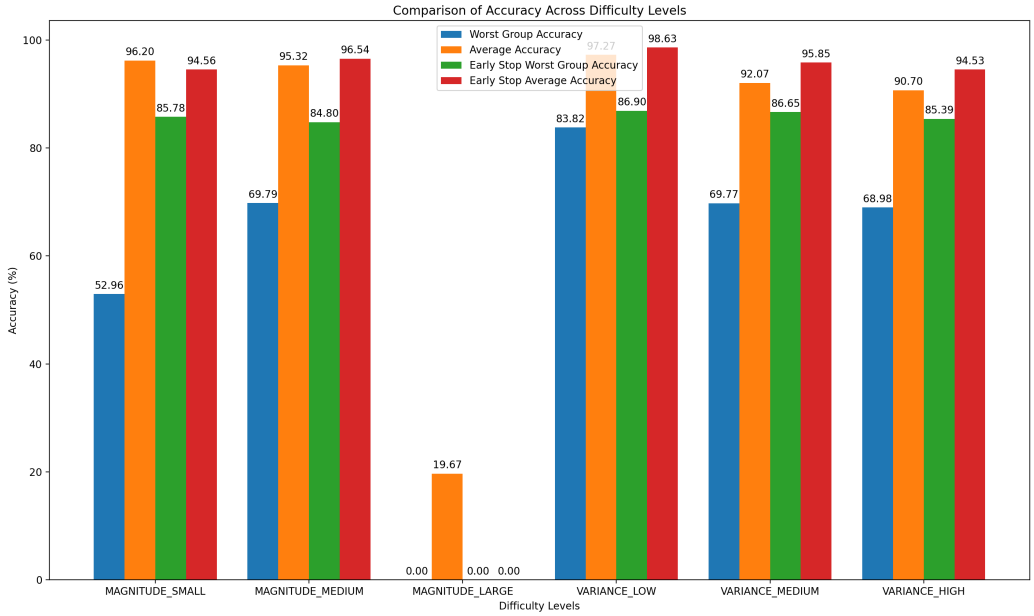

Figure 2: Comparison of Accuracy Across Difficulty Levels. This bar chart compares the performance metrics across different difficulty levels (e.g., *Magnitude Small*, *Magnitude Medium*, *Magnitude Large*, *Variance Low*, *Variance Medium*, and *Variance High*). Metrics include *Worst Group Accuracy*, *Average Accuracy*, *Worst Group Early Stop Accuracy*, and *Early Stop Average Accuracy*, all represented as percentages.

batch_size	worst_group_accuracy	average_accuracy	early_stopping_worst_group_accuracy	early_stopping_average_accuracy
64	0	20.15	0	18.77
128	0	63.41	30.67	77.11
256	0	40.22	0	33.82

Figure 3: Ablation study on batch size

does not perform as thought as well. However, some interesting findings may worth our attention. While this strategy does not improve spurious correlation metrics, it improves average accuracy. We could infer that 1. parameters across different rankings lead to overfit, and most of them centers at the middle ranking weights; 2. parameters across different rankings contribute together to learn spurious correlation.

VARIANCE_LOW

Figure 5 shows with a similar structure with Figure 4, but the conclusion is different than the hardest version of dataset.

In the training stage, we could see masking does not place too much difference in the overall accuracy. However, it somewhat influences worst group accuracy. The middle ranking weights have a higher influence over worst group accuracy while the high ranking weights barely have no influence.

In the evaluation stage, weights in different ranking has shown different importance in both overall accuracy and worst group accuracy. By the gap between masked version and baseline version, we could safely infer that the low ranking parameters matter the most, especially for learning spurious correlations correctly, while the middle weights do not contain too much useful information.

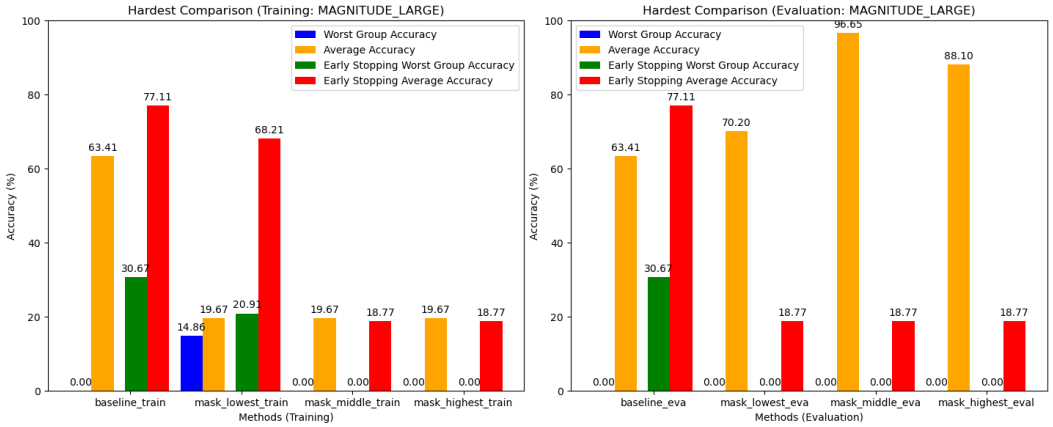

Figure 4: Hardest Comparison (Training and Evaluation): MAGNITUDE_LARGE. This chart compares Worst Group Accuracy, Average Accuracy, Early Stopping Worst Group Accuracy, and Early Stopping Average Accuracy for different training and evaluation methods.

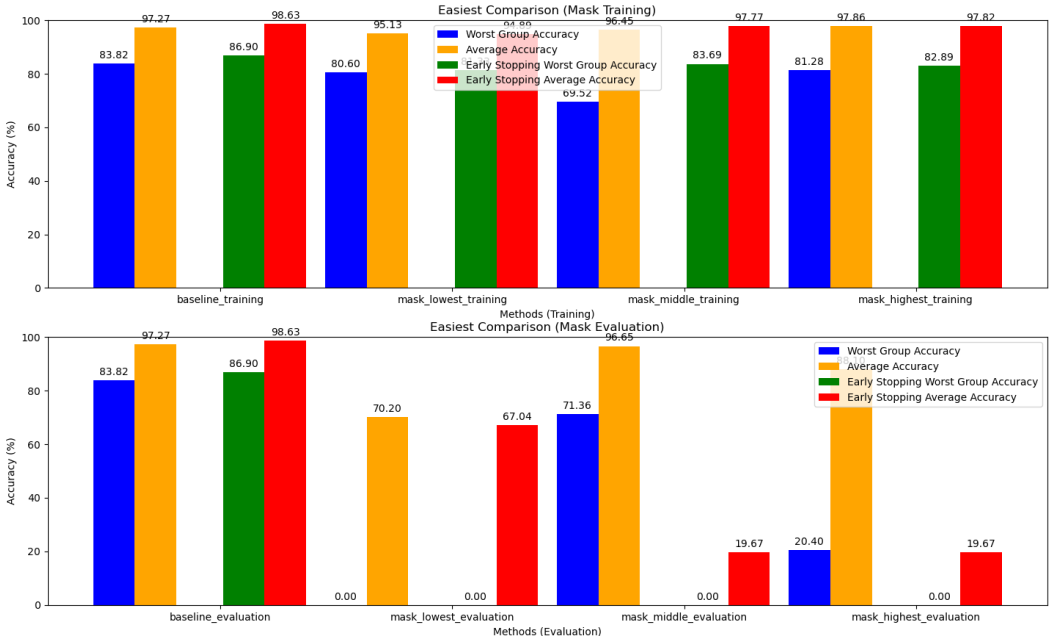

Figure 5: Easiest Comparison (Mask Training and Evaluation). This chart compares Worst Group Accuracy, Average Accuracy, Early Stopping Worst Group Accuracy, and Early Stopping Average Accuracy across different training and evaluation methods.

6 Conclusion

Based on our experiments, it is safe to conclude that network Pruning could be used to explore the effectiveness of model parameters. By applying network pruning strategy with three scales of ranking in two different stages, we found that, for hardest version of dataset, parameters across different scales of ranking tend to work together to learn features and distinguish core features with spurious features, and their impact on overall accuracy and worst group accuracy could be in conflict. While in easiest version, different scales of weights tend to have different degrees of influences over both overall accuracy and worst group accuracy. These findings provide an interesting direction for following researchers to explore under the setting of unsupervised learning.

In general, our work makes use of fewer parameters during training and evaluation, and provides insights for explainability of how parameters contribute to accuracy.

A Appendix A: Future Directions

To extend this work in the future, we aim to apply customized masking strategies, explore the influence of network pruning across different layers, and elaborate on the influence of our method across all levels of datasets.

While some parameter rankings seem more important, no single scale of weights can determine performance in all cases. Therefore, a more flexible and dynamic pruning strategy could help improve interpretability in contrastive learning settings.

Moreover, although studies show that the last layer plays the most critical role in prediction, the gap between full-epoch accuracy and early-stop accuracy suggests that some middle layers may have significant impacts. Future experiments on all six levels of datasets and additional datasets could provide deeper and more systematic insights into the general framework of parameter importance.

B Appendix B: Introduction to the Baseline Models

Group Balancing (GB) and GroupDRO (GROUP_DRO) utilize group-labeled training data, while Just Train Twice (JTT), Cluster, Correct-n-Contrast (CNC), Environment Inference for Invariant Learning (EIIL), and Separate Early and Resample (SPARE) do not. Spread Spurious Attributes (SSA) uses the Group-Labeled Validation Set. EIIL leverages logits and labels from a pre-trained model to infer group partitions that maximize loss variance, enabling robust group-wise evaluation.

Cluster identifies patterns and groups similar data points based on certain features. CNC is a contrastive approach that directly learns representations robust to spurious correlations Zhang et al. [2022]. JTT, as its name suggests, involves first training a standard model (ERM) to identify misclassified data points, followed by retraining a new model that upweights ("upsamples") these difficult examples during a second training phase.

Group Balancing samples every mini-batch to ensure equal representation from all groups Joshi et al. [2024]. GroupDRO (GDRO) samples group-balanced training batches and minimizes the empirical worst-group training loss Joshi et al. [2024]:

$$\mathbf{w} \in \arg\min_{\mathbf{w}} \max_{g \in \mathcal{G}} \mathbb{E}_{(x_i, y_i) \in g} \left[\ell(f(\mathbf{w}, \mathbf{x}_i), y_i) \right].$$

SSA uses group-labeled validation data to train a group label predictor in a semi-supervised manner Joshi et al. [2024]. SPARE clusters the neural network's output early in training and leverages inferred group information to train a robust model with balanced group sampling, mitigating spurious correlations Joshi et al. [2024].

References

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Spuco: A framework for spurious correlation discovery and mitigation. *Journal of Machine Learning Research*, 25:1–25, 2024.

Xinyu Lin, Yiyan Xu, Wenjie Wang, Yang Zhang, and Fuli Feng. Mitigating spurious correlations for self-supervised recommendation. *arXiv preprint arXiv:2212.04282*, 2022.

Baharan Mirzasoleiman. Lecture #13: Neural network pruning. 2024a. CS 260D Large-scale Machine Learning, Lecture Notes.

Baharan Mirzasoleiman. Lecture #11: Data selection for machine learning. Lecture Slides, CS 260D, UCLA, 2024b. Available upon request or distributed in class.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. *arXiv preprint arXiv:2202.05189*, 2022.

Michael Zhang, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. 2022. `https://cs.stanford.edu/~mzhang/correctncontrast.pdf`.