Effect of Neural Network Purning on Spurious Correlation

Muskan Rizwan Shaikh
muskanrs@ucla.edu

Syam Sundar Kirubakaran
syamk@ucla.edu

Akshat Mehta
akshatmehta@ucla.edu

Soham Kulkarni
sohamkulkarni@ucla.edu

Department of Computer Science, UCLA, Los Angeles, CA, 90024

Abstract

One of the major factors that affect the accuracy of a large neural network while training on data at scale, is that, it learns a lot of undesirable correlation between a class irrelevant features and class label. Sometimes these models become so huge that training and inference time becomes extremely high. So, it's extremely important to address both of these major concerns to create a robust neural network. Therefore, we propose to understand the effect of neural network pruning on model trained on spuriously correlated dataset(dataset that has undesirable features) to observe performance fluctuations through various experiments both on supervised and contrastive learning settings.

Code: `https://github.com/Muskan-R-S/Large-Scale-ML-Project` Zip: `https://syam.work/assets/ucla/260d/final_project.zip`

1 Introduction

The implementation of machine learning in real-world scenarios faces a lot of challenges, including data quality, feature selection, resource management, data availability, and the delicate balance between computational efficiency and model accuracy. To ensure the efficient deployment of machine learning models we need to explore strategies for optimizing the resources. It is crucial to find the right balance to make sure that we spend less on computing and still keep the accuracy of the machine learning models high. One intriguing challenge in machine learning is the presence of spurious correlations in datasets. Spurious correlations refers to the undesirable correlation between a class irrelevant feature and a label. Deep neural networks can be negatively impacted by these correlations present in the data, a standard ERM model when trained on a dataset that has spurious correlations could learn to classify examples based on the spurious feature instead of learning the important features. This could result in a high average accuracy of the model, but a low worst group accuracy. We aim to study through out project how model optimization techniques like the neural network pruning impacts the accuracy of the model when there are spurious correlations present in the dataset. Neural network pruning aims to improve the efficiency of a neural network by reducing its size and complexity. It involves identifying and removing certain components from the neural network such as weights, neurons, or filters, to create a more compact model without significantly sacrificing performance. This is important when we want to use models in scenarios where memory resources are constrained, like the edge devices or mobile applications.

Preprint. Under review.

2 Related Work

The field of machine learning has witnessed extensive research in mitigating the impact of spurious correlations and improving the robustness of learned representations across various domains. Addressing the challenge of spurious correlations has been a focal point in recent research due to their detrimental effects on model generalization. A range of techniques has been developed, including data reweighting, adversarial training, and specialized training strategies, as evidenced in the following studies: Studies Joshi et al. [2023] Youbi Idrissi et al. [2021] have emphasized the vulnerabilities of supervised learning models to spurious correlations. Supervised learning models tend to exploit features that exhibit correlations with the target variable but may lack causal relevance, leading to biased predictions, especially concerning minority subgroups or specific contexts.

Joshi et al. [2023] delves into the repercussions of spurious features in supervised learning models, highlighting their impact on model fairness and performance across different groups. This study investigates the ways in which neural networks could amplify biased decision-making based on spurious correlations.

Adversarial training methods have been explored to enhance model robustness against spurious correlations Kirichenko et al. [2022]. These approaches involve incorporating adversarial examples during model training to encourage the learning of more generalized representations less prone to over-relying on spurious correlations.

Additionally, specialized training strategies have been developed, such as worst-group-accuracy optimization Youbi Idrissi et al. [2021], aiming to build classifiers robust to worst-case performance across different data groups. This research explores the efficacy of simple data balancing techniques in enhancing worst-group-accuracy, emphasizing the significance of group information in model selection and training.

Recent interest in self-supervised learning (SSL) methods Cadet et al. [2023] has raised questions about how such techniques handle spurious correlations in learned representations. Self-supervised learning aims to extract meaningful representations from unlabeled data, but concerns persist about the potential entanglement of spurious features in learned representations. The study reveals that SSL models exhibit biases towards simpler features that may be spuriously correlated, leading to disparities in the captured representations. The proposed method endeavors to mitigate spurious information from SSL-learned representations without requiring explicit group or label information, ultimately enhancing the model's performance on downstream tasks.

Some of the recent works show that the penultimate layer is important from a pruning point of view in supervised learning settings when we are working with spurious datasets. For e.g., Lee et al. [2023] shows that pruning the irrelevant neurons in the last layers after deploying an attention model on images (to learn model predictions) led to increased worst-group accuracy. Idrissi et al. [2022]show that simple data balancing can achieve S.O.T.A. accuracy, and access to group info is very critical during model selection, for improving worst-group accuracy.

However, despite these contributions, a gap persists in understanding how pruning affects data with spurious correlation or data imbalance. This study seeks to bridge this gap by performing an analytical analysis of the causal relationship between pruning of different layers of a neural network. We aim to see how we can further extrapolate the results to other models.

3 Problem Formulation

The overarching aim is to investigate how the intentional reduction of network parameters through pruning techniques influences the occurrence and mitigation of spurious correlations—a phenomenon where the model erroneously identifies patterns that do not reflect genuine underlying relationships in the data. This research seeks to address the pressing need for efficient and interpretable neural networks by assessing the trade-off between model complexity and the emergence of spurious correlations. The formulation involves defining specific metrics for spurious correlation, devising appropriate pruning methodologies, and establishing a comprehensive experimental framework to rigorously evaluate the network's performance in the presence of varying degrees of pruning. This study aspires to contribute valuable insights into the delicate balance between model optimization and the potential pitfalls of spurious correlation in neural networks.

4 Method

To study the effects neural network pruning on the spurious correlations we first perform supervised learning on a dataset that has high spurious correlations and then perform unstructrured magnitude pruning, where we remove a percentage of the lowest weights. We prune each layer one by one and see how this effects the accuracies. We then compare these results to contrastive learning by following a very similar approach of training on spurious dataset and observing the change in accuracy as we prune different layers of the network.

4.1 Supervised Learning:

Dataset Preparation:

The experiment utilizes the SpuCo MNIST dataset for the training, validation, and testing processes. SpuCo MNIST contains images categorized into different classes to simulate spurious feature correlations.

Model: To perform supervised learning on the SpuCo MNIST dataset, the model SpuCoModel LeNet was selected.

Model Training Strategies: To get a better worst group accuracy we trained using the Evan Zheran Liu [2021] JTT method. The training happens in the following sequence. First we train via standard ERM, then construct the error set of training examples that are misclassified by the ERM. After which we upsample the examples in the error set. Finally we train the final model on a dataset including these upsampled examples via ERM.

Model Evaluation:

The evaluation phase includes the assessment of model performance using a separate validation dataset. Evaluators are used to measure model accuracy, particularly focusing on worst-group accuracy, average accuracy, and prediction accuracy related to spurious attributes.

Model Duplication and Pruning:

The study duplicates the trained models and performs pruning experiments to understand the impact of pruning on model performance. Pruning is conducted on different layers of the models, varying the pruning amounts to evaluate the resulting model's worst-group accuracy, average accuracy, and the model's predictive capability regarding spurious attributes.

4.2 Contrastive Learning:

Dataset Preparation:

The experiment utilizes the a custom MNIST dataset for the pretraining, training and validation. The dataset is created by adding colors(yellow, red, green and blue) to the white tints of MNIST dataset. Samples from the custom MNIST dataset is shown below. It's also important to note that we've added augmentations such as Horizontal Flip and Random Crop which are required in the context of contrastive learning.

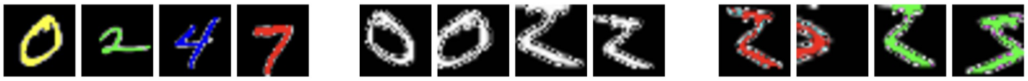

Figure 1: (1) Spurious MNIST dataset ; (2) MNIST with Augmentations ; (3) Spurious MNIST with Augmentations (Horizontal Flip and Random Crop)

Approach:

We obtain the MNIST data and add spurious features followed by performing the required augmentations as shown. The contrastive learning model has two main components: (1) Encoder: Uses an Multi-layer Perceptron that has 3 Convoluational Layers. (2) Projection Head: Has a linear layer that applies a linear transformation to the incoming data.

First, we perform pretraining in which we perform 2 augmentations of the same image which is formally known as TwoCropTransform and then contrast one with respect to the other augmentation. This is repeated for all the dataset entries. We formulate the contrastive objective function that encourages similar instances to have similar embeddings and dissimilar instances to have different embeddings. In other words, similar images are clustered together and pushed away from dissimilar images. We then perform linear probing to distinctly separate out clusters from each other by drawing linear boundaries.

Notice that we store a snapshot of the model along with all its configuration and weights after pretraining phase so that it can be reused for multiple linear probes while pruning on different layers and with different prune probabilities.

5 Experiments

Evaluation Metrics:

Let us look at the evaluation metrics used in the experiments in further detail:

1. Worst-Group Accuracy: Worst-group accuracy assesses how well the model performs on the group of data that does not have the spurious attribute. A lower worst-group accuracy suggests that the model struggles with these challenging subgroups, indicating susceptibility to the effects of spurious attributes.
2. Average Accuracy: The average accuracy provides an overall view of the model's performance across all classes or groups within the dataset. A higher average accuracy indicates better generalization and proficiency in handling various classes or groups.
3. Predictive Capability for Spurious Attributes: This evaluates the accuracy of the model in predicting the spurious attribute. A low accuracy in predicting the spurious attribute indicates that the model does not rely on the spurious correlation for making predictions. This metric allows us to comment on the fairness of the model.

Experiment Settings:

1. All Layers Pruning Experiment: Pruning experiments are carried out across all layers of the model by adjusting the pruning amount iteratively from 0 to 1. The resulting model's accuracy metrics are recorded at each pruning interval.
2. Layer-Specific Pruning Experiments: Layer-specific pruning experiments are conducted individually on each layer of the model. Similar to the previous experiment, pruning amounts are adjusted incrementally, and the models' performance is assessed at different intervals.

Model Architecture:

Model architectures for supervised and contrastive is given as follows: (respectively)

Layer	Nomenclature
Convolution Layer 1	Layer 1
Convolution Layer 2	Layer 2
Fully Connected Layer 1	Layer 3
Fully Connected Layer 2	Layer 4

Layer	Nomenclature
Encoder::Conv1	Layer 1
Encoder::Conv2	Layer 2
Encoder::Conv3	Layer 3
ProjectionHead::Linear1	Layer 4
ProjectionHead::Linear2	Layer 5

5.1 Supervised Learning:

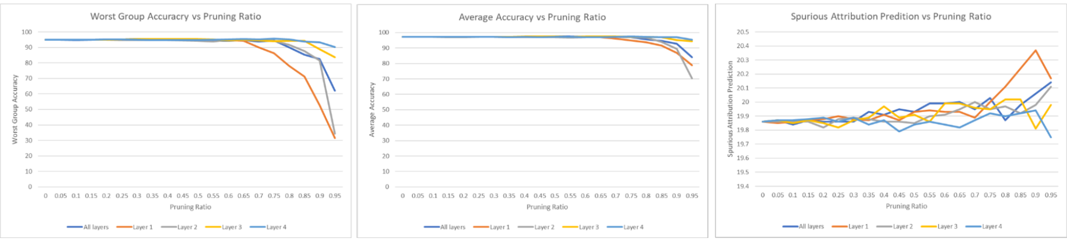

Figure 2: Spurious Magnitude - Small; ERM - Yes; JTT - Yes; Pruning - 0 to 1

We can notice that interesting things happens on purning at larger magnitude. Post 60%, we can also notice a drop in accuracy.

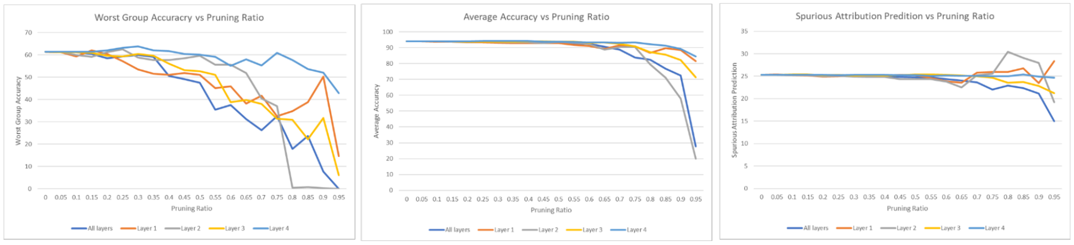

Figure 3: Spurious Magnitude - Large; ERM - Yes; JTT - Yes; Pruning - 0 to 1

We can notice that on increasing the magnitude of spurious correlations, accuracy doesn't remain stable and of all the layers, the last layer produces the most promising results both in the case of worst group accuracy and average accuracy.

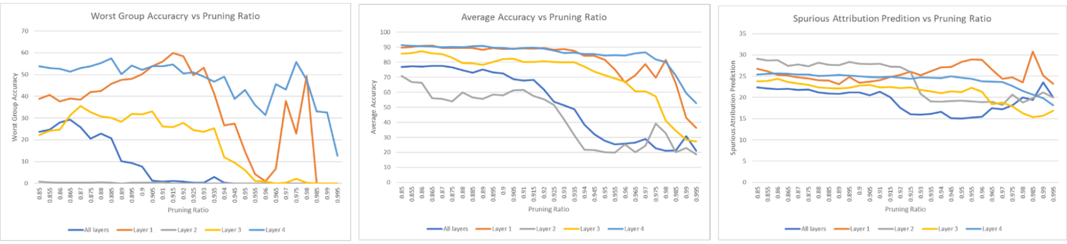

Figure 4: Spurious Magnitude - Large; ERM - Yes; JTT - Yes; Pruning- 0.85 to 1

We can notice that even with higher pruning ratio, last layer produces the best result compared to pruning other layers.

Observation Verdict:

1. Through our experiments, we observed that the model accuracies remain relatively stable until a substantial pruning threshold is reached. Specifically, we observed a noticeable decline in accuracies when the pruning percentage exceeded about 70 percent.
2. Overall, Our experiments show that pruning from the last layer yields better results and the worst group accuracy does drops the least as compared to the other layers, after substantial pruning. This outcome suggests a viable strategy: prioritizing more aggressive pruning for the last layer while applying more conservative pruning to other layers, achieving a favorable balance without compromising worst group accuracies significantly.

5.2 Contrastive Learning:

First, we perform contrastive learning on plain MNIST dataset that has no spurious features and you can evidently find that validation accuracy is always higher than training and this is not the case when we learn from spurious MNIST dataset, which means that spurious features are learnt heavily than the core features.

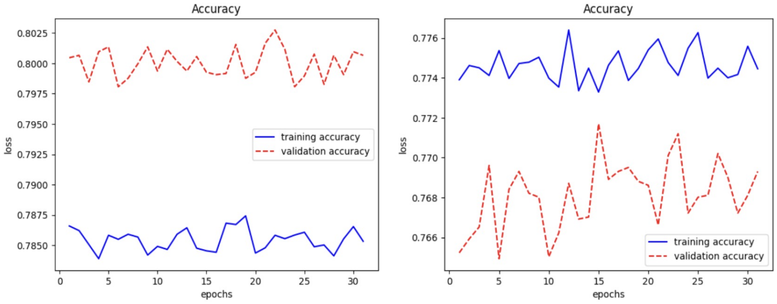

Figure 5: MNIST vs Spurious MNIST Accuracies (unpruned)

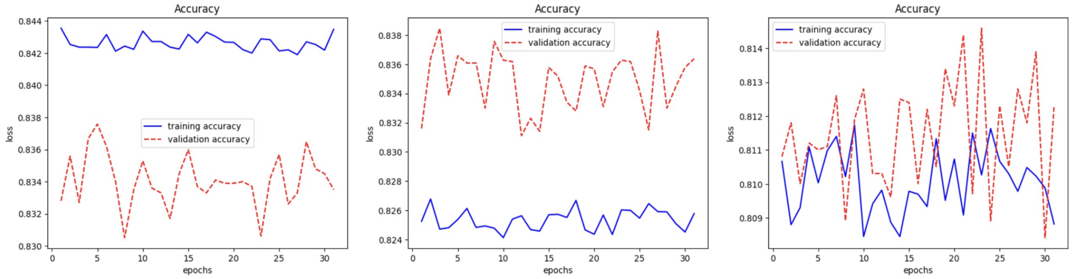

Figure 6: Prune different Layers (layer 1, 2 and 3) of the model trained on Spurious MNIST with 0.6 magnitude

While keeping the magnitude of pruning at 0.6 and pruning different layers (Layer 1, Layer 2 and Layer 3), we observe the following:

1. Pruning the Layer 1: This has no significant effect as the graph look similar to the graph in Fig 6 but it is important to note that the accuracy range of validation is high than that of the one in Fig 6.
2. Pruning the Layer 2: This pushes the validation accuracy above the training accuracy which is desirable. We also notice the highest validation accuracy here when pruning layer 2 which is approx. 83%
3. Pruning the Layer 3: The validation and training accuracies are brought close to each other and overlap in many instances. Also, notice that the range of accuracies in the graph is lowest compared to the previous layer pruning.

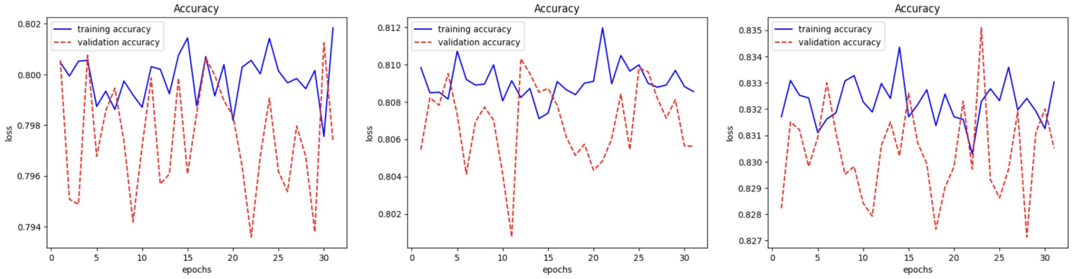

Figure 7: Prune different Layers (layer 1, 2 and 3) of the model trained on Spurious MNIST with 0.9 magnitude

While keeping the magnitude of pruning at 0.9 and pruning different layers (Layer 1, Layer 2 and Layer 3), we can observe that in all three cases, the training and validation accuracies are close to each other but the maximum validation accuracy is obtained while pruning the last layer which is approximately 83.5%.

Observation Verdict: On purning different layers with higher magnitude we can observe that pruning the last layer produces the maximum validation accuracy where as with lower magnitude we can observe that Layer 2 (middle layer) produces the maximum accuracy in the context of contrastive learning.

6 Conclusion

In conclusion, the study on the "Effect of Neural Network Pruning on Spurious Correlation" reveals the nuanced relationship between pruning techniques and the spurious correlations. While neural network pruning proves effective in reducing redundancy and enhancing efficiency, the extent and methodology of pruning significantly influence the model's susceptibility to false associations. We were able to concretely arrive at a conclusion that pruning at higher ratios on the last layer of the network produces the most effective results both in terms of supervised and contrastive learning.

7 Contributions:

1. Muskan Rizwan Shaikh: Project Ideation, Presentation, Code, Experiments and Write-up: Overall Supervised Learning, Introduction, Methodology, and Results.
2. Syam Sundar Kirubakaran: Project Ideation, Presentation, Code, Experiments and Write-up: Overall Contrastive Learning, Abstract, Methodology, and Results.
3. Akshat Mehta: Project Ideation, Presentation, Code and running experiments, Write-up: Related work, Methodology, Experiment, and results.
4. Soham Kulkarni: Project Ideation, Presentation, tried adaptive pruning code in Supervised Learning setting, Write-up: Literature Review/ Related Work

References

Xavier Cadet, Ranya Aloufi, Alain Miranville, Sara Ahmadi-Abhari, and Hamed Haddadi. Evaluating the robustness of self-supervised representations to background/foreground removal, 06 2023.

Annie S. Chen Aditi Raghunathan Pang Wei Koh Shiori Sagawa Percy Liang Chelsea Finn Evan Zheran Liu, Behzad Haghgoo. Just train twice: Improving group robustness without training group information, 09 2021.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark a more realistic dataset, 06 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 04 2022.

Seongmin Lee, Ali Payani, and Duen Horng Chau. Towards mitigating spurious correlations in image classifiers with simple yes-no feedback. 2023. URL https://api.semanticscholar.org/CorpusID:259941195.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy, 10 2021.