SEVERING SPURIOUS CORRELATIONS WITH DATA PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have been shown to learn and rely on spurious correlations present in the data that they are trained on. Reliance on such correlations can cause these networks to malfunction when deployed in the real world, where these correlations may no longer hold. To overcome the formation of such correlations, recent studies propose approaches that yield promising results. These works, however, study settings where the strength of the spurious signal is significantly greater than that of the core, invariant signal, making it easier to detect the presence of spurious features in individual training samples and allow for further processing. In this paper, we identify new settings where the strength of the spurious signal is relatively weaker, making it difficult to detect any spurious information while continuing to have catastrophic consequences. We also learn that spurious correlations are formed primarily due to only a handful of all the samples containing the spurious feature and develop a novel data pruning technique that identifies and prunes small subsets of the training data that contain these samples. Our proposed technique does not require information regarding the sample-wise presence or nature of spurious information, or human intervention. Finally, we show that such data pruning attains state-of-the-art performance on previously studied settings where spurious information is identifiable.

1 INTRODUCTION

Deep neural networks have shown promising results on a variety of benchmarks and applications. Their reliability and by extension, ability to solve increasingly challenging problems remains questionable as these networks have been shown to exhibit several failure modes in practice. Of these, the inability to adapt to distributional shifts due to the reliance on spurious features is of strong concern as it makes them unreliable for deployment in the real world. Spurious Correlations are correlations that a network forms between simple, weakly predictive spurious features present in a fraction of the training data and the class label. These correlations are problematic as a network may prefer them over strongly predictive invariant correlations when making a prediction. Thus, in the event of a distribution shift where spurious features either no longer exist or become correlated with a different task, which is a common phenomenon in the real world, these networks begin to malfunction (Arjovsky et al., 2019; Sagawa et al., 2020a;b; Nagarajan et al., 2021; Geirhos et al., 2020).

To promote the learning of invariant features, several recent works have proposed techniques that yield promising results. For instance, Sagawa et al. (2020a) aim to up-weight training samples without spurious features, but with the assumption that one would be aware of which samples contain spurious features. Kirichenko et al. (2022) work with the same assumption where they re-train the last layer of a spuriously biased network on an unbiased dataset and obtain state-of-the-art results on popular benchmarks. Moayeri et al. (2023) follow a similar technique of utilizing human supervision to determine the strength of spurious signals in each training sample and fine-tune a biased network on samples where spurious signals are weaker. Attaining such information regarding sample-wise presence of spurious features during training, however, is unrealistic. To overcome this problem, Liu et al. (2021); Zhang et al. (2022); Ahmed et al. (2021); Creager et al. (2021) aim to infer the sample-wise presence of spurious features based on a biased network's outputs and they follow this with further processing such as up-weighting of samples without spurious features or aligning representations of samples with and without spurious features, typically at the cost of overall testing accuracy.

All past works that study spurious correlations and aim to promote invariant learning perform experiments on settings where the strength of the spurious signal is significantly higher than the core, invariant signal. This makes it easy to identify the presence of spurious features in each sample within the training set, which is, however, not always the case. In this paper, we aim to tackle novel settings where the strength of the spurious signal is relatively weaker, making it impossible to identify such information while rendering existing techniques ineffective. We then show that spurious correlations are formed primarily due to a few key samples in the training set and present a novel data pruning technique to identify and prune a small subset of the training that contains these samples. Finally, we show that such data pruning attains state-of-the-art performance on previously studied settings where information regarding the sample-wise presence of spurious features is easily identifiable. We summarize our contributions below:

Contributions

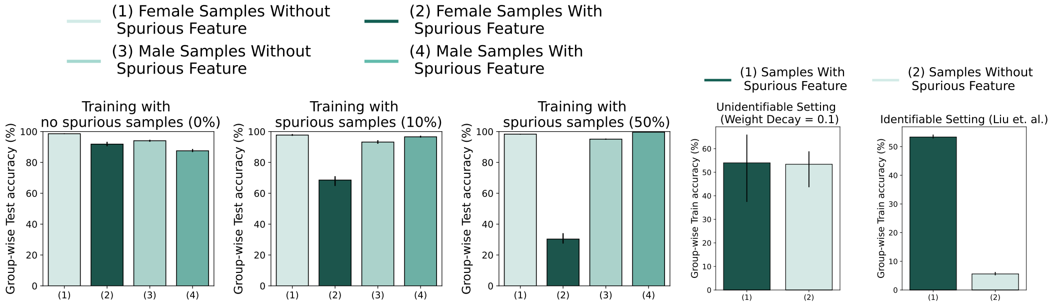
- We identify settings where it is impossible/difficult to identify sample-wise presence of spurious features. This renders past approaches that mitigate spurious correlations as ineffective.
- We discover that spurious correlations are formed primarily due to a handful of all the samples containing spurious features. Based on this insight, we propose a simple and novel data pruning technique that identifies and prunes a small subset of the data that contains these samples.
- We show that such data pruning attains state-of-the-art results even in previously studied settings where information regarding the sample-wise presence of spurious features is identifiable.

2 BACKGROUND AND RELATED WORK

Consistent with past literature, we study the supervised classification setting where $S = \{(x_i, y_i)\}_{i=1}^N$ denotes the training dataset of size N and network is trained to learn a mapping between x_i (input) and y_i (class label) using empirical risk minimization (Vapnik, 1998). Every training sample $s \in S$ contains a core feature (c_i) that represents its class (y_i). A fraction of all samples within a class contain the spurious feature (a_i) associated with that class. Core (or invariant) features are causal to the ground truth label y_i and are fully predictive of the task, as they are present in all samples. Spurious features are not causal to the ground truth labels and are partially predictive of the task, as they are present in only a fraction of all samples of a class.

Spurious Correlations. The correlations a network forms between spurious features and class labels. Such correlations are undesirable as they are not causal to class labels and can disappear during testing or become associated with a different task, causing these networks to malfunction (Arjovsky et al., 2019; Nagarajan et al., 2021; Shah et al., 2020; Tsipras et al., 2019; Sagawa et al., 2020b;a; Kirichenko et al., 2022; Jain et al., 2023; Ye et al., 2022). To make a model robust against spurious correlations, Sagawa et al. (2020a) suggest using information regarding the sample-wise presence of spurious features during training to directly minimize loss on samples without spurious features. Since such information is often unavailable, more recent approaches attempt to infer group information with the help of an ERM trained model, followed by subsequent sample up-weighting or class-wise representational alignment (Sohoni et al., 2020; Liu et al., 2021; Zhang et al., 2021; Ahmed et al., 2021; Creager et al., 2021). Recent work has also shown that a model learns both general and spurious features and just re-training the last layer with balanced data where strong spurious correlations do not hold can improve robustness on data with explicit spuriousness (Kirichenko et al., 2022). Moayeri et al. (2023) perform a similar re-training, where they fine-tune the penultimate layer of a trained model on those samples with minimal spuriosity, where spuriosity is determined with human supervision.

Simplicity Bias and Spurious Correlations. Recent work has shown that deep neural networks have a greater affinity towards simpler features than more complicated ones (Shah et al., 2020; Morwani et al., 2023; Tiwari & Shenoy, 2023). Shah et al. (2020) show that in the presence of two fully predictive features, a network would choose to fully ignore the more complicated features in favor of



(a) Training with just a *few* Male samples with spurious features causes heavy reliance on spurious features, indicated by low test accuracy on Female samples with Spurious Features.

(b) Training with greater regularization may not help identify samples with the spurious feature.

Figure 1: Spurious Information is Often Unattainable.

the simpler set of features. In settings where the simpler feature is not fully predictive (Spurious), the network still relies strongly on these features (Sagawa et al., 2020b; Kirichenko et al., 2022), leading to the formation of spurious correlations.

Feature Difficulty. Consistent with deep learning literature (specifically, those works concerned with spurious correlations), difficulty of learning a feature is determined by the following three factors: (1) Proportion (or Frequency) of training samples containing the spurious feature (Sagawa et al., 2020b; Shah et al., 2020; Kirichenko et al., 2022), (2) Area Occupied and Position (if it is centered or not) in the training sample (Moayeri et al., 2022) and (3) The amount of noise in its signal (Sagawa et al., 2020b; Ye et al., 2022). A feature which is present in a large portion of all training samples, occupies a lot of area, is centered, and has little to no variance, is easy to learn. On the other hand, a feature which is present in a small portion of all training samples, occupies little area, is not centered, and has a lot of noise/variance, is hard to learn.

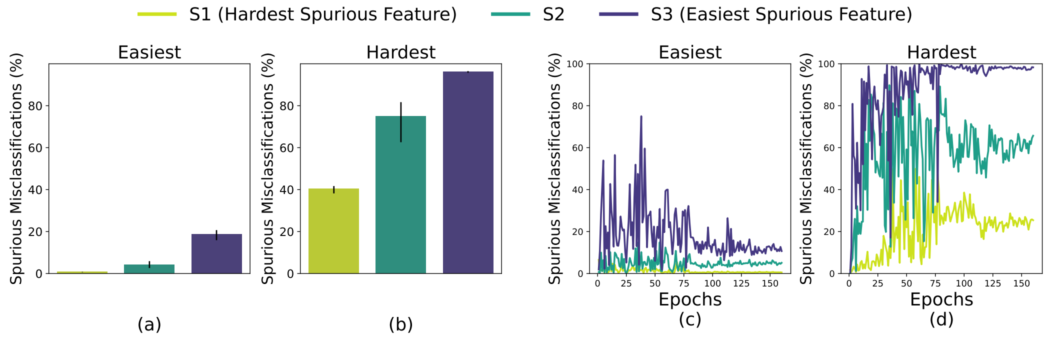
Our work is concerned with two different widely studied areas in deep learning: out-of-distribution generalization and data pruning.

Data Pruning. A large part of deep learning's recent success has been attributed to large datasets that models are trained on. However, with increasing datasets sizes, computational costs have risen significantly. This raises an important question: Are all samples in a dataset equally important for attaining good testing accuracy? In other words, is it possible to reduce samples in the training data without impacting generalizability? Recent work has shown that it is possible to remove a large fraction of a dataset without sacrificing test accuracy. Most approaches create sample-wise metrics that measure how important it is to maintain a sample in the training set and remove the least important samples by their definition (Paul et al., 2021; Toneva et al., 2019; Feldman & Zhang, 2020; Sorscher et al., 2022).

While existing work aims to prune samples to reduce computational costs, our work focuses on pruning samples to mitigate spurious correlations formed during training. This direction, to the best of our knowledge, is *novel*.

3 SPURIOUS INFORMATION IS OFTEN UNATTAINABLE

Suppose one wants to build a gender classifier using the CelebA dataset (Liu et al., 2015). A small fraction of the training samples in the Male class contain eyeglasses (10%), which takes the form of the spurious feature in our setting. None of the samples in the Female class contain eyeglasses in the training set. Thus, the Male class is spuriously correlated with eyeglasses. Note that in our setting, only a small minority of all Male samples contain the spurious feature, making the strength of the spurious signal relatively weaker. This is different from settings studied in literature, where the spurious feature is present in a majority of the samples (often 95-97% of all samples of that class) and the strength of the spurious signal is significantly higher than the strength of the core, invariant signal. We observe that under normal training conditions without the existence of spurious samples, test



(a) (b) (c) (d)

Figure 2: Introducing spurious features in 100 samples with the easiest core features (`Easiest`) causes little to no reliance on spurious features, indicated by low Spurious Misclassifications. Introducing the *same* spurious features in 100 samples with the hardest core features (`Hardest`) causes heavy reliance on spurious features, indicated by high Spurious Misclassifications.

accuracy of Female samples with the spurious feature is high as shown in Fig. 1a (Left). Introduction of just a few Male samples with the spurious feature is capable of reducing the test accuracy of Female samples with eyeglasses significantly, as shown in Fig. 1a (Middle and Right).

In settings where the strength of the spurious signal is significantly greater than the strength of the core, invariant signal, it becomes easy to differentiate between groups of samples containing spurious features and those that do not (Liu et al., 2021; Ahmed et al., 2021; Zhang et al., 2022; Creager et al., 2021; Yang et al., 2024). This is because the network relies very strongly on the spurious features during training, making it easy to differentiate between such groups based on representational differences or on a network's ease of learning these samples. For instance, Liu et al. (2021); Zhang et al. (2022) train a network with high regularization (greater weight decay) such that a network correctly classified samples with easy spurious features during training and misclassifies those without, as shown in Fig. 1b (Right). In our setting, however, the spurious feature is not present in a majority of the training samples and so the strength of the spurious signal is relatively weaker. This makes it impossible to identify which samples contain spurious features using past approaches, even when 50% of all male samples contain eyeglasses - see Fig. 1b (Left). Thus, we identify the primary failure mode of existing approaches as follows:

- *Attaining information regarding the presence of spurious features becomes impossible/difficult when the strength of the spurious signal is not exceptionally greater than the strength of the core, invariant signal in the training set.*

Based on this observation, we propose the following problem statement:

- **Problem Statement:** *How does one sever spurious correlations in settings where attaining spurious information is difficult or impossible?*

4 EXPERIMENTAL DESIGN

In this paper, we study both settings: those where information regarding group information is identifiable and those where such information is unidentifiable. We also move beyond simple binary classification tasks that have been the focus of all recent works that study spurious correlations. Below, we detail the experimental design for this paper. Please note that we provide additional details in the Appendix.

TESTBED

- **CIFAR-10S.** We create a testbed based on the CIFAR-10 dataset (Krizhevsky, 2009) where we synthetically introduce spurious features in a fraction of one class' $(c1)$ training samples. We introduce the same spurious feature in all samples of a different class $(c2)$ during testing.

The degree of spurious feature reliance is estimated by measuring the number of samples of $(c2)$ that are misclassified as $(c1)$ during testing. Since spurious features are synthetically introduced, this setting allows us to:

1. Vary the strength of the spurious feature relative to the core, invariant feature.
2. Accurately compute the difficulty of learning the core, invariant feature of individual training samples before the introduction of spurious features.

Note that in this setting, the spurious feature takes the form of a line running through the center of the images of class $c1$ and we vary the strength of the spurious signal by increasing the region the spurious feature covers or by increasing the proportion of samples that contain the spurious feature.

UNIDENTIFIABLE BENCHMARKS

- **CelebA.** (Liu et al., 2015) We utilize the same experimental setting studied in Sec. 3, where we build a gender classifier in which the Male class is spuriously correlated with `Eyeglasses`.
- **Hard ImageNet.** (Moayeri et al., 2022) The Hard ImageNet dataset is a 15 class classification task where all classes have a spurious feature associated with them and certain classes share similar spurious features. For instance, both `Ski` and `Dog Sled` classes contain similar spurious features (`Snow`, `People`, `Trees`, and `Hills`). In our experiments, we measure the degree of spurious feature reliance by measuring the numbers of `Dog Sled` samples that are misclassified as `Ski` during testing. Note that in our experimental setting, no `Skis` are misclassified as `Dog Sled`.

STANDARD BENCHMARKS (i.e., IDENTIFIABLE)

- **Waterbirds.** (Sagawa et al., 2020a) The Waterbirds task is binary image classification task where the goal is to classify an image of a bird as `landbird` or `waterbird`. In this setting, the class `landbird` is spuriously correlated with `Land` backgrounds while the class `waterbird` is spuriously correlated with `Water` backgrounds.
- **MultiNLI.** (Williams et al., 2018) The MultiNLI task is a classification task with three classes where the goal is to classify the second sentence in a pair of sentences as entailed by, neutral with, or contradicts. Consistent with the setting in Sagawa et al. (2020a), a large fraction of the `contradicts` class contains negation words while the other two only contain a few samples with negation words, making the `contradicts` class spuriously correlated with negation words and the other two with the lack thereof.

EVALUATION.

Current practice in deep learning utilizes Worst-Group Accuracy (WGA) to assess the degree of spurious feature reliance in binary classification tasks. WGA computes the accuracy of test samples that contain the spurious feature associated with the other class during training. While suitable for simple binary classification tasks, WGA becomes insufficient to asses the reliance of spurious features in settings with multiple classes. This is because WGA cannot differentiate between loss in test accuracy due to spurious correlations, or due to lack of learnability of invariant correlations stemming from limited capacity or insufficient training data. In such settings, we measure the degree of spurious feature reliance through *Spurious Misclassifications*, i.e. the percentage of samples of one class $(c1)$ containing the spurious feature of another class $(c2)$ that are misclassified as $(c2)$ during testing. *Lower Worst Group Accuracy indicates heavy reliance on spurious correlations while high Worst Group Accuracy indicates little to no reliance on spurious correlations. A high number of Spurious Misclassifications indicates heavy reliance on spurious correlations while a low number Spurious Misclassifications indicate little to no reliance on spurious correlations.*

5 SPURIOUS CORRELATIONS ARE FORMED FROM A FEW KEY SAMPLES

In this section, we aim to form a deeper understanding of relationship between spurious correlations and standard training paradigms. Deep neural network training relies on the Empirical Risk Minimization (Vapnik, 1998) principle, where during training, the model attempts to minimize risk or loss on the training data. From recent works, we understand that spurious correlations are learned because the network uses these weakly predictive but simple spurious features to further minimize training risk (Sagawa et al., 2020a; Geirhos et al., 2020; Kirichenko et al., 2022). In this section, we attempt to understand if this simple statement encapsulates the entire relationship between spurious correlations and deep neural network training.

We understand that spurious correlations are formed because standard training paradigms are prone to accepting any features that help minimize risk or loss during training. However, different samples have different contributions to overall training loss due to differences in difficulty of learning the features used to make correct classification. Does this imply that samples containing harder core/invariant features contribute more to the formation of spurious correlations? Or do all samples contribute equally to the formation of spurious correlations, as a network might simply prefer a simpler feature over a more complex one?

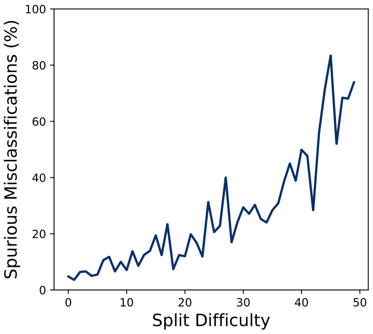


Figure 3: Spurious feature reliance exhibits polynomial growth with increasing sample difficulty.

To answer this question, we study a variant of the CIFAR-10S setting described in Sec. 4. In this experiment, we first train a ResNet20 on standard CIFAR-10. Early in training, we compute training error for each sample to estimate the difficulty of learning these samples, as is done in Paul et al. (2021). Such sample-wise training error is computed as $||p(w, x) - y||_2$, where $p(w, x)$ is the probability distribution given by the network for sample x, w denotes the network parameters early in training, and y is the one hot encoding of the ground truth value. We specify the epoch at which these values are computed in the Appendix. Trivially, the greater the error, the more difficult that sample is to learn. Since the standard CIFAR-10 does not contain significant spurious cues that can impact difficulty of learning, we estimate the difficulty of learning the core/invariant feature per sample as the difficulty of learning that sample. Next, we train a ResNet20 on two settings: One where we introduce the spurious feature in samples with the lowest training error (samples with *easy invariant* features) and another where we introduce spurious features in samples with the highest training error (samples with *hard invariant* features). We term these settings as `Easiest` and `Hardest`, respectively. In each setting, we vary the strength of the spurious signal by increasing the amount of area it takes up in each image. Spurious feature $S1$ takes up the least amount of area and is the most difficult to learn while spurious feature $S3$ takes up the most amount of area and is the easiest spurious feature to learn. In other words, a network will have less reliance on spurious feature $S1$ as it is difficult to learn and more reliance on $S3$ as it is easier to learn. $S2$ is in between them. Note that in both experimental settings, we introduce spurious features in only 100 samples (2% of class $c1$, 0.2% of the training set), and we maintain the proportion of samples containing the spurious features in all settings.

Samples with simple core features do not contribute to spurious correlations. In the setting where we introduce spurious features in samples with low training loss (`Easiest`), we find that Spurious Misclassifications tend to zero (Fig. 2(a)). This is true even if the strength of the spurious signal is increased. More interestingly, the network learns the spurious features present in these samples in the first couple of epochs of training but overcomes them as training converges. This is evidenced by high Spurious Misclassifications in the first half of training followed by low Spurious Misclassifications in the second half of training, as shown in Fig. 2(c). In other words, the network learns to ignore weakly predictive but simpler spurious features and prefers strongly predictive but complex core features in those samples that contain relatively easier invariant features.

Samples with hard core features are primary contributors to spurious correlations. In the setting where we introduce spurious features in samples with high training loss (`Hardest`), we find that Spurious Misclassifications are significantly higher, as shown in Fig. 2(b). This is also true for spurious features that are harder to learn ($S1$). Unlike what we observe in the previous setting, the network is unable to overcome the spurious features present in these samples, as we observe that Spurious Misclassification either remains the same or increases in the second half of training, as shown in Fig. 2(d). It is important to note that although only 100 of all samples of class $c1$ contain the spurious feature (2% of class $c1$, 0.2% of the training set), almost all samples of class $c2$ (100% of

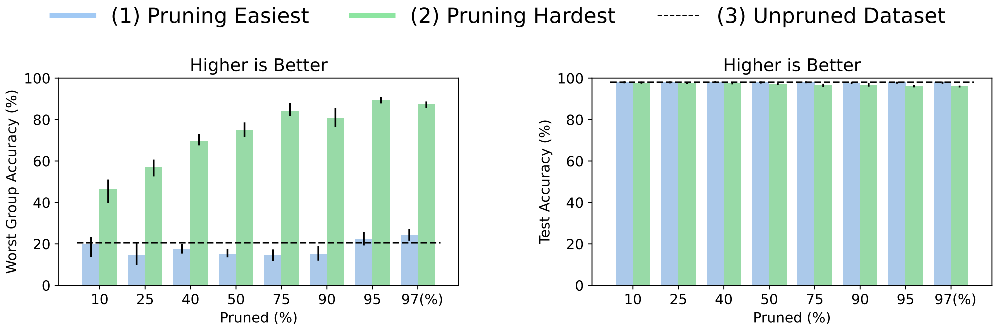


Figure 4: Excluding only a handful of training samples with spurious features and hard core features mitigates spurious correlations in the CelebA setting. This is indicated by high Worst Group Accuracies. Excluding up to 97% of all training samples with spurious features and easy core features shows no improvements in worst group accuracy.

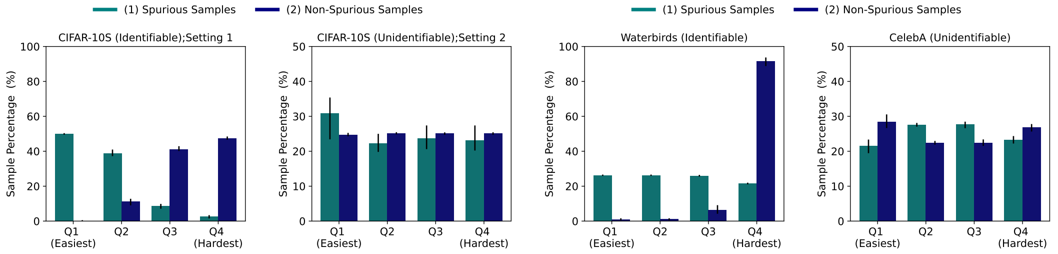
class $c2$) are spuriously misclassified during testing (Fig. 2(b)). This is in sharp contrast to the setting where we introduce ***the same*** spurious features in samples with easy invariant features (`Easiest`), where almost ***none*** of the samples of class $c2$ are spuriously misclassified (Fig. 2(a)).

Spurious feature reliance exhibits polynomial growth with increasing sample difficulty. We extend our original experimental setting to introduce the spurious feature in 100 samples ranging from the easiest to the hardest, instead of only considering the Easiest and Hardest 100 samples. Intuitively, as we scale up the difficulty of the 100-sample split to which we inject the spurious features, we observe an increase in spurious misclassifications, as shown in Fig. 3. Interestingly, we observe that spurious feature reliance grows polynomially as we scale up the difficulty of the core features of the 100 samples, implying that those samples with hard core features contribute significantly more to spurious feature reliance than those with easy core features.

Excluding a few key samples during training severs spurious correlations. From the experiments above, we learn that samples with harder core features contribute far more to spurious feature reliance than samples with easy core features. Consider a setting where spurious features are uniformly distributed across samples containing easy and hard core features. In such a setting, what if we simply exclude a handful of the hardest training instances that contain spurious features in them? Does this simple exclusion sever the spurious correlations that would otherwise be learned by the network? In Fig. 4, we show that by simply excluding a few samples containing the ***spurious*** feature with ***hard invariant*** features in the CelebA setting studied in Sec. 3 (5% of all samples with spurious features in that class, 1% of the total train set), we observe significant improvements in Worst Group Accuracy (WGA: Female samples with spurious features). On the other hand, excluding up to 97% of the easiest samples containing the spurious feature shows no improvements in WGA. Within the pool of samples containing spurious features, difficulty of learning the invariant feature is similar to the difficulty of learning the sample. This is because spurious features in our setting exhibit very low variance. For instance, in the MultiNLI setting, the spurious feature comprises the same set of negation words across the training data. Additionally, we note that on pruning these samples, we do not observe significant drops in overall testing accuracies, implying that these samples do not contribute significantly to generalizability either (Fig. 4).

6 SEVERING SPURIOUS CORRELATIONS WITH DATA PRUNING

Based on the observations presented above, we develop a novel data pruning method to sever spurious correlations by identifying and pruning subsets of the training data that contain the few key samples that contribute to spurious feature reliance. To do so, we first understand how the presence of spurious features impacts the difficulty of learning a sample in the training set and how it impacts the training distribution when observed through the lens of sample difficulty.



(a) Training distribution variance based on the strength of the spurious feature in CIFAR-10S Testbed. Grouped by Quartiles, sorted by difficulty.

(b) Training distribution variance based on strength of spurious feature in identifiable and unidentifiable settings. Grouped by Quartiles, sorted by difficulty.

Figure 5: Impact of strength of spurious signal on sample difficulty.

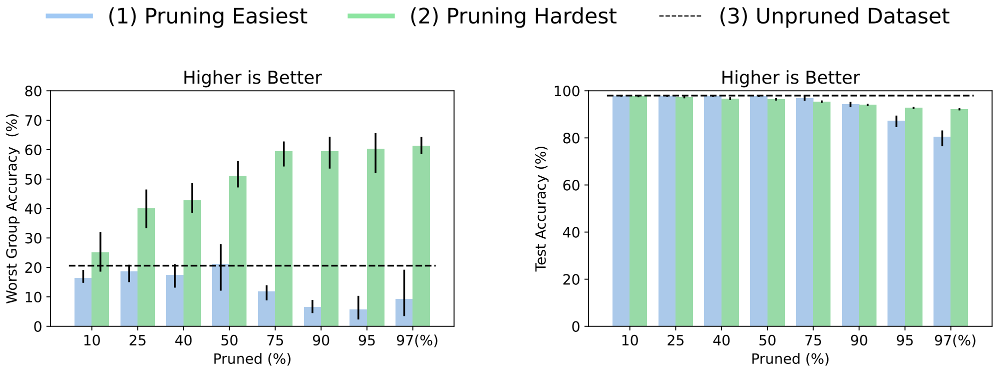


Figure 6: Excluding a small fraction of all hardest samples in the dataset mitigates spurious correlations in the CelebA setting. This is indicated by high Worst Group Accuracies.

6.1 THE IMPACT OF SPURIOUS FEATURES ON TRAINING DISTRIBUTION

Consider the CIFAR-10S setting described in Sec. 4. We create two versions of this setting: **Setting 1**: the strength of the spurious signal is significantly greater than the strength of the core, invariant signal; **Setting 2**: the strength of the spurious signal is only marginally greater than the strength of its invariant counterpart. In Setting 2, spurious information is unattainable/unidentifiable. In both settings, we introduce spurious features into samples at random such that both samples with easy and hard core features contain spurious features in them. We observe in Setting 1, all of the samples that contain the spurious feature lie in the first half of the data distribution when observed through the lens of sample difficulty. In Setting 2, however, samples containing spurious features are distributed uniformly (Fig. 5a). Note that in both settings, spurious misclassifications are significant (58.0% and 13.76% in Setting 1 and Setting 2, respectively).

We extend these findings to the benchmarks discussed in Sec. 4. In settings where the strength of the spurious signal is significantly greater than the strength of the invariant feature, (`Identifiable settings`), samples containing the spurious feature occupy a large majority of the first half of the training distribution. On the other hand, in the ***unidentifiable*** benchmarks, samples containing spurious features are ***uniformly distributed*** across the training distribution when viewed through the lens of sample difficulty (Fig. 5b). Note that since Hard ImageNet has multiple spurious features, we show that this setting is unidentifiable by showing images with and without these spurious features spread uniformly across the difficulty spectrum in the Appendix.

6.2 DATA PRUNING IN UNIDENTIFIABLE SETTINGS

From Sec. 6.1, we observe that in settings where the strength of the spurious signal is not significantly greater than the strength of the invariant signal, i.e. in settings where attaining information regarding the sample-wise presence of spurious features is impossible, samples containing spurious features are uniformly distributed across the training distribution when observed through sample difficulty. In other words, the presence of spurious features does not have a significant impact on the training distribution. We also know that samples containing hard core features that also contain the spurious feature are primary contributors to the formation of spurious correlations. Thus, to mitigate spurious correlations, one would only have to prune the ***hardest*** samples in the training set, as this subset of the data would contain samples with ***spurious*** features that have ***hard core*** features.

In Figs. 6 and 7, we show that by simply pruning a small subset of the hardest samples in the training set, one can overcome spurious correlations. Note that instead of pruning the hardest samples globally, we prune an equal proportion of the hardest samples per class to account for differences in difficulty per class.

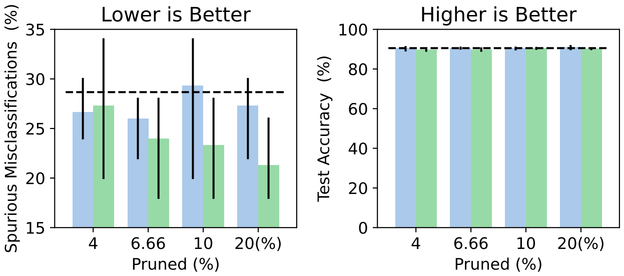


Figure 7: Excluding a small fraction of all hardest samples in the dataset mitigates spurious correlations in the Hard ImageNet setting. This is indicated by low Spurious Misclassifications.

6.3 DATA PRUNING IN IDENTIFIABLE SETTINGS

Unlike in unidentifiable settings, in settings where the strength of the spurious signal is significantly greater than the strength of the invariant signal, samples with spurious features are not uniformly distributed when sorted by sample difficulty. This is because the presence of strong spurious information enables the network to understand samples with hard core features better. In such settings, it is not feasible to simply prune the hardest samples of the training data as this will primarily prune samples that do not contain the spurious features (Figs. 5a & 5b). Liu et al. (2021); Ahmed et al. (2021); Creager et al. (2021); Zhang et al. (2022); Yang et al. (2024) show that in settings where the strength of the spurious signal is significantly greater than the strength of the invariant signal, it is possible to identify which samples contain spurious features in them and which ones do not. For such settings, we work with group labels as is done in Sagawa et al. (2020a); Kirichenko et al. (2022); Deng et al. (2023); Nam et al. (2022) and show that by simply pruning those spurious samples containing the hardest core features, we attain state-of-the-art performances as shown in Table 1. The term group label refers to labels that indicate the presence or absence of spurious features in each training sample within a class. Note that in settings with heavy class imbalances, such as Waterbirds, we prune such that the total number of samples in both classes is the same. We also compare our method with two popular techniques that do not require group labels during training, `JTT` (Liu et al., 2021) and `CnC` (Zhang et al., 2022) [1]. It is

Table 1: We present Worst Group and Mean accuracies. Consistent with literature, Mean accuracy is reported by weighting groups based on their prevalence in the unpruned training dataset. Simple data pruning attains state-of-the-art performance on standard benchmarks (Identifiable Settings). Group labels column represents availability of group labels in training and validation sets.

	Waterbirds (%)		MultiNLI (%)		Group Labels	
Method	Worst%	Mean%	Worst %	Mean%	Train	Val
ERM	74.81 (0.7)	98.10 (0.1)	65.9 (0.3)	82.8 (0.1)	✗	✗
CnC (Zhang et al., 2022)	88.5 (0.3)	90.9 (0.1)	-	-	✗	✓
JTT (Liu et al., 2021)	86.7	93.3	72.6	78.6	✗	✓
gDRO (Sagawa et al., 2020a)	86.0	93.2	**77.7**	81.4	✓	✓
DFR$^{\text{Tr}}$ (Kirichenko et al., 2022)	90.2 (0.8)	97.0 (0.3)	71.5 (0.6)	82.5 (0.2)	✓	✓
PDE (Deng et al., 2023)	90.3 (0.3)	92.4 (0.8)	-	-	✓	✓
Ours	**90.93 (0.58)**	92.48 (0.72)	75.88 (1.62)	81.07 (0.25)	✓	✓

---

[1] Note that we do not compare our results with Idrissi et al. (2022). In their experimental results for Waterbirds, the base ERM model has a higher worst-group accuracy than mean accuracy by 10%. This is very different from the norm where worst group accuracy is significantly lower than mean accuracy.

interesting to note that such data pruning attains state-of-the-art performances across a wide range of pruning sparsities, as observed in Fig. 8. This suggests, in addition to all previously observed results, that if one wants to ensure that the model they obtain after training is robust to spurious correlations, they only need to ensure that they are careful to remove a few training instances that are hard for the network to understand. This is in contrast to existing methods that severe spurious correlations which are more complex and at times, computationally expensive (Sagawa et al., 2020b; Ahmed et al., 2021; Ye et al., 2022; Zhang et al., 2022; Kirichenko et al., 2022; Moayeri et al., 2023; Deng et al., 2023).

7 CONCLUSION

Summary. We have shown that, in practice, attaining information regarding the presence and nature of spurious features is often impossible. This renders all existing techniques that show promise in overcoming spurious correlations as ineffective. We also discover that spurious correlations are formed due to only a small fraction of all samples containing the spurious feature and develop a novel data pruning technique that overcomes spurious correlations by pruning a small subset of the training data that contains these samples. Finally, we show that such data pruning attains state-of-the-art performance on standard benchmarks where information regarding spurious features is easily available.

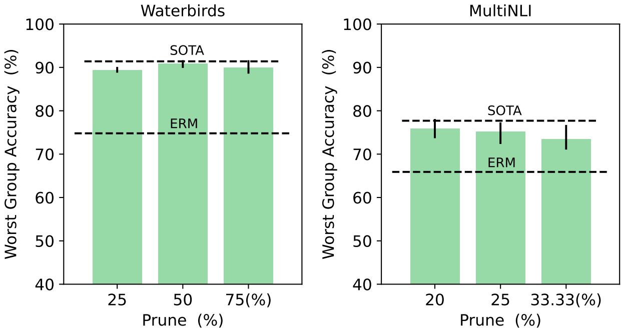


Figure 8: Our data pruning approach works well across a wide range of sparsities. Sparsities reported before class balancing, if applicable. We report SOTA results from Kirichenko et al. (2022) and Sagawa et al. (2020a) respectively. Note that Kirichenko et al. (2022) obtain state-of-the-art results on Waterbirds with the help of a held-out group balanced dataset (DFR$^{\text{Val}}$) and thus, we exclude it from our comparision in Table 1.

Outlook. With current practice of training models on increasing large datasets, it has become critical to develop techniques that identify samples that can cause these models to malfunction without requiring human intervention. In the future, we will further develop this work to take into account other failure modes that are commonly exhibited by these deep neural networks.

REFERENCES

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron C. Courville. Systematic generalisation with group invariant predictions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL `http://arxiv.org/abs/1907.02893`.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 2021.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/0506ad3d1bcc8398a920db9340f27fe4-Abstract-Conference.html`.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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(11):665–673, 2020.

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 (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 11–13 Apr 2022. URL `https://proceedings.mlr.press/v177/idrissi22a.html`.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=99RpBVpLiX`.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=rJlnB3C5Ym`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425. URL `https://doi.org/10.1109/ICCV.2015.425`.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In *NeurIPS*, 2022.

Mazda Moayeri, Wenxiao Wang, Sahil Singla, and Soheil Feizi. Spuriosity rankings: Sorting data to measure and mitigate biases. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=jSuhnO9QJv`.

Depen Morwani, jatin batra, Prateek Jain, and Praneeth Netrapalli. Simplicity bias in 1-hidden layer neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=PTvxck0QDE`.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Jun Hyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=_F9xpOrqyX9`.

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, Zachary 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 Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20596–20607, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020b.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nimit Sharad Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34330–34343. PMLR, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://aclanthology.org/N18-1101`.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961. PMLR, 2024. URL `https://proceedings.mlr.press/v238/yang24c.html`.

Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. *CoRR*, abs/2210.11075, 2022. doi: 10.48550/arXiv.2210.11075.

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron C. Courville. Can subnetwork structure be the key to out-of-distribution generalization? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12356–12367. PMLR, 2021. URL `http://proceedings.mlr.press/v139/zhang21a.html`.

Michael Zhang, Nimit Sharad Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 2022.

A APPENDIX

A.1 TRAINING DETAILS

CIFAR-10S. We use the ResNet20 implementation from Liu et al. (2019) that we train for 160 epochs. The network is optimized using SGD with an initial learning rate 1e-1 and weight decay 1e-4. The learning rate drops to 1e-2 and 1e-3 at epochs 80 and 120 respectively. We main a batch size of 64. Sample difficulty is computed after the 10th epoch.

CelebA. We use an ImageNet pre-trained ResNet-50 from PyTorch (Paszke et al., 2019) that we train for 25 epochs. The network is optimized using SGD with a static learning rate 1e-3 and weight decay 1e-4. We main a batch size of 64. Sample difficulty is computed after the 10th epoch.

Hard Image-Net. We use an ImageNet pre-trained ResNet-50 from PyTorch (Paszke et al., 2019) that we train for 50 epochs. The network is optimized using SGD with a static learning rate 1e-3 and weight decay 1e-4. We maintain a batch size of 128. Sample difficulty is computed after the 1st epoch.

Waterbirds. We use an ImageNet pre-trained ResNet-50 from PyTorch (Paszke et al., 2019) that we train for 100 epochs. The network is optimized using SGD with a static learning rate 1e-3 and weight decay 1e-3. We maintain a batch size of 128. Sample difficulty is computed after the 1st epoch.

MultiNLI. We use a pre-trained BERT model that we train for 20 epochs. The network is optimized using AdamW using a linearly decaying starting learning rate 2e-5. We maintain a batch size of 32. Sample difficulty is computed after the 5th epoch.

A.2 ADDITIONAL EXPERIMENTAL DETAILS

CIFAR-10S. We follow a similar approach to Nagarajan et al. (2021) for adding a spurious line where pixel values for a vertical row of pixels in the middle of the first channel are set to the maximum possible value (255) before normalization and before any augmentations. We use the same augmentations generally used from training on the original CIFAR-10 Krizhevsky (2009).

CelebA. In this setting, we maintain 5000 Female Samples without Eyeglasses and 2500 Male samples with Eyeglasses and 2500 Male samples without Eyeglasses.Consistent with implementations in Sagawa et al. (2020a); Liu et al. (2021), we do not use any augmentations.

Hard ImageNet. In this setting, we maintain 58 Dog Sled samples with minimal spurious features and 100 Ski samples randomly drawn from the dataset. All remaining classes are maintained the same.We use the same augmentations used for training on ImageNet.

Waterbirds. We use the original Waterbirds setting commonly used in practice (Sagawa et al., 2020a; Liu et al., 2021; Zhang et al., 2022; Kirichenko et al., 2022).We use the augmentations used in Kirichenko et al. (2022) when training, which are similar to the augmentations used for training on ImageNet.

MultiNLI. We use the original MultiNLI setting commonly used in practice (Sagawa et al., 2020a; Liu et al., 2021; Kirichenko et al., 2022). Consistent with implementations in Sagawa et al. (2020a); Liu et al. (2021); Kirichenko et al. (2022), we do not use any augmentations.

A.3 HARD IMAGENET: VERIFYING UNIDENTIFABILITY

We show that our Hard ImageNet setting is an unidentifiable setting by presenting information similar to Fig. 5a and Fig. 5b. But since there exist multiple spurious features, namely `Snow`, `People`, `Trees`, and `Hills`, we simply show that samples that contain these features and those that do not are scattered uniformly across the difficulty spectrum (Fig. 9).

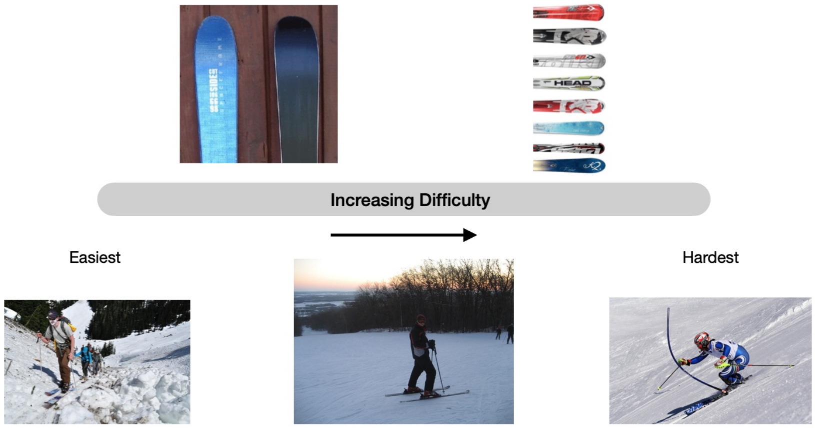


Figure 9: Training distribution variance based on strength of spurious feature in identifiable and unidentifiable settings.