Identifying Spurious Features through Forgetting Scores

Author 1 **Mirelle George** **Author 3** **Author 4**
UID: 205692340
msgeorge21@ucla.edu

Abstract

When spurious features are present in data sets, test accuracy is often diminished. Current leading methods for identifying examples that contain spurious features are limited in examining the behavior of examples only after the final epoch of training, as in the Just Train Twice (JTT) method. Here, we present a method for identifying examples with spurious features by first tracking their forgetting scores, a concept proposed by Toneva et al. (2019), throughout all epochs of training and then classifying the examples with low forgetting scores to be those with spurious features. We can then upsample the examples with high forgetting scores and then continue to train the subsequent data set to increase robustness. Our experiments demonstrate that this forgetting-based upsampling and further training slightly outperforms JTT's test accuracy.

1 Introduction

Machine learning models often struggle with spurious correlations - undesirable relationships between class-irrelevant features and labels that can lead to poor generalization. While these spurious features may help improve training accuracy, they typically harm test performance, especially for minority groups that don't exhibit these correlations. Current approaches for identifying and mitigating spurious correlations, such as Just Train Twice (JTT), rely on examining model predictions only at the final epoch of training, potentially missing valuable information from the training trajectory.

In this work, we propose a novel method for identifying examples with spurious features by leveraging forgetting scores throughout the entire training process. A forgetting event occurs when a data point is misclassified after being previously classified correctly, and the forgetting score counts these events during training. Our key insight is that examples with spurious features tend to have significantly lower forgetting scores compared to examples without them, as the model learns and maintains these easier-to-exploit correlations early in training.

We evaluate our approach on the SpuCo MNIST dataset, where some examples contain spurious colored backgrounds that correlate with digit labels. By using forgetting scores to identify and upsample minority groups, we demonstrate improved test accuracy compared to JTT baselines. This work suggests that forgetting dynamics can provide a more robust signal for detecting spurious correlations compared to examining predictions at a single point in training.

2 Literature Review

A "forgetting event" is described as an occurrence when a training example switches from being correctly classified to incorrectly classified over the course of learning. Examples that are never forgotten across training are considered unforgettable examples and examples that are consistently forgotten across training are considered forgotten examples. Toneva et al. (2019) hypothesizes that certain examples consistently forgotten over training, if they exist, must not be similar to other examples from the same task. To explore this further, researchers analyze the proportion of "forgettable" and "unforgettable" examples for a given task and assess how these examples affect the decision boundary of a model and generalization error. Toneva et al. (2019)'s experimental findings in-

dicate that: a) a large number of unforgettable examples are stable across different seeds and neural architectures; b) examples with noisy labels and images with "uncommon" features are among the most forgotten examples; c) even when a solid portion of the least-forgotten examples is removed from the training data, the model still performs competitively on the test set.

Machine learning model performance can be poor when the training and test distributions do not align. A common type of distribution shift is the subpopulation shift, which is the case when the proportion of some subpopulation varies differently in the training and test data. One case is when the data contains spurious correlations, which is when domain-dependent features in the training data are strongly correlated with labels, but these correlations do not generalize. For example, as provided by Joshi et al. (2024), in the training data, if many images of a 'swallow' appear on a 'sky' background, an image classifier may learn that the swallow category is associated to the sky background instead of the bird. Therefore, during test time, the model will have a poor worst-group accuracy on swallows that do not appear on a sky background. If majority and minority groups of examples with and without spurious features are known, methods to mitigate these spurious correlations can be implemented, such as upweighting or upsampling the group through group-known robust methods. Deep neural networks often take advantage of domain-specific (spurious) features that are common in the majority of examples within a class during training and this reliance on spurious features results in poor test accuracy for minority groups that do not exhibit these features. Despite the recent efforts to remedy spurious correlations, the lack of standardized benchmark and more complex datasets limits reproducible evaluation and accurate comparison of the proposed solutions. To address this, Joshi et al. (2024) systematically benchmarks 8 SOTA algorithms, by training over 5K models across 6 vision datasets, including our 3 newly introduced datasets that have more challenging image classification tasks with more complex spurious features. The performance of group inference methods deteriorates when the spurious feature is more complex.

3 Approach

We employ the SpuCo MNIST dataset to test our hypothesis that lower forgetting scores are associated with examples that have spurious features. The SpuCo MNIST dataset contains images of colored digits, where some examples have additional colored backgrounds that are spurious features. The task is to recognize the correct digit in the image. We initialize a data set that contains 25 groups, where 5 groups contain the spurious colored background feature and 20 do not. Then, we train these examples for ten epochs using a 2-layer multi-layer perceptron model and record the prediction for each example for each epoch. After training, we record the corresponding forgetting score, ranging from 0 through 6, for every example (6 corresponding to examples that are never predicted correctly). We subsequently upsample the examples identified as lacking spurious features (part of "minority groups") based on each example's forgetting score and then train a new model given this modified training set.

4 Experiments and Results

4.1 Forgetting Score Groupings

We find that a much lower forgetting score strongly distinguishes majority groups with spurious features from minority groups without, as shown in Figures 1, 2, and 3. The majority group images on the diagonal of each heatmap (e.g. digit 0 with color 0) all have low average forgetting scores less than 1. In contrast, the other minority groups have higher average forgetting scores greater than 1. This distinction is more evident with higher spurious correlation strength.

We also find that forgetting scores offer more information than JTT. Forgetting scores capture information across training epochs, while JTT is limited to prediction correctness in the final epoch. Figures 4, 5, and 6 compare forgetting scores with JTT groupings based

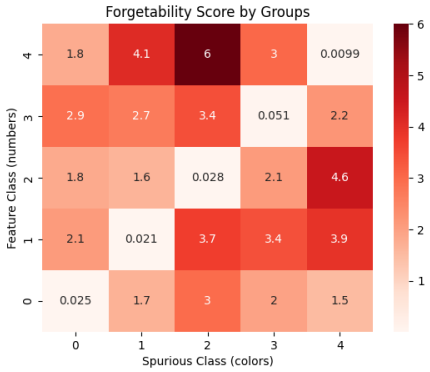

Figure 1: Forgettability for data with spurious correlation strength 0.9

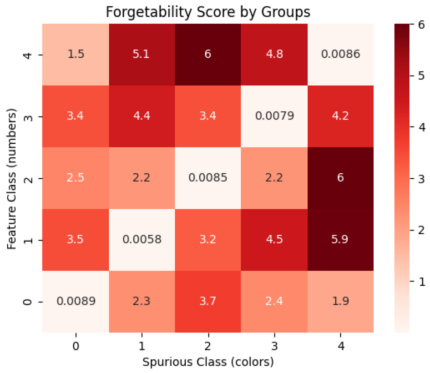

Figure 2: Forgettability for data with spurious correlation strength 0.95

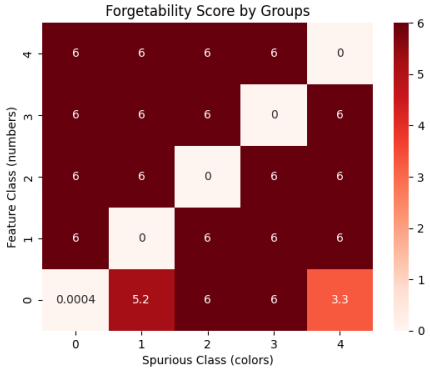

Figure 3: Forgettability for data with spurious correlation strength 0.995

on prediction correctness after initial epochs of training. In particular, Figures 4 and 5 highlight missclassification that can be mitigated with forgetting scores. The left sides of these two graphs highlight how JTT incorrectly groups some data in the majority groups as part of the minority group. These data points misclassified by JTT have a lower average forgetting score. Thus, this incorrect JTT grouping can be avoided via a forgetting-based grouping. Similarly, the right sides of the two graphs show JTT also misclassifies some minority group data as part of majority groups. If a forgetting-score threshold is used to group these points instead, this misclassification can be avoided.

4.2 Forgetting-based Upsampling

After confirming that forgetting-score distinguishes between majority and minority groups, we group data using two different methods. The first is directly into 7 groups by forgetting score from 0 through 6. We refer to this as "forgetting-based". The second, inspired by JTT's two groups, splits the data into two based on if the forgetting score is greater than 0. This method, which we call "forgetting-based binary", follows from the forgetting scores of the majority groups being close to 0. For each grouping method, we then upsample all forgetting-based groups in order to train robustly. Figures 7, 8, and 9 show the results of training on forgetting-based upsampled data in comparison to the results of JTT robust training. We find that forgetting-based robust training slightly outperforms JTT's test accuracy. Furthermore, we see a comparable to slightly improved worst-group (when using ground-truth groupings by digit and color) test accuracy with both forgetting-based methods compared to JTT, as shown in Table 1.

5 Conclusion and Future Works

Overall, we find forgetting scores to be promising in mitigating spurious correlations. We find that majority groups with spurious features have much lower forgetting scores than minority groups without. Thus, upsampling

Spurious Correlation Strength	Method	Worst-Group Accuracy
0.9	JTT	76.47
	Forgetting-based	76.07
	Forgetting-based binary	76.07
0.95	JTT	36.52
	Forgetting-based	**58.59**
	Forgetting-based binary	**58.59**
0.995	JTT	2.02
	Forgetting-based	**4.01**
	Forgetting-based binary	**4.01**

Table 1: Worst-group accuracy of after different robust training methods.

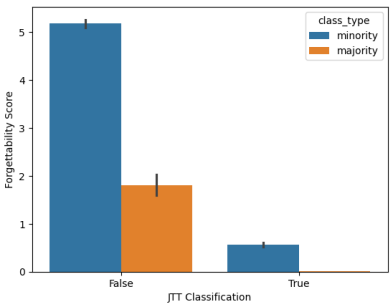

Figure 4: Forgettability versus JTT Results with spurious correlation strength of 0.9

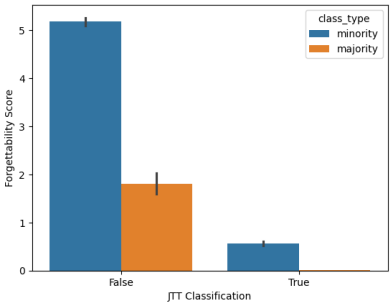

Figure 5: Forgettability versus JTT Results with spurious correlation strength of 0.95

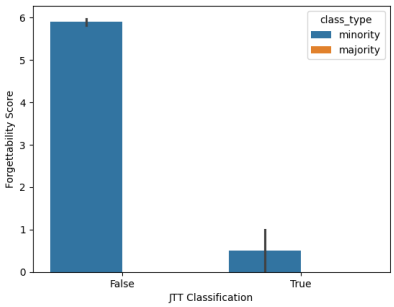

Figure 6: Forgettability versus JTT Results with spurious correlation strength of 0.995

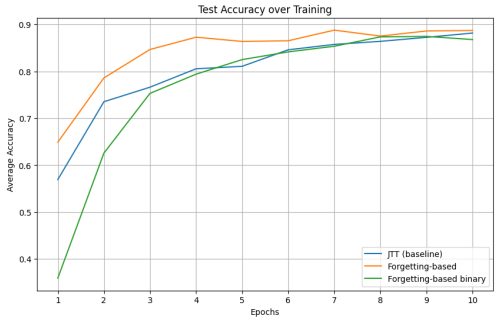

Figure 7: Test Accuracy over Training with Spurious Correlation Strength of 0.9

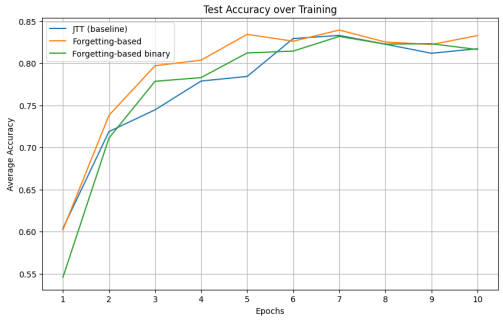

Figure 8: Test Accuracy over Training with Spurious Correlation Strength of 0.95

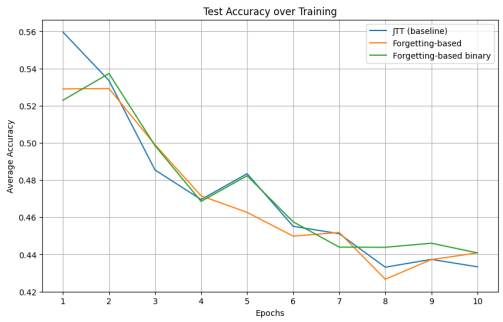

Figure 9: Test Accuracy over Training with Spurious Correlation Strength of 0.995

data based on forgetting score mitigates the issue of learning spurious features; we find that this method slightly improves test performance and worst-group test accuracy compared to JTT. Furthermore, forgetting scores offer more information to mitigate spurious correlations than JTT as forgetting scores highlight differences between majority and minority groups that JTT fails to capture. Future research may explore robust training based on other forgetting-based grouping methods, such as using different forgetting score thresholds to divide into groups.

References

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. 2024. Towards mitigating more challenging spurious correlations: A benchmark new datasets.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning.