Classification Bias on a Data Diet

Tejas Pote, Mohammed Adnan, Yigit Yargic, Yani Ioannou,

UNIVERSITY OF CALGARY

Introduction

- Modern deep learning models achieve overwhelming generalization performance in image classification tasks on vision datasets like CIFAR-10 and CIFAR-100.
- However, we observe a bias in classification accuracies of different classes even when the datasets are balanced across classes.
- This classification bias has been seen to aggravate in settings like model pruning [3] and adversarial training [1].
- We analyze the effect of data sparsification techniques, particularly data diet [2], on classification bias in class-balanced datasets like CIFAR-10 and CIFAR-100.

Data Sparsification

- While pruning of model parameters has been extensively studied, pruning the dataset by identifying important training samples has received attention only recently.
- The primary motivation for pruning data has been to study the effect of individual samples and population sub-groups on training dynamics.
- By reducing samples in the training data, these techniques additionally offer huge savings in computation.
- With most of these techniques leading to imbalanced data distributions, it is worth thinking about how they affect the classification bias.

Data Diet

- Data diet [2] is a state-of-the-art data pruning technique which identifies important samples early in training using simple scores.
- The importance of training samples is decided with the help of two scores: '**GraNd (Gradient Normed)**' and '**EL2N (Error L2-Norm)**'
- **GraND Scores**: The GraNd score of a training example at time t in training is defined as $\mathbb{E}_{w_t}||g_t(x,y)||_2$, where $g_t(x,y)$ is the gradient of the loss evaluated over the sample (x,y) at training step t.
- **EL2N Scores**: The EL2N score for a particular training example at time t in training is given by $\mathbb{E}||p(w_t,x)-y||_2$, where $p(w_t,x)$ is a probability distribution over the output classes of the confidence of prediction.
- These scores are evaluated for every training sample and they are ranked according to decreasing values of the scores.
- Depending on the desired level of sparsity, an appropriate fraction of samples is chosen.

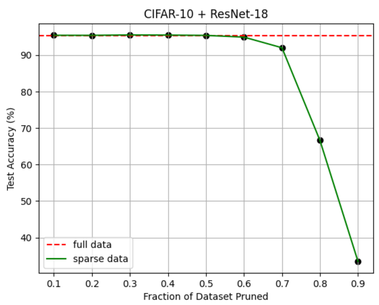

Figure 1. CIFAR-10/ResNet-18

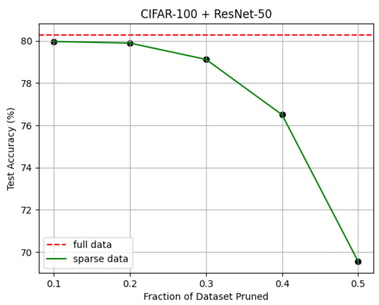

Figure 2. CIFAR-100/ResNet-50

- Figures 1 and 2 show the variation of overall model accuracies with data sparsity. In both cases, for data sparsities within a certain range, training over sparse data is able to match the performance of training over full dataset.
- Beyond a certain level of data sparsity, the generalization performance drops.
- The analysis of overall accuracies suggests that a considerable number of points can be dropped from the traning data.
- However, this analysis does not provide any idea about the classification bias due to data diet.

Class-wise Accuracies

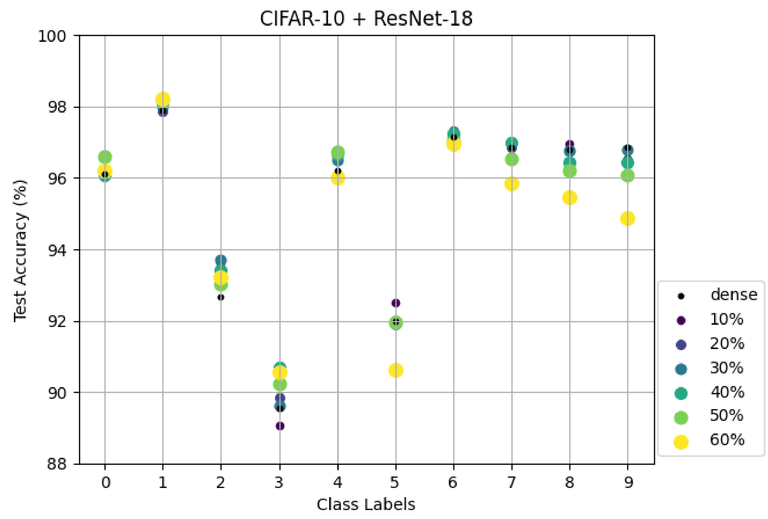

Figure 3. CIFAR 10/ResNet 18 Class-wise Accuracies: A plot of accuracies of individual classes with increasing data sparsity for ResNet-18 trained on CIFAR-10. An increase in data sparsity leads to increase in the accuracy of most of the classes.

- We observe the trend in the accuracies of the individual classes to analyse the classification bias.
- Figure 3 shows the variation of class-wise accuracies with increasing data sparsity on CIFAR-10 trained with ResNet-18.
- Classes 3 (cat) and 5 (dog), whose performance is worst over dense data show improved accuracies at moderate levels of data sparsity.
- Above 50%, we see a drop in the accuracy of most of the classes. This is in agreement with the trend observed for generalization performance, where the model performance starts degrading above 50% data sparsity.

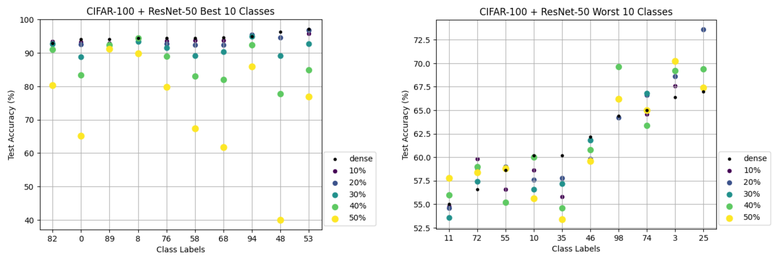

Figure 4. CIFAR 100/ResNet 50 Class-wise Accuracies: A plot of the test accuracies of 10 best (left) worst (right) performing classes with increasing data sparsity for ResNet-50 trained on CIFAR-100.

- For the ResNet-50 architecture trained over CIFAR-100, we analyse the class-wise accuracies of the 10 best and worst performing classes over the dense model.
- Figure 4 shows the variation of accuracies of these classes with changing levels of sparsity.
- While the best performing classes do not suffer a lot in test accuracy, many worst performing classes show better accuracies with increasing data sparsity.
- Although these observation hint at reduced classification bias with increasing data sparsity, we need more compelling evidence to justify this observation.

Standard Deviation of Class-wise Test Accuracy

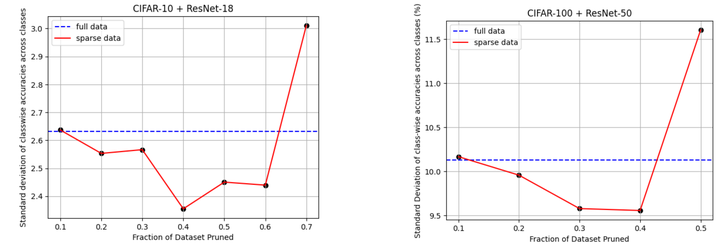

Figure 5. Standard Deviation of Class-wise Test Accuracies Across All Classes: for different data sparsities when ResNet-18 is trained on CIFAR-10 (left) and ResNet-50 is trained on CIFAR-100 (right).

- In order to study the bias on quantifiable terms, we calculate the standard deviation of class-wise accuracies across all classes.
- A lower value of standard deviation suggests that the deviations of per-class accuracies from the mean per-class accuracy is small, which indicates a lower classification bias.
- Figure 5 shows the plot of standard deviation of per-class test accuracy across classes with increasing data sparsity.
- For both the models, there is a range of data sparsities for which the standard deviation of per-class accuracies is lower than that with training over dense data. It is interesting to note that for these exact range of sparsity values, the overall test accuracies of the models trained on sparse data are comparable with that trained on dense data.

Class-wise Data Keep Ratio

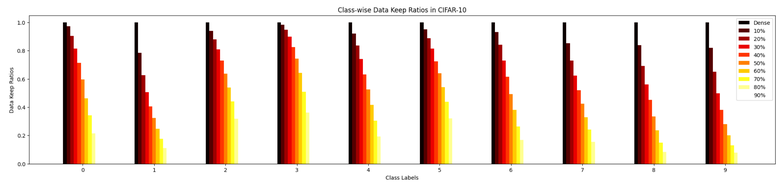

Figure 6. CIFAR-10/ResNet-18 Data Keep Ratio: A plot of data keep ratio across all the classes with increasing data sparsity.

- The class-wise data keep ratio is simply the ratio of training samples retained in a class in a sparse model over the number of samples in the class in the dense model.
- Figure 6 shows the variation of class-wise data keep ratio with data sparsity. Interestingly, classes 3 (cat) and 5 (dog) have the highest keep ratio across all the models.
- This suggests that both these classes contain more difficult samples to learn than other classes, consistent with the intuition that these two classes likely share the most visual features within the CIFAR-10 set of classes [1].

Insights

- Our analysis shows that data sparsification techniques like data diet can help reduce classification bias.
- More interestingly, our results suggest that training on imbalanced data distributions — when created by informed data pruning algorithms like data diet — can result in image classification models with more consistent class-wise generalization performance.

References

[1] Benz, P., Zhang, C., Karjauv, A., and Kweon, I. S. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pp. 325–342. PMLR, 2021.

[2] Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

[3] Tran, C., Fioretto, F., Kim, J.-E., and Naidu, R. Pruning has a disparate impact on model accuracy. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

CPAL LESS IS MORE

CalgaryML Lab: