Classification Bias on a Data Diet

Tejas Pote*
Indian Institute of Technology, Kharagpur
tejaspote915@iitkgp.ac.in
Adnan Ahmad†
University of Calgary
adnan.ahmad@ucalgary.ca
Yigit Yargic†
University of Calgary
yigit.yargic@ucalgary.ca
Yani Ioannou
University of Calgary
yani.ioannou@ucalgary.ca

Data diet is a data pruning method that retains important samples in data which are identified early in training using simple scores. For training on class-balanced datasets, although data diet drastically reduces the dataset size, it results in class-imbalance in data. With a pre-conceived notion of class-imbalance being one of the potential reasons for classification bias in models trained on image classification tasks, we try to understand if data diet exacerbates class bias as compared to training with the full training dataset. In contrast to our belief, our analysis of data diet hints towards an interesting direction in understanding how to reduce the inter-class generalization performance discrepancy observed when training even on class-balanced datasets like CIFAR-10 and CIFAR-100. In essence, our study underscores that we can achieve models with better class-wise generalization performance by training on imbalanced data distributions created by informed data pruning algorithms like data diet.

1. Introduction

Deep learning models like residual neural networks [1] have been able to obtain overwhelming performance on image classification tasks on class-balanced datasets like CIFAR-10 and CIFAR-100 [2]. However, although the overall generalization performance is impressive, the class-wise performance has been observed to be somewhat skewed even on such class-balanced datasets — showing consistently better generalization performance for some classes and worse for others. This class-wise discrepancy has been observed to be similar across different model architectures for a given dataset [3].

Recently, substantial work has been done recently in literature to identify important samples in training data [4–7]. The primary motivation for these works is to understand the effect of individual samples and population subgroups in data on training dynamics. However, these ideas have been extended to *data pruning*, where deep learning models are trained on the subset of training data identified as important for learning. Such data pruning methods promise large savings in the computation necessary for training by drastically reducing the samples in training data set — often in a way that affects class-balance. This poses an interesting question — do data pruning methods exacerbate classification bias?

Prior works on data pruning have mostly focused on the overall generalization performance of the model, but have not investigated the effect of pruning on classification bias. It is imperative to study the effect of data pruning on model bias before deployment of such models. In this work, we analyze

[*] Work done as a part of MITACs Globalink Research Internship.
[†] Equal contribution

Preprint. Work in progress.

the effect of data sparsification on the class-wise distribution of training data and subsequently the classification bias on a per-class basis. We consider the benchmark datasets CIFAR-10 and CIFAR-100 in which all the classes have an equal number of training and test samples. We select import training samples and create a subset of training data using "Data Diet" [8], a state-of-the-art data sparsification technique.

We highlight the following insights from our analysis:

- Models trained with class-balanced data distributions are susceptible to biased classification, we speculate this is due to varying difficulty of learning features to represent different classes. This is in contrast to the common misconception that models trained on balanced data distributions are not biased.
- Data sparsification techniques like Data Diet are effective at identifying the difficult classes to learn using global pruning criterion.
- Depending on the dataset, there is a range of data sparsities for which we can achieve classification with lower bias than training with the entire training dataset.
- Our empirical observations suggest that contrary to the conventional notion, training models on imbalanced datasets can lead to lower classification bias if that imbalanced dataset is carefully selected considering the difficulty of training samples/classes.

We first present a brief overview of 'Data Diet' in section 3 along with the terminologies used. Next we present an in-depth analysis of bias mitigation through metrics like generalization performance (section 4.2), class-wise accuracies (section 4.3) and class-wise data keep ratios (4.4).

2. Related Work

2.1. Classification Bias in Deep Learning

Among the many challenges faced while deploying deep learning models for image classification, a crucial one is classification bias, where certain groups in data are poorly classified compared to others. Imbalanced training data distributions are one of the most common causes of models making biased predictions with respect to specific classes [9]. A number of techniques have been proposed to tackle classification bias by creating balanced data distributions. This includes random oversampling and undersampling [10], sampling with data cleaning [11–13], clustering based oversampling [14], etc. Classification bias has also been analysed in a number of different settings like model pruning [15] and adversarial training [3]. Hooker et al. [16, 17] observed that subsets of samples in the training data are disproportionately affected by model compression techniques like pruning and quantization. Blakeney et al. [18] incorporate knowledge distillation for tackling bias induced due to network pruning.

2.2. Data Sparsification

Data sparsification techniques have received a lot of attention recently, primarily to study the effect of training data on learning dynamics. A series of works is on *coresets*, in which the goal is to identify weighted subsets of training data with a predetermined tolerance in performance degradation [6, 19–21]. Toneva et al. [7] identify examples which transition from being correctly classified to misclassified during the course of training and conjecture that examples which are rarely forgotten can be omitted. Much work has been done on identifying difficult examples. For instance, [4] suggest 'Variance of Gradient' as an effective metric for identifying difficult examples. [5] introduce prediction depth, which represents the number of hidden layers beyond which a network's prediction becomes deterministic, as a measure of sample difficulty. 'Selection Via Proxy'(SVP) [6] train a small proxy model for selecting important examples in the training data.

3. Data Diet

Data Diet is a data pruning technique which identifies important samples in the training data earlier during training [8]. Consider the dataset $S = \{(x_i, y_i)\}_{i=1}^N$ sampled from a distribution D. A neural network is trained over these samples which has logit outputs denoted by $f_w(x) \in \mathbb{R}^K$ where w are the parameters of the network and K is the number of output classes. Data diet introduces two scores: '**GraNd (Gradient Normed)** ' and '**EL2N (Error L2-Norm)**' to compute the importance of each individual sample.

Gradient Normed Score (GraNd) The GraNd score of a training example at time t in training is defined as $\chi_t(x, y) = \mathbb{E}_{w_t}||g_t(x, y)||_2$, where $g_t(x, y)$ is the gradient of the loss evaluated over the sample (x, y) at training step t. At any given training step, the contribution of a training example (x, y) to the decrease in loss on any other example is bounded by the GraNd score. This suggests that examples with smaller GraNd scores have a smaller influence on learning than those with higher scores. We refer to [8] for a theoretical proof of this statement.

Error L2-Norm Score (EL2N) The EL2N score for a particular training example at time t in training is given by $\mathbb{E}||p(w_t, x) - y||_2$. Here, $p(w, x)$ is a probability distribution over the output classes constructed by applying the softmax function 'σ' over the logit outputs given by $p(w, x) = \sigma(f(w, x))$, where

$$\sigma(z_1, ..., z_K)_k = \frac{\exp\{z_k\}}{\sum_{i=1}^K \exp\{z_i\}}$$

This score is motivated from the definition of the GraNd score. For any training example (x, y), let $\psi_t^k(x) = \nabla_{w_t} f_t^{(k)}(x)$ be the kth component of the logit gradient. The GraNd score can be formulated as

$$\chi_t(x, y) = \mathbb{E}||\sum_{k=1}^K \nabla_{f^{(k)}} \ell(f_t(x), y)^T \psi_t^{(k)}(x)||_2,$$

under the cross-entropy loss ℓ, $\nabla_{f^{(k)}} \ell(f_t(x), y)^T = p(w_t, x)^{(k)} - y_k$, which is simply the error in prediction. Paul et al. [8] empirically observe that the logit gradients saturate after a few epochs into training. Thus, the GraNd score for a training example can be well approximated by the corresponding EL2N score.

4. Experiments

4.1. Experimental Setup

We present our analysis training ResNet-18 and ResNet-50 models on the CIFAR-10 and CIFAR-100 datasets respectively. Note that Paul et al. [8] used ResNet-50 for their CIFAR-100 experiments, motivating our usage despite the fact ResNet-50 is not normally trained on CIFAR-10/100. During training a batch size of 128 is used, with an initial learning rate of 0.1, which is decayed by a factor of 5 at epochs 60, 120 and 160. Both the ResNet-18 and ResNet-50 models are trained for a total of 200 epochs. We independently train 10 models each for ResNet-10/100 with different initializations on the entire dataset and compute the EL2N score at the 20th epoch as suggested in [8]. These 10 scores are then averaged over different random initializations. The EL2N scores obtained are then used to rank the examples in the dataset. The training data is then pruned to a desired sparsity level based on these scores and ResNet-18/ResNet-50 models are trained on the pruned dataset from scratch.

4.2. Generalization Performance

Figure 1 shows the variation of test accuracy of the models as we increase the level of pruning in the data. For the ResNet-18 model trained on CIFAR-10, the test performance is marginally better than the dense dataset performance for data sparsities ranging from 10 to 50%. Beyond 50% data sparsity, the test accuracy starts falling below the dense dataset accuracy, though the drop is not significantly

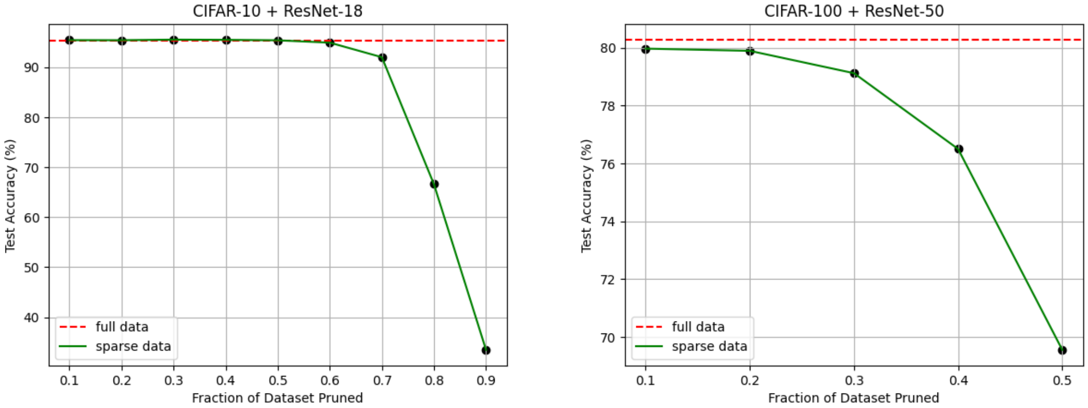

Figure 1: **Test Accuracy across Different Data Sparsities**: ResNet-18 trained on CIFAR-10 (left) and ResNet-50 trained on CIFAR-100 (right). In both cases, there is range of sparsities for which training over sparse data yields comparable or better generalization performance than that over dense training dataset.

large even up to 70% data sparsity. Thus for smaller models like ResNet-18, one can achieve similar generalization with data sparsity as high as 50%.

On the other hand, the test accuracy is just at par with the dense accuracy at extremely low levels of sparsity ($\sim$10%) for ResNet-50 model trained on CIFAR-100. As evident from Figure 1, the drop in accuracy is nominal for data sparsity below 40%. Beyond that, the degradation in model performance is considerably larger. These observations suggest that for both the models, there is range of data sparsity values where the training performance over sparse data is comparable to that over dense data.

4.3. Class-wise accuracies

For both models, we compute the test accuracies of each of the output classes over 5 different random initializations and average them over the intializations. Figure 2 shows the plot of the per-class accuracies of CIFAR-10 trained with ResNet-18 with increasing level of sparsity in data. An immediate observation from the plot is that classes 3 (cat) and 5 (dog) report the worst performance among all classes for models trained on dense as well as sparse data. The accuracies of these classes are seen to increase as the data pruning ratio is increased from 10 to 50%. 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 Figure 1, where the model performance starts degrading above 50% data sparsity.

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 3 shows the variation of accuracies of these classes with changing levels of sparsity. The test accuracies of the best performing classes are not affected much within 40% data sparsity. On the other hand, we observe that almost half of the worst performing classes report better accuracies than dense at 10% data sparsity. Although increasing sparsity leads to degradation in generalization performance, we observe improvement in the accuracies of the worst performing classes over the dense model. In fact, while there is a huge drop in the overall accuracy of the model at 50% sparsity, we observe that majority of the worst performing classes in Figure 3 report better accuracies than the model trained on dense data.

4.3.1. Standard Deviation of Per-Class Test Accuracy Across all Classes

Simply observing these trends in the per-class test accuracies does not give us a clear picture about classification bias. In order to study the bias on quantifiable terms, we calculate the standard deviation of per-class accuracy across all classes. Ideally, if there is no bias in classification, the prediction accuracy for all classes should be same. A high bias indicates that certain classes are

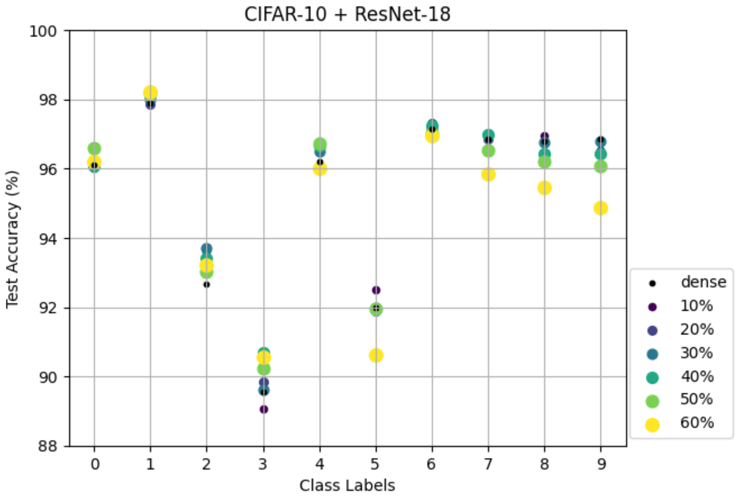

Figure 2: **CIFAR 10/ResNet 18 Class-wise Accuracies**: A plot of accuracies of individual classes with increasing data sparsity for ResNet-18 trained on CIFAR-10. An increase in data sparsity leads to increase in the accuracy of most of the classes.

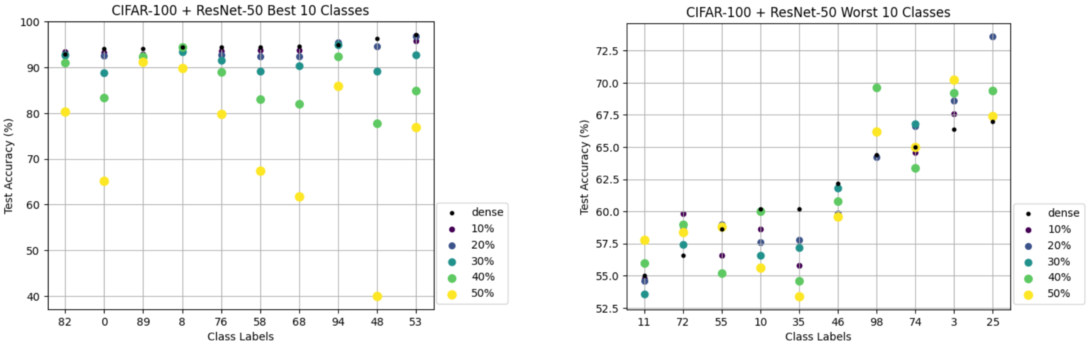

Figure 3: **CIFAR 100/ResNet 50 Class-wise Accuracies**: A plot of the test accuracies of 10 best (left) worst (right) performing classes with increasing data sparsity for ResNet-50 trained on CIFAR-100. The classes which show best performance over dense data do not suffer a lot in test accuracy in for moderate data sparsity. However, many classes which show worst performance over dense data show improvement with increasing data sparsity.

poorly classified compared to others. We previously calculated the per-class accuracies for both the models by averaging over 5 different initializations. We compute the standard deviation of these per-class accuracies across the mean per-class accuracy. 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. Using this metric allows us to compare the variance in the -per-class bias between two models irrespective of their overall test accuracies.

Figure 4 shows the standard deviation of class-wise accuracies across different data sparsities for both the models. For both the models, we observe that there is a range of sparsity values for which the standard deviation across all the class-wise accuracies is lower than that obtained using dense data, suggesting overall the biases for specific classes is reduced. As shown in Figure 1, 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. Thus, we observed empirically that data diet may reduce classification bias within a certain range of data sparsity as compared to the dense training dataset.

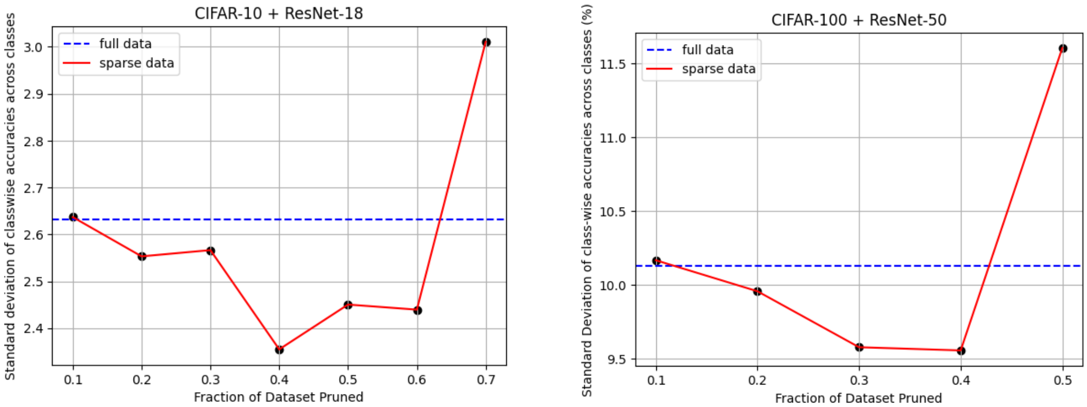

Figure 4: **Standard Deviation of Class-wise Test Accuracies Across All Classes**: for different data sparsities when ResNet-18 is trained on CIFAR-10 (left) and ResNet-50 is trained on CIFAR-100 (right). For both the models, we observe that in the range of data sparsities where training on sparse data yielded better performance the standard deviation of class-wise test accuracies across all classes is lower than that on dense data, suggesting lower classification bias for that data sparsity range.

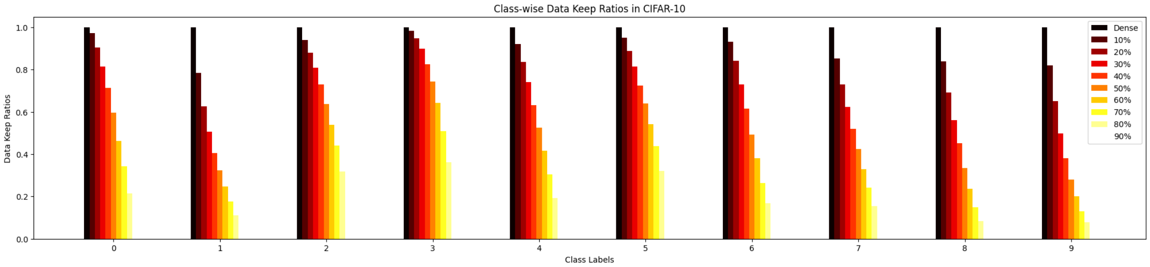

Figure 5: **CIFAR-10/ResNet-18 Data Keep Ratio**: A plot of data keep ratio across all the classes with increasing data sparsity. CIFAR-10 classes 3 (cat) and 5 (dog) retain the most number of samples over all sparsity values. This suggests they are difficult to classify, which is reflected by the trends in class-wise accuracy.

4.4. Class-wise Data Keep Ratios

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. The ranking of the training samples based on EL2N scores is done globally while pruning. Intuitively a higher EL2N score suggests that the sample is difficult to learn. Figure 5 shows that 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 [3]. This observation is well-supported by Figure 2 where the accuracies of classes 3 and 5 are observed to be lower than other classes. Although the data pruning is done globally, the EL2N scores somehow create an appropriate skewness in the data distribution, which roughly highlights the easy and difficult classes to learn.

5. Discussion

Data diet is a data pruning method that reduces the training dataset size by removing the easiest training data samples to classify, irrespective of class label. When training on a class-balanced dataset, data diet will often result in a drastically reduced, but class-imbalanced training dataset. With imbalanced training data distributions often being the leading offenders for learned bias in models trained on image classification tasks, our study set out to answer the question of if data diet exacerbates class bias as compared to training with the full training dataset.

Counter to our expectations, our results suggest data diet may provide an interesting direction in understanding how to reduce the inter-class generalization performance discrepancy observed when training even on class-balanced datasets like CIFAR-10 and CIFAR-100. In particular, in the case of CIFAR-10, the classes *cat* and *dog* (class labels 3 and 5 respectively) are often observed to have accuracies lower than the other classes across different models. The semantic similarity between the two classes has been speculated to be a probable reason for the misclassifications [3].

Data diet [8] ranks training samples based on EL2N scores (see section 3), i.e. the norm of the error vectors between the predicted and true labels. The samples retained (those with higher EL2N scores) are the training samples most difficult to learn. Figure 4 suggests that pruning data with data diet may suppress the class-wise biases learned, indicated by the lower standard deviation of per-class accuracies over all classes, over a range of moderate sparsities.

In summary, our results suggest that in fact 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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009.

[3] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 325–342. PMLR, 2021.

[4] Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients, 2021. URL https://openreview.net/forum?id=fpJX0O5bWKJ.

[5] Robert John Nicholas Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WWRBHhH158K.

[6] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

[7] 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 *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

[8] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

[9] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, oct 2002. ISSN 1088-467X.

[10] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[11] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, jun 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007735. URL https://doi.org/10.1145/1007730.1007735.

[12] Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In Silvana Quaglini, Pedro Barahona, and Steen Andreassen, editors, *Artificial Intelligence in Medicine*, pages 63–66, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-48229-1.

[13] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer, 1997.

[14] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explor. Newsl.*, 6(1):40–49, jun 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007737. URL https://doi.org/10.1145/1007730.1007737.

[15] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=11nMVZK0WYM.

[16] Sara Hooker, Aaron C. Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget. *arXiv: Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:226812844.

[17] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily L. Denton. Characterising bias in compressed models. *ArXiv*, abs/2010.03058, 2020. URL https://api.semanticscholar.org/CorpusID:222178157.

[18] Cody Blakeney, Nathaniel Huish, Yan Yan, and Ziliang Zong. Simon says: Evaluating and mitigating bias in pruned neural networks with knowledge distillation. *ArXiv*, abs/2106.07849, 2021. URL https://api.semanticscholar.org/CorpusID:235436182.

[19] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. SCG '05, page 126–134, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139918. doi: 10.1145/1064092.1064114. URL https://doi.org/10.1145/1064092.1064114.

[20] Myunggwon Hwang, Yuna Jeong, and Wonkyung Sung. Data distribution search to select coreset for machine learning. In *The 9th International Conference on Smart Media and Applications*, SMA 2020, page 172–176, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389259. doi: 10.1145/3426020.3426066. URL https://doi.org/10.1145/3426020.3426066.

[21] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. *Advances in neural information processing systems*, 29, 2016.