Bias in Pruned Vision Models: In-Depth Analysis and Countermeasures

Eugenia Iofinova
IST Austria

Alexandra Peste
IST Austria

Dan Alistarh
IST Austria & Neural Magic

Abstract

Pruning—that is, setting a significant subset of the parameters of a neural network to zero—is one of the most popular methods of model compression. Yet, several recent works have raised the issue that pruning may induce or exacerbate bias *in the output of the compressed model. Despite existing evidence for this phenomenon, the relationship between neural network pruning and induced bias is not well-understood. In this work, we systematically investigate and characterize this phenomenon in Convolutional Neural Networks for computer vision. First, we show that it is in fact possible to obtain highly-sparse models, e.g. with less than* 10% *remaining weights, which do not decrease in accuracy nor substantially increase in bias when compared to dense models. At the same time, we also find that, at higher sparsities, pruned models exhibit higher uncertainty in their outputs, as well as increased correlations, which we directly link to increased bias. We propose easy-to-use criteria which, based only on the* uncompressed model*, establish whether bias will increase with pruning, and identify the samples most susceptible to biased predictions post-compression. Our code can be found at* https://github.com/IST-DASLab/pruned-vision-model-bias.

1. Introduction

The concept of "bias" in machine learning models spans a range of considerations in terms of statistical, performance, and social metrics. Different definitions can lead to different relationships between bias and accuracy. For instance, if bias is defined in terms of accuracy disparity between identity groups, then accuracy in the "stronger" group may have to be reduced in order to reduce model bias. Several sources of bias have been identified in this context. For example, bias in datasets commonly used to train machine learning models [4,5,53] can severely impact outputs, and may be difficult or even impossible to correct during training. The choice of model architecture, training methods, evaluation, and deployment can create or exacerbate bias [2,42,43].

One potential source of bias which is relatively less investigated is the fact that machine learning models, and in particular deep neural networks, are often *compressed* for efficiency before being deployed. Seminal work by Hooker et al. [29] and its follow-ups, e.g. [28, 38] provided examples where model compression, and in particular pruning, can exacerbate bias by leading models to perform poorly on "unusual" data, which can frequently coincide with marginalized groups. Given the recent popularity of compression methods in deployment settings [13,18,19,27] and the fact that, for massive models, compression is often necessary to enable model deployment, these findings raise the question of whether the bias due to compression can be exactly characterized, and in particular whether bias is an inherent side-effect of the model compression process.

In this paper, we perform an in-depth analysis of bias in compressed vision models, providing new insights on this phenomenon, as well as a set of practical, effective criteria for identifying samples susceptible to biased predictions, which can be used to significantly attenuate bias.

Our work starts from a common setting to study bias and bias mitigation [28, 29, 40, 50]: we study properties of sparse residual convolutional neural networks [25], in particular ResNet18, applied for classification on the CelebA dataset [41]. Then, we validate our findings across other CNN architectures and other datasets. To study the impact of sparsity, we train highly accurate models with sparsity ranging from 80% to 99.5%, using the standard gradual magnitude pruning (GMP) approach [18, 21, 22, 55]. We consider bias in dense and sparse models from two perspectives: *systematic bias*, which refers to consistent errors in the model output, and *category bias*, which refers to violations of fairness metrics associated with protected groups.

On the positive side, our analysis shows that the GMP approach can produce models that are highly sparse, i.e. 90-95% of pruned weights, without significant increase in any bias-related metrics. Yet, this requires care: we show that *shared, jointly-trained* representations are significantly less susceptible to bias, and so careful choices of training procedure are needed for good results. On the other hand, at very high sparsities (95%-99.5%) we do observe non-trivial increase in category bias for the sparse models, for specific protected attributes. We perform an in-depth study of this phenomenon, correlating increase in bias with increased uncertainty in the model outputs, induced by sparsity. Leveraging insights from our analysis, we provide a simple set of criteria and techniques based on threshold calibration and overriding decisions for sensitive samples, which we show to have a significant effect on bias reduction. The latter only use information found in the original dense model.

2. Methodology

2.1. Notions of Bias

We now define the notions of bias we will use in the rest of the paper. We emphasize these categories should not be seen as exclusive: instead, they allow us to study different aspects of the given phenomena.

Systematic Bias. A standard, broad meaning of bias is *systematic error* [12]: for example, we can measure whether models are biased toward overconfidence in their predictions, or if they tend to generalize poorly to data from a shifted distribution. We call this *Systematic Bias*; a full list of the metrics we use is given in section 2.3.

Category Bias. A complementary approach to defining bias centers around the notion of *subgroup/category* of samples in the dataset. Here, bias refers to violations of group fairness metrics with respect to given categories [2] for instance by measuring differences in false positive, false negative, or error rates across subgroups. Other related metrics are worst subgroup performance [47], or the standard deviation of accuracy across identity categories [40].

Inherent to these definitions is that the choice of attributes that define the subgroups must be *meaningful* in a sociological context and *relevant* to the model's application. For example, it is appropriate to measure the accuracy difference with respect to race and gender in facial identification software, since even a moderate difference in accuracy can lead to discrimination in real-world settings. Models that are highly-accurate on standard metrics, e.g. top-1 accuracy, may still be considered biased, for instance with respect to demographic parity. In order to distinguish the concept of *bias* from that of *fairness*, here we focus on *algorithmic bias*, which we define as cases in which a model amplifies bias found in the training data. A classic example is when a model tends to have worse accuracy on samples from poorly-represented subgroups of the dataset. We call this type of bias *Category Bias*.

These notions are complementary: category biases are likely associated with systematic biases, and therefore, studying systematic bias can help us understand cases where models show socially-relevant category bias. This is a common assumption that is frequently used to study bias, for instance in the work on compression-identified exemplars of [28, 29], which first identifies a consistent set of examples on which compressed models frequently struggle, and then demonstrates that these are enriched for certain identity groups. Generally, we are also interested in understanding the relationship between statistical notions of bias, examined via specific metrics, and potential systematic bias across protected categories.

2.2. Category Bias Metric: Bias Amplification

Following prior work [29, 54], we consider datasets where samples are classified according to binary attributes, and use a subset of these as "identity" attributes. For this, we introduce as our main metric a variant of *Bias Amplification (BA)* [54]. Intuitively, bias amplification will measure the extent to which correlations between identity categories and predicted attributes in the training data are exaggerated by the model. While positive correlation between an identity category and a predicted attribute can be reasonable (a model can predict that women wear earrings more frequently than men), models that *amplify* such input relationships in their output may be stereotyping, by relying on identity markers as a proxy for other attributes.

To encode this formally, we compute bias amplification. We define the function $N(\cdot)$ to provide the *count* of the number of samples with a specific binary attribute value, e.g. Young $= 1$, over a given sample set. We then define the bias b of a binary attribute $A \in \{0, 1\}$ with respect to a binary identity category $I \in \{0, 1\}$ as

$$b = \frac{N(A = 1, I = 1)}{N(A = 1)},$$

if the attribute and identity category are positively correlated in the training data, and

$$b = \frac{N(A = 1, I = 0)}{N(A = 1)}, \text{ otherwise.}$$

The bias *amplification* is then the difference between the bias computed on the predicted attribute $\tilde{A}$ and the true value of the attribute A, computed on the test set:

$$BA = \frac{N(\tilde{A} = 1, I = 1)}{N(\tilde{A} = 1)} - \frac{N(A = 1, I = 1)}{N(A = 1)},$$

if the predicted attribute is positively correlated with the identity category, and

$$BA = \frac{N(\tilde{A} = 1, I = 0)}{N(\tilde{A} = 1)} - \frac{N(A = 1, I = 0)}{N(A = 1)},$$

if the predicted attribute is negatively correlated with the identity category. We do not compute the Bias Amplification on any attribute that is not significantly biased toward either value of the identity category, or if some combination of the predicted and protected attribute is very infrequent (e.g., occurring less than 10 times in the test data).

Discussion. This metric has several advantages. Firstly, it is clear that high BA values signal stereotyping by the model. Unlike the original BA metric of [54], our definition uses the label distribution in the *test data* as the true baseline for the predicted label distribution of the model, allowing us to separate the effect of the model itself from the effect of the underlying data, and also allowing us to test the model for bias in settings where the test distribution does not closely resemble the training distribution.

Additionally, BA is not directly affected by other possible biases in the model, such as a tendency to underpredict rare attributes. Moreover, unlike direct false-positive/negative analysis, BA directly takes into account predictions over both values of the protected attribute, and can be meaningfully aggregated across attributes.

2.3. Systematic Bias Metrics

We use several other fine-grained metrics to measure the systematic bias of dense and sparse models.

Threshold Calibration Bias (TCB). On many datasets, the majority of attributes are not evenly split across samples: e.g., for CelebA, the average imbalance is 80%/20%. We measure the change (typically, decline) of the proportion of predictions into the less common value of the attribute using the default threshold. Note that values near 1 show minimal TCB, while values away from 1 in either direction show higher TCB.

$$TCB = \begin{cases} \frac{N(\tilde{A}=1)}{N(A=1)}, & \text{if Mean}(A) < 0.5 \\ \frac{N(\tilde{A}=0)}{N(A=0)}, & \text{otherwise.} \end{cases}$$

Uncertainty and Calibration. Attribute predictions after applying the sigmoid function range between 0 and 1. For a converged model, they tend to cluster around the extremes, with some smaller number of predictions falling nearer the center of the interval. We consider prediction values between 0.1 and 0.9 to be *uncertain*. These uncertainty metrics simply compute the proportion of predictions that fall into the uncertain interval. We then check if the uncertainty correctly estimates the proportion correct by bucketing [8, 45]. The prediction range is split into ten equal-width buckets, and average per-bucket difference of the confidence and the proportion correct. These are then weighted by the bucket size and aggregated. The weighted average difference of the accuracy and confidence of the buckets is presented as the Expected Calibration Error (ECE).

$$ECE = \sum_{m=1}^{10} \frac{|B_m|}{\sum_{n=1}^{10} |B_n|} \left| \text{acc}(B_m) - \text{conf}(B_m) \right| .$$

Label Interrelation. Finally, we look at the strength of relationship between predicted labels on the various attribute. Specifically, for each attribute A, we train a linear regression using all other attributes as the features and A as the variable to be predicted; the coefficient of determination (R^2) of this model tells us the extent to which the model output for A can be predicted from the model outputs of the other attributes in a co-trained model. Note that this does not imply a causal relationship - we cannot say that the model is using some of the attributes to predict others. Rather, a high interrelation suggests that the hidden feature layer is less expressive, forcing a closer relationship between linear classifiers using it as the features.

2.4. Evaluation Setup

CelebA Setup. In our primary study, we focus on ResNet18 [25] models that predict human-annotated binary attributes from cropped-and-centered photos of celebrities in the CelebA dataset [41].

CelebA attribute prediction is frequently used for bias measurement [28, 29, 40, 50]. This is in part due to its size and widespread availability. Yet, CelebA is an imperfect proxy for real-world human photographs, as it skews substantially in both age and skin color, as well as make-up, hairstyles, and overall presentation of the human subjects. As previous works have looked at both models that jointly co-train all or most CelebA attributes [40, 50] and models that train only a single attribute [29], we conduct both types of experiments. For the all-in-one/joint training, we train a ResNet18 model with 40 logistic classifiers after the fully-connected layer. Additionally, we train models with a single head for 7 CelebA attributes: Blond, Smiling, Oval Face, Big Nose, Mustache, Receding Hairline, Bags Under Eyes.

We validate our results by repeating our experiments on the ResNet50 and MobileNetV1 [30] architectures, as well as on structured sparsity (2:4, 1:4 and 1:8) sparsity patterns, which are better supported by current NVIDIA hardware [44]. We also validate some of our findings on the uncropped CelebA dataset, as well as on the iWildcam [3] and Animals with Attributes2 [51] datasets.

For CelebA, we use four attributes for computing Category Bias: “Male”, “Young”, “Chubby”, and “Pale Skin”[1]. These attributes were chosen because they loosely correspond to categories traditionally used to measure bias and discrimination. Examples of these categories can be found in Appendix M. **In the rest of the paper, we use “categories” to refer to these four attributes when they are used as the group identifier to compute BA, and “attribute” to refer to any CelebA attribute that is used as a prediction target.**

Model Architectures. For both ResNet and MobileNet models, we use the standard model architecture, with only one fully-connected layer and a logit transformation following the convolutional blocks, and Binary Cross-Entropy loss. Unlike other studies using CelebA [50], we found that including an additional fully-connected layer did not improve accuracy. Nor did it increase accuracy to initialize with ImageNet weights as in [40, 50], and therefore all models were randomly initialized following [24]. Consis-

[1] The choices to present gender as a binary attribute, and the specific words to describe the attributes were chosen by the creators of the CelebA dataset. We continue their use here to avoid confusion and enable comparisons with other works.

tent with other work, we use the cropped-and-centered version of the dataset described in [41], and perform training data augmentations consistent with [50]. We also validate on the uncropped version. We report results after running each experiment from 5 random seeds.

Model Compression. We perform unstructured pruning, by gradually removing the lowest magnitude weights during training, known as Global Magnitude Pruning (GMP) [18, 21, 22, 55]. GMP is a standard baseline, which, despite its simplicity, is competitive with more complex approaches [17, 18, 34, 35, 49]. We prune all ResNet18 models to 80%, 90%, 95%, 98%, 99%, and 99.5% sparsity. Following earlier work [31], we considered two variants of GMP. The main variant starts from a random initialization (RI), and gradually removes parameters after the tenth training epoch, while simultaneously training the model [55]; we refer to this setup as GMP-RI. The second variant starts from a pre-trained dense model, then gradually removes parameters with the lowest global magnitude while continuing to finetune the model at a lower learning rate; this second variant will be referred to as GMP-PT. We train models using SGD with momentum, with the exception of pre-trained (PT) pruning, for which we found Adam [33] to yield better results. We use the model state at the end of the epoch which reached highest performance on a held-out validation dataset. All the experiments presented are performed for ResNet18 models under the GMP-RI setup; we provide additional validation for GMP-PT in Appendix E, which supports our conclusions.

Our setup makes some complementary choices relative to prior work [28, 29]. Specifically, we prune weights by magnitude *globally* as opposed to *per-layer*. This will allow us to reach much higher sparsity levels relative to [28, 29] before model breakdown. Further, we chose relatively long model training times (100 epochs for 40-attribute dense and GMP-RI models, 80 epochs for GMP-PT models, and 20 epochs for all single-attribute models), as this leads to both higher accuracy and lower bias metrics.

Accuracy Results. Using GMP and an extended training schedule, we are able to obtain sparse models that match or outperform the dense baseline, both in terms of accuracy and ROC-AUC values, even at high ($\geq$ 99%) sparsities, while providing substantial improvements in theoretical FLOPs (computed as in [14]), and practical inference speed on CPU when using the DeepSparse inference engine [10]. We present our results for dense and sparse (GMP-RI) models trained to predict all 40 attributes in Table 1, which show that sparse models can outperform the dense one, even at high sparsities. This is also confirmed by the more robust AUC metric, which is agnostic to the prediction threshold; at all sparsity levels, except for 99.5%, we can observe a slight improvement in AUC scores over the dense models. We observe a similar trend regarding the quality of sparse models over the dense baseline with single-attribute training. This is in contrast to previous work [29], which observes a degradation of sparse models over dense even at 90% sparsity. We believe our improved results are due to the use of a better pruner (global over uniform layer-wise magnitude pruning), and improved training schedule. Nevertheless, they further motivate our study of properties of sparse models, beyond accuracy.

Metric	Dense			Sparsity (%)			
		80	90	95	98	99	99.5
Accuracy (%)	90.4	90.8	91.0	91.3	91.5	91.5	91.1
AUC (%)	80.5±0.2	81.0 ± 0.2	81.3 ± 0.3	81.5 ± 0.2	81.5 ± 0.2	81.0 ± 0.1	79.7 ± 0.1
Inf. FLOPs (B)	3.64	1.40	0.998	0.683	0.386	0.241	0.145
Inf. items/sec	130	138	181	234	318	373	403

Table 1. Average Accuracy AUC, estimated inference FLOPs, and inference times on CPU (using the DeepSparse Engine [36]) for ResNet18 models jointly trained on all 40 binary attributes. We report results after running each experiment from 5 random seeds. For better readability, we present AUC scores as percentages. We omit variances for the accuracies, as they are all ≤ 0.1.

Additionally, we examined randomly-selected images in each category manually, to validate the quality of the human ratings and the images presented to the automated classifier (see Appendix M for screenshots). We provide our tool, the example viewer, for the convenience of other researchers.

3. The Effects of Sparsity on Bias

3.1. Baseline: Analysis of Dense Models

Systematic Bias in Uncompressed Models. Examining bias in dense models, we find that, when jointly-trained across all attributes, they tend to under-predict the less prevalent output value for each attribute, with an average TCB of 0.9. Models trained on a single attribute have a worse under-prediction error than jointly-trained models at lower sparsities; for instance, predictions for Oval Face had a TCB of 0.84 when trained jointly with all other attributes, but 0.52 when trained singly. Additionally, dense models were overconfident with respect to the prediction probability, with an average ECE of 0.054 for jointly-trained models. Single-attribute dense models showed higher uncertainty (Figure 3 and Appendix C), despite having higher accuracy than jointly-trained models.

Category Bias in Uncompressed Models. Dense models exhibit non-trivial bias amplification (BA), for both singly and jointly-trained attributes. The results show two trends. The first, shown in Figure 1 (left), is that BA is substantially higher with respect to specific categories: for instance, with respect to Male and Young, relative to Chubby and Pale Skin. The attributes with highest BA value for dense joint training are Double Chin (Male, 0.053), Wavy Hair (Male, 0.047), Wearing Necktie (Young, 0.046), Pointy Nose (Male, 0.045), Chubby (Male, 0.043), and Oval Face (Male, 0.042). (See Appendix J for a full table.) These attributes rank in the top five for several identity categories, suggesting that they are prone to correlations.

The second trend is that single-attribute training shows a much higher BA than joint training. (See the bottom row of Figure 3, 0% sparsity.) For instance, BA with respect to ‘Male’ is about three times higher when training singly rather than jointly in the case of Oval Face and Big Nose (0.15 vs 0.04 and 0.11 vs 0.03).

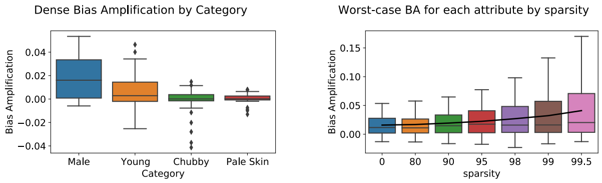

Figure 1. (Left) Bias Amplification by category for dense ResNet18 CelebA models. (Right) Distribution of Worst-Case Bias Amplification across identity categories, for all attributes and sparsities, CelebA on ResNet18.

Discussion. It appears that both compressed and uncompressed models are still prone to bias amplification. From the point of view of our analysis, the presence of bias in the dense model allows us to compare against sparse models.

Manual Review of Celeb-A Samples. It is tempting to ascribe intuitive explanations to the above correlations. However, examining the above attributes more closely, we observe that they have low accuracy and high uncertainty values. Inspecting randomly chosen images, we noticed that attributes such as Pointy Nose often appear difficult to classify, even for human raters. Others, such as Wearing Necktie, are often *impossible to observe directly* on the *cropped version* of the image typically used for this task[2]. Finally, an inspection of images shows that Wearing Lipstick appears difficult to judge from the appearance of the mouth, without relying on indirect information, such as the person’s gender, or presence of other makeup. Thus, even though we do not detect large bias amplification for this attribute, we consider this measurement unreliable. See Appendix M for examples from these categories.

3.2. The Effect of Sparsity on Systematic Bias

Figure 2 shows the effect of pruning CelebA models jointly-trained on all attributes on systematic bias, in the random initialization (RI) setup. First, notice that, as we increase model sparsity, accuracy stays largely unchanged. Yet, other characteristics of the model change considerably. Threshold Calibration Bias (TCB) worsens with sparsity for jointly trained models, with an ever-lower proportion of predictions of the less popular value of each attribute. (Consider that the average TCB for dense models is 0.90, while for 99.5%-sparse models it is 0.81.) Uncertainty goes up considerably for almost every attribute, roughly doubling from dense to 99.5%-sparse models.

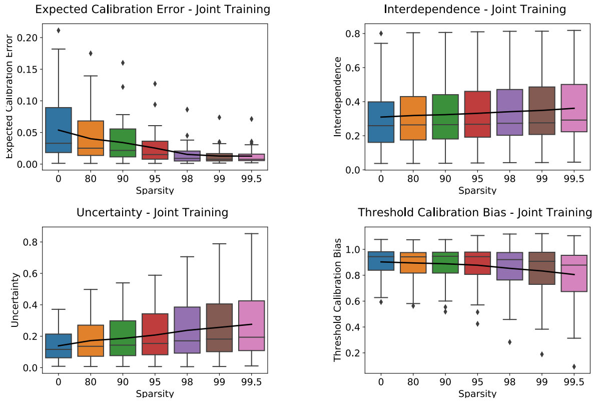

Figure 2. Systematic Bias metrics (TCB, ECE, Interdependence) of ResNet18 models jointly trained on all CelebA attributes. The thick black line denotes the mean value at each sparsity level. In this and all boxplots, the horizontal line represents the median across all CelebA attributes, the edges of the box denote the 25^{th} and 75^{th} quartiles, and dots indicate all points more than 2.5 times the distance from the mean to the respective quartile.

Combining these two observations, we note that in our experiments, jointly-trained sparse models are *better calibrated* than dense with an average ECE of 0.013 for jointly-trained 99.5% sparse models versus 0.054 for dense models. (Note that [8] observe similar behavior of ECE for Lottery Tickets [16], at lower sparsity, and on different datasets.) Finally, label interdependence increases with sparsity, from an average R^2 of 0.31 to 0.36, suggesting that the more compact feature representation in sparse models results in greater entanglement between the features for every attribute.

For singly-trained models, uncertainty is largely unchanged as sparsity increases, perhaps due to already having high values in the dense model, relative to the jointly-trained model. In effect, jointly-trained models have lower uncertainty than singly-trained ones at lower sparsities, but roughly equal uncertainty at higher sparsities. (See Figure 3 and Appendix C for full data.) Threshold Calibration Bias confirms this trend: TCB is roughly constant with sparsity for singly-trained models, but gets worse (decreases) for jointly-trained models. Thus, jointly-trained models are less miscalibrated at lower sparsities relative to singly-trained ones, but similarly miscalibrated at higher sparsities.

3.3. The Effect of Sparsity on Category Bias

Next, we focus on the effect of sparsity on bias amplification. Here, the expectation is that, if sparse models exhibit more bias, for instance by picking up on spurious correlations, bias amplification should increase. We first examine this trend in Figure 1 (right), for jointly-trained models. We

[2] Human raters were asked to assign labels using the uncropped version of the image.

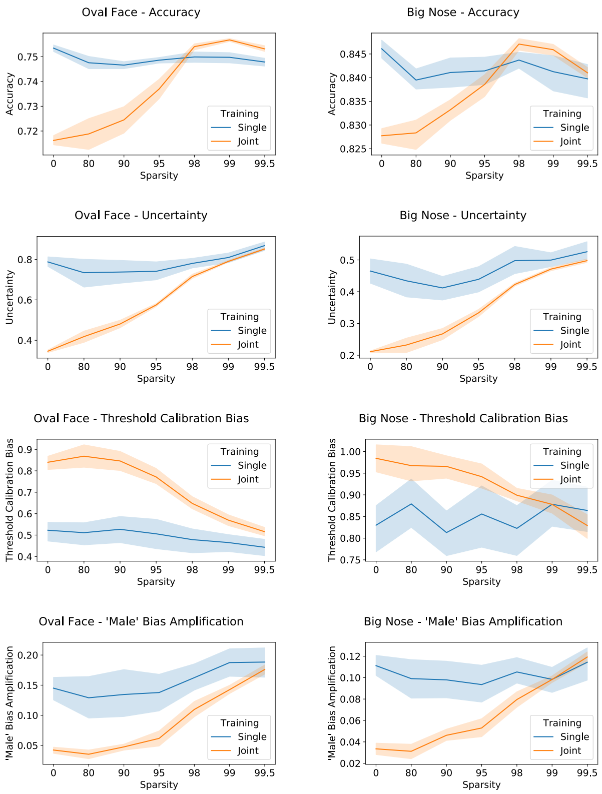

Figure 3. Effect of single versus joint training of attributes on accuracy (first row), uncertainty (second row), Threshold Calibration Bias (third row), and Bias Amplification for the 'Male' attribute (fourth row), on the ResNet18 CelebA model, predicting Oval Face (left) and Big Nose (right).

observe that BA presents a slight increase w.r.t. sparsity between 90 and 95%, after which the increase is more pronounced. The values for BA at the highest sparsity levels are largely determined by the BA values of dense models, with a coefficient of determination $R^2 = 73.2$.

In contrast, when we examine runs with *single-attribute training* (bottom row of Figure 3 and Appendix C), we observe that, in this case, sparsity has very little effect on bias amplification for the hidden 'Male' category, which stays roughly constant, within noise bounds. However, recall from our previous discussion that the baseline (dense) bias amplification is significantly higher for single-attribute training relative to jointly-trained attributes. Specifically, BA for *dense singly-trained models* is roughly as high as for *99.5%-sparse jointly-trained* models. One interpretation is that the additional prediction heads of the jointly-trained models encourage a more robust feature representation which *discourages bias at low sparsity*; at high sparsity, however, the compactness of representation induces more bias. Thus, switching to singly-trained attributes may be a good strategy at high sparsity levels.

Another observation is the high correlation between the evolution of *uncertainty* (second row in Figure 3), TCB (third row), and that of bias amplification (fourth row), relative to the sparsity increase. Specifically, the increase in output uncertainty is linked to stronger bias amplification.

We further investigated whether co-training the identity category with the attribute of interest encourages more diversity in the representation. In this case, we observed a very similar trend regarding BA as for singly-trained attributes, which indicates that the source of bias goes beyond the relationship between the two attributes. These results are shown in Appendix D.

3.4. Injecting Backdoor Features in Sparse Models

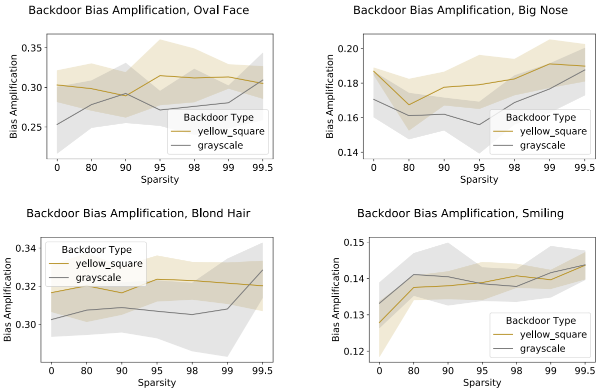

Figure 4. Effect on BA of adding a backdoor feature when performing single-attribute training for four attributes.

To study the amplification of bias by sparse models, we artificially introduce bias in the data through specific modifications to the samples, via "backdoor attacks". We then measure the effects on a similarly "backdoored" test set, for dense and sparse models for single-attribute prediction. We follow a similar approach to [48, 50] for backdooring: we apply a fixed transformation—grayscaling of the entire image [50], or inserting a small yellow square [48] — to the majority of training samples with a positive label, and to a smaller subset of those having the negative label. On the test set, we keep an even ratio of backdoored samples. We perform both the grayscale and yellow square backdoor attacks when training with four separate attributes: Blond, Smiling, Oval Face and Big Nose. We use a backdooring split of 95% positive /5% negative for Blond and Smiling, and 65% positive /35% negative for Oval Face and Big Nose. The smaller split prevents the model from simply memorizing the backdoor on harder tasks.

Targeted backdoors enable us to better control and isolate the source of bias introduced in the models. We consider category bias, and focus on bias amplification (BA) as our main metric. Specifically, in the definition of BA described in Section 2.1 we consider backdooring as our identity category, *i.e.* if a sampled is backdoored, then it has identity category 1, and 0 otherwise.

Our results in Figure 4 show that, as expected, BA increases substantially for all models considered. Moreover, we observe that bias is slightly amplified with sparsity, for example on the Big Nose or Smiling attributes. Overall, our study on bias for backdoored models results in similar conclusions to the "clean" single label experiments. For example, when examining the BA scores for single label training in Figure 3, we notice that the values have only a slight increase with sparsity. This suggests that bias is more likely to follow from less diverse feature representations, whereas here the relationship with sparsity is weaker.

4. Mitigating Sparsity-Induced Bias

4.1. Threshold Calibration

Inspired by our earlier observation that sparser models tend to show worse threshold calibration bias, we consider what happens when we adjust the thresholds to better fit the true distribution of each attribute. We note that the decision to adjust the threshold is not clear-cut; the logistic loss encourages the correct prediction, rather than the correct *ranking* for each attribute. Further, the threshold adjustment does not take the identity feature into account, and should not be confused with fairness-aware threshold adjustments [23]. Instead, we set a single threshold for each attribute so that the predictions are correctly calibrated on the original CelebA validation set.

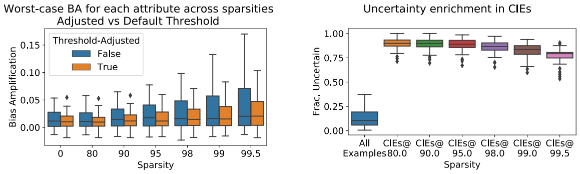

Figure 5. (Left) Effect of threshold calibration on ResNet18 models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

The results of threshold calibration are shown in Figure 5 (Left). Despite the fact that the threshold adjustment process is agnostic to identity categories, this simple correction reduces the bias amplification across all sparsities, almost eliminating bias effects at up to 90% sparsity.

4.2. Overriding Sensitive Samples

Since the additional bias amplification in sparse models must be due to test samples whose classification has changed between dense and sparse models, we examine these examples more closely. We focus on Compression-Identified Exemplars (CIEs) [28, 29], which are the test examples on which the modal dense label across multiple training runs disagrees with the modal sparse label, regardless of which one is correct. For each sparsity, we compute the CIEs across five runs each of the dense and sparse models. Our results in Figure 5 (Right) show that CIEs are greatly enriched for prediction uncertainty, suggesting that improving the predictions of these examples may assist in reducing BA, especially in the sparse models. However, CIEs are expensive to compute due to requiring multiple models for consensus, and are specific to the sparsity level.

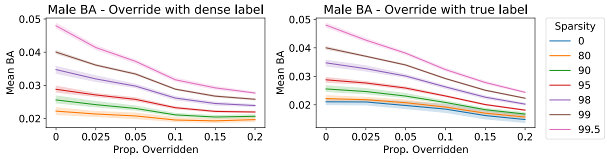

Figure 6. Effect of label overrides on Male Bias Amplification.

Prediction overrides, where a fixed label for a small subset of data is distributed along with the model, and selected over the model prediction at inference time, are common in model deployment. Inspired by our observation that CIEs are highly enriched for uncertain examples, we propose to prioritize the highest-uncertainty data as classified by a dense model, in cases where the dense model already shows positive BA. We replicate this setting on the test dataset. This is consistent with standard practice for override prioritization to improve accuracy, since the most uncertain examples are presumed to have the highest chance of having the wrong label.

We consider two possible override labels: the correct label, which simulates human overrides, and the dense label, which simulates the best possible label if human labeling is impractical. We apply these overrides to all sparse labels and measure the bias amplification. Our results (Figure 6 and Figure B.1) show that overrides with both human and dense labels substantially decrease the bias amplification of models of all sparsities. For instance, using manual overrides for the most uncertain 5% of examples lowers the mean BA of the 99.5% sparse model by 23%, and replacing the top 10% lowers the mean BA by 35%. This suggests that the use of uncertainty-based override pipelines is an effective tool for reducing bias amplification on sparse models, even when only the dense model is used to set prioritization.

5. Additional Validation

We emphasize the fact that the above observations have been validated on additional datasets and models, so our findings hold generally. We discuss these experiments briefly below, and present them in full in the Appendix.

Additional Validation on CelebA. We experiment with the setup where pruning starts from a pretrained model, for which we include the results in Appendix E, showing similar results. We additionally prune to N:M (2:4, 1:4 and 1:8) sparsity patterns [44] in Appendix F, with similar results to lower-sparsity models pruned without this restriction. Experiments validating our results for singly- and jointly-trained attributes on the MobileNetV1 architecture [30] can be found in Appendix G, showing the same trends, but at slightly lower sparsities. We additionally validate the joint training results on the ResNet50 architecture in Appendix H, with very similar results to ResNet18. Finally, we repeat the ResNet18 joint training experiments using the *uncropped* CelebA dataset, which ensures that features such as the presence of neckwear are available to the model (as they were to the human labellers). We discuss these results in Appendix I.

Additional Datasets. We further validated our findings on two additional datasets. The Animals with Attributes (AwA) dataset [51] serves as a useful validation for our observations regarding the effect of sparsity on bias in binary prediction (Appendix K). The challenging iWildcam dataset [3] validates our observations regarding increased uncertainty relative to sparsity in the context of multiclass classification (Appendix L).

6. Related Work

Fairness, Bias, and Bias Mitigation. A number of fairness metrics have been proposed, including individual fairness, which requires that individuals with similar characteristics receive similar outcomes, and group fairness, which requires parity along some metric between individuals in commonly-identified groups [2]. Many works propose techniques to remove or mitigate bias in general [47, 50], while [40] mitigates accuracy bias on compressed models. Notably, [50] proposes the use of synthetic benchmarks such as backdooring images. Backdooring is also used by [48] for evaluating bias in transfer learning.

Bias Due to Compression. Seminal work by Hooker et al. [28, 29] initiated the study of compression-induced bias, showing that bias can be amplified by model pruning, and isolate the influence of Compression Identified Exemplars (CIEs) as rare examples in the training data. Our work significantly extends this research, by examining compression effects via Bias Amplification, and showing that highly-sparse models may in fact be bias-free for moderate $\leq 90\%$ sparsities, using joint training, global pruning, and additional finetuning. In addition, we provide strategies for bias mitigation that do not require knowledge of identity categories, nor any information about compressed models.

Recent work by Chen et al. [8] studies pruning effects from four aspects: generalization/robustness to distribution shifts, prediction uncertainty, interpretability, and loss landscape, for pruned models obtained via variants of the Lottery Ticket Hypothesis (LTH) approach [7, 9, 16]. They show that LTH-pruned models match (or slightly outperform) dense models across all these categories. Our work is related in that they also study prediction uncertainty for models, noticing that sparse LTH models can be competitive with dense ones in terms of uncertainty, measured as ECE. Yet, the focus of our work is different: we perform an in-depth comparison of bias effects, specifically focusing on the high-sparsity range, where we exhibit and carefully analyze the emergence of bias. In addition, we provide a set of techniques for characterizing and mitigating bias in pruned models, which is beyond the scope of [8].

Good et al. [20] studies the relative distortions in the recall of a model in relationship with sparsity, and proposed a gradient-based pruning method to decrease the negative effect of sparsity on this metric. Other works analyze the variance in classification error among classes as a proxy for bias in sparse models [6], while others [32, 52] use knowledge distillation [26] to decrease the misalignment between sparse and dense models. By comparison, our study focuses on characterizing and mitigating bias given a fixed compression scheme, for which we propose different metrics, as well as detection criteria and countermeasures.

Systematic Bias. Finally, systematic bias is an important avenue of research that compliments our work by using more sophisticated techniques to identify and categorize hard-to-learn examples [1,11,15,46]. However, these works use finer-grained definitions of systematic bias, and do not consider model compression.

7. Conclusion

We performed an in-depth study of bias in sparse models, and showed that it is possible to obtain highly-sparse models without loss in accuracy or AUC. However, these models have higher uncertainty compared to dense ones, and the predicted labels are more interdependent. Bias amplification is often substantially exacerbated at high sparsities ($\geq 95\%$) and the bias of individual attributes in sparse models correlates well with their bias in the dense baseline. However, the effect we observe on both systematic and category bias is influenced by the training setting, i.e. joint or individual attribute training. In future work, we plan to examine the impact of different compression approaches (pruning and quantization techniques) on our bias metrics, more complex countermeasures for mitigating the bias we have shown to arise in highly-compressed models, and further applications, such as language modelling.

Acknowledgments

The authors would like to sincerely thank Sara Hooker for her feedback during the development of this work. EI was supported in part by the FWF DK VGSCO, grant agreement number W1260-N35. AP and DA acknowledge generous ERC support, via Starting Grant 805223 ScaleML.

References

[1] Robert J. N. Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In *NeurIPS*, 2021. 8

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`. 1, 2, 8

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020. 3, 8, 40

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 2021. 1

[5] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 1

[6] Cody Blakeney, Nathaniel Huish, Yan Yan, and Ziliang Zong. Simon says: Evaluating and mitigating bias in pruned neural networks with knowledge distillation. *arXiv preprint arXiv:2106.07849*, 2021. 8

[7] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained BERT networks. *arXiv preprint arXiv:2007.12223*, 2020. 8

[8] Tianlong Chen, Zhenyu Zhang, Jun Wu, Randy Huang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Can you win everything with a lottery ticket? *Transactions on Machine Learning Research*, 2022. 3, 5, 8

[9] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 8

[10] DeepSparse. NeuralMagic DeepSparse Inference Engine, 2021. 4

[11] Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. 8

[12] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Technical report, Department of Computer Science, Oregon State University, 1995. 2

[13] Erich Elsen, Marat Dukhan, Trevor Gale, and Karen Simonyan. Fast sparse convnets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14629–14638, 2020. 1

[14] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning (ICML)*, 2020. 4

[15] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. ICLR, 2022. 8

[16] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. 5, 8

[17] Elias Frantar, Eldar Kurtic, and Dan Alistarh. M-FAC: Efficient matrix-free approximations of second-order information. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 4

[18] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019. 1, 4

[19] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021. 1

[20] Aidan Good, Jia-Huei Lin, Hannah Sieg, Mikey Ferguson, Xin Yu, Shandian Zhe, Jerzy Wieczorek, and Thiago Serra. Recall distortion in neural network pruning and the undecayed pruning algorithm. *ArXiv*, abs/2206.02976, 2022. 8

[21] Masafumi Hagiwara. A simple and effective method for removal of hidden units and weights. *Neurocomputing*, 6(2):207 – 218, 1994. Backpropagation, Part IV. 1, 4

[22] Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 1, 4

[23] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016. 7

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8

[27] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *arXiv preprint arXiv:2102.00554*, 2021. 1

[28] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019. 1, 2, 3, 4, 7, 8, 16

[29] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising Bias in Compressed Models. *arXiv:2010.03058*, 2020. 1, 2, 3, 4, 7, 8, 16, 40, 41

[30] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3, 8, 20

[31] Eugenia Iofinova, Alexandra Peste, Mark Kurtz, and Dan Alistarh. How well do sparse imagenet models transfer? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[32] Vinu Joseph, Shoaib Ahmed Siddiqui, Aditya Bhaskara, Ganesh Gopalakrishnan, Saurav Muralidharan, Michael Garland, Sheraz Ahmed, and Andreas Dengel. Going beyond classification accuracy metrics in model compression. *arXiv preprint arXiv:2012.01604*, 2020. 8

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[34] Eldar Kurtic and Dan Alistarh. Gmp*: Well-tuned global magnitude pruning can outperform most bert-pruning methods. *arXiv preprint arXiv:2210.06384*, 2022. 4

[35] Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *arXiv preprint arXiv:2203.07259*, 2022. 4

[36] Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Bill Nell, Nir Shavit, and Dan Alistarh. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In *International Conference on Machine Learning (ICML)*, 2020. 4

[37] Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. Soft threshold weight reparameterization for learnable sparsity. In *International Conference on Machine Learning (ICML)*, 2020. 11

[38] Lucas Liebenwein, Cenk Baykal, Brandon Carter, David Gifford, and Daniela Rus. Lost in Pruning: The Effects of Pruning Neural Networks beyond Test Accuracy. *Conference on Machine Learning and Systems (MLSys)*, 2021. 1

[39] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification, 2022. 11

[40] Xiao-Ze Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. *ArXiv*, abs/2207.10888, 2022. 1, 2, 3, 8

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1, 3, 4

[42] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 2021. 1

[43] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating Algorithmic Bias through Fairness Attacks. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 1

[44] Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378*, 2021. 3, 8, 18

[45] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 3

[46] Nazneen Rajani, Weixin Liang, Lingjiao Chen, Meg Mitchell, and James Zou. Seal: Interactive tool for systematic error analysis and labeling. *arXiv preprint arXiv:2210.05839*, 2022. 8

[47] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 2, 8

[48] Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*, 2022. 6, 8

[49] Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 4

[50] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, Prem Qu Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 4, 6, 8, 11

[51] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 8, 39

[52] Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. Beyond preserved accuracy: Evaluating loyalty and robustness of bert compression. 2021. 8

[53] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. 2020. 1

[54] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. 2

[55] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017. 1, 4, 11