SELECTIVE BRAIN DAMAGE: MEASURING THE DISPARATE IMPACT OF MODEL PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural network pruning techniques have demonstrated it is possible to remove the majority of weights in a network with surprisingly little degradation to top-1 test set accuracy. However, this measure of performance conceals significant differences in how different classes and images are impacted by pruning. We find that certain examples, which we term *pruning identified exemplars (PIEs)*, and classes are systematically more impacted by the introduction of sparsity. Removing PIE images from the test-set greatly improves top-1 accuracy for *both* sparse and non-sparse models. These hard-to-generalize-to images tend to be mislabelled, of lower image quality, entail abstract representations, atypical examples or require fine-grained classification.

1 INTRODUCTION

Between infancy and adulthood, the number of synapses in our brain first grow and then fall. Synaptic pruning improves efficiency by removing neurons that are redundant and strengthening synaptic connections that are useful for the environment (Rakic et al., 1994). Despite losing 50% of all synapses between age two and ten, the brain continues to function (Kolb & Whishaw, 2009; Sowell et al., 2004). "Use it or lose it" is frequently used to describe the environmental influence of the learning process on synaptic pruning, however there is little scientific consensus on *what* exactly is lost (Casey et al., 2000).

In this work, we ask what is *lost* when we prune a deep neural network. Work on pruning deep neural networks has demonstrated a remarkable ability to sparsify a model to a fraction of the original weights while giving up minimal test-set accuracy (Cun et al., 1990; Hassibi et al., 1993b; Han et al., 2015; Ullrich et al., 2017; Liu et al., 2017; Louizos et al., 2017; Collins & Kohli, 2014; Weigend et al., 1991; Nowlan & Hinton, 1992; Lee et al., 2018b). Gale et al. (2019) show it is possible to prune 90% of all weights in a ResNet-50 network (He et al., 2015) trained on ImageNet (Deng et al., 2009) and only lose less than 3% absolute in top-1 test set accuracy. The ability to prune networks with seemingly so little degradation to accuracy is puzzling. In this work, we address the following questions:

- *Are certain types of examples or classes disproportionately affected by pruning?*
- *How does pruning impact robustness such as sensitivity to image corruptions (blur, noise, contrast) and adversarial examples?*

Answers for these question can provide intuition into the role of additional capacity in deep neural networks and, perhaps more important, provide a principled framework for articulating the trade-offs incurred by compressing deep neural networks. Many of the most promising use cases for compressed models occur in sensitive domains, such as improving access to health care by using machine learning driven diagnostics on mobile phones (Esteva et al., 2017). Pruned or compressed models are frequently favored for deploying deep neural networks onto devices because reducing the number of network weights lowers energy consumption, memory footprint, and latency (Reagen et al., 2016; Chen et al., 2016; Theis et al., 2018; Kalchbrenner et al., 2018; Valin & Skoglund, 2018). For tasks where incorrect predictions can harm human welfare, it is critical that we understand when a machine learning model is qualified to make decisions on real world inputs. To our knowledge, this is the first work to shed light on the trade-offs pruning incurs by considering new measures beyond test accuracy.

The primary findings of our work can be summarized as follows:

1. *In some settings, pruning has a non-uniform impact across classes; a fraction of classes are disproportionately and systematically impacted by the introduction of sparsity. .*
2. *The examples most impacted by pruning, which we term Pruning Identified Exemplars (PIEs), are more challenging for both sparse and non-sparse models to classify.*
3. *Pruning makes deep neural networks less robust to natural adversarial examples and common image perturbations (such as blur, fog and contrast).*

For (1) and (2), we establish consistent findings for different standard architectures on CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009). We identify *Pruning Identified Exemplars* (PIEs) as images where the modal label differs between a set of sparse and non-sparse models. We find that removing PIEs from the test-set improves top-1 accuracy for *both* a fully parameterized non-sparse model as well as a sparse model (Figure 4).

Toward finding (3), we measure changes to model sensitivity to both common image corruptions and natural adversarial examples using two open source robustness benchmarks: ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-A (Hendrycks et al., 2019).We find that both sparse and non-sparse models are brittle and sensitive to ImageNet-A and C, and that this brittleness is amplified at highers level of sparsity (Table 7).

Our findings provide important insights about when pruned models are qualified to make decisions on real world inputs. Our PIE methodology identifies a tractable subset of images which are more challenging for sparse and non-sparse models. Our work suggest that PIE could be useful for tasks where it is important to choose not to classify certain examples when the model is uncertain (Bartlett & Wegkamp, 2008; Cortes et al., 2016b;a; Cortes et al., 2017), to aid interpretability as a case based reasoning tool to explain model behavior (Kim et al., 2016; Gurumoorthy et al., 2017; Caruana, 2000; Hooker et al., 2018) or to surface atypical examples for further human inspection (Leibig et al., 2017).

2 A Comparison of Sparse and Dense Models

We consider a supervised classification problem where a deep neural network is trained to approximate the function F that maps an input variable X to an output variable Y, formally $F : X \mapsto Y$. The model is trained on a training set of N images $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, and at test time makes a prediction y_i^* for each image in the test set. The true labels y_i are each assumed to be one of C classes, such that $y_i = [1, ..., C]$.

A reasonable response to our desire for more compact representations is to simply train a network with fewer weights. However, as of yet, starting out with a compact dense model has not yielded competitive test-set performance. Instead, research has centered on training strategies where models are initialized with "excess capacity" which is then subsequently removed through a pruning process. A pruning method $\mathcal{P}$ identifies the subset of weights to remove (i.e. set to zero). A sparse model function, $\hat{F}_t^{\mathcal{P}}$, is one where a fraction $t \in [0, 1]$ of all model weights are set to zero. Equating weight value to zero effectively removes the contribution of a weight as multiplication with inputs no longer contributes to the activation. A non-sparse model function, $\hat{F}_0$, is one where all weights are trainable.

ImageNet and CIFAR-10 Setup We consider two classification tasks and models; a wide ResNet model (Zagoruyko & Komodakis, 2016) trained on CIFAR-10 and a ResNet-50 model (He et al., 2015) trained on ImageNet, both using batch normalization (Ioffe & Szegedy, 2015). A key goal of our analysis is to produce findings that are not anecdotal as would be the case when analyzing one trained model of each type. Instead we independently train a population of 30 models for each experimental setting. We train for $32,000$ steps (approximately 90 epochs) on ImageNet with a batch size of 1024 images and $80,000$ steps with a batch size of 128 for CIFAR-10. For ImageNet, the baseline non-sparse model obtains a mean top-1 accuracy of 76.68% and mean top-5 accuracy of 93.25% across 30 models. For CIFAR-10, mean baseline top-1 accuracy is 94.35%.

We prune over the course of training to obtain a target end sparsity level $t \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. For example, $t = 0.9$ indicates that 90% of model weights are removed by pruning, leaving a maximum of 10% non-zero weights.

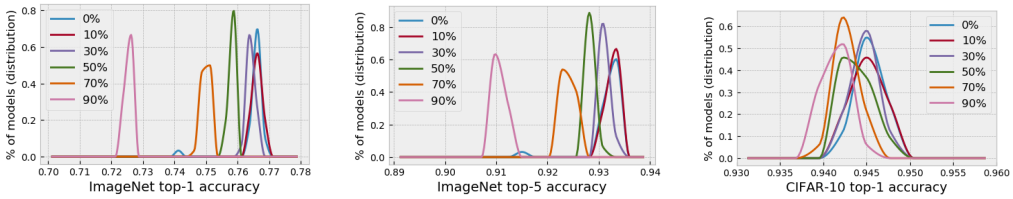

Figure 1: Distributions of top-1 and top-5 model accuracy for populations of independently trained sparse and non-sparse models on ImageNet and CIFAR-10. The distributions for CIFAR-10 top-5 accuracy (not shown) are tightly clustered and overlapping in $[0.9965, 1.0]$. The distributions are fairly tight with one outlier for the ImageNet baseline model.

true label:	dog	airplane	airplane	automobile	truck	horse
non−sparse:	dog	airplane	airplane	automobile	automobile	cat
sparse:	cat	ship	ship	truck	cat	horse

Figure 2: Visualization of pruning identified exemplars (PIE_{30}) for the CIFAR-10 dataset. This subset of impacted images is identified by considering a set of 30 non-sparse wide ResNet models and 30 models trained to 30% sparsity.

Across all experiments, we use magnitude pruning as proposed by Zhu & Gupta (2017) to identify the weights to remove. Magnitude pruning is a simple rule-based method that thresholds weights at zero that fall below a certain absolute magnitude. It has been shown to outperform more sophisticated Bayesian pruning methods and is considered state-of-the-art across both computer vision and language models (Gale et al., 2019). The choice of magnitude pruning also allowed us to specify and precisely vary the final model sparsity for purposes of our analysis, unlike regularizer approaches that allow the optimization process itself to determine the final level of sparsity (Liu et al., 2017; Louizos et al., 2017; Collins & Kohli, 2014; Wen et al., 2016; Weigend et al., 1991; Nowlan & Hinton, 1992). Although the ability to precisely vary sparsity is required for this experimental framework, we note that our methodology can be extended to other methods. Figure 1 shows the distributions of model accuracy across model populations for the non-sparse and sparse models for ImageNet and CIFAR-10.

2.1 Class Level Performance

Comparisons of the effects of different pruning algorithms or levels of sparsity on classification tasks such as ImageNet have largely centered on top-line metrics such as top-1 or top-5 test-set accuracy averaged across classes, but reliance on these measures hides detail in the model's performance. Comparing only top-1 model accuracy between a baseline and a pruned model amounts to assuming that class accuracies are expected to maintain their relative relationships to the top-1 model accuracy before and after pruning. In this work, we consider whether this is a valid assumption. *Is relative performance unaltered by pruning or are some classes impacted more than others?*

For a given model, we compute the class accuracy β_t^c for class $c \in \mathcal{C}$ and sparsity $t \in [0, 1]$. We compute overall model accuracy from the set of class metrics:

$$\beta_t^{\mathcal{M}} = \frac{1}{N_{\mathcal{M}}} \sum_{c \in \mathcal{C}} \beta_t^c * N_c$$

where N_c is the number of examples in class c and $N_{\mathcal{M}}$ is the total number of examples in the data set. If the impact of pruning was uniform, we would expect each class accuracy to shift by the same number of percentage points as the difference in top-1 accuracy between the sparse and non-sparse model. This forms our **null hypothesis** (H_0) – the shift in accuracy for class c before and after

pruning is the same as the shift in top-1 accuracy. For each class c we consider whether to reject H_0 and accept the **alternate hypothesis** (H_1) that pruning disparately affected the class's accuracy in either a positive or negative direction:

$$H_0 : \beta_0^c - \beta_0^{\mathcal{M}} = \beta_t^c - \beta_t^{\mathcal{M}}$$
$$H_1 : \beta_0^c - \beta_0^{\mathcal{M}} \neq \beta_t^c - \beta_t^{\mathcal{M}}$$

Neural net training is most often done in a non-deterministic fashion, and we consider each model k in its population of $K = 30$ models to be a sample of some underlying distribution. Given a class c and a population of K models trained at a sparsity t, we construct a set of samples S_t^c of the mean-shifted class accuracy as $S_t^c = \{\beta_{t,k}^c - \beta_{t,k}^{\mathcal{M}}\}_{k=1}^K$.

Evaluating whether the difference between a sample of mean-shifted class accuracy from sparse and non-sparse models is "real" amounts to determining whether two data samples are drawn from the same underlying distribution, which is the subject of a large body of goodness of fit literature (D'Agostino & Stephens, 1986; Anderson & Darling, 1954; Huber-Carol et al., 2002). In this work, we use a two-sample, two-tailed, independent Welch's t-test (Welch, 1947) to determine whether the means of the samples S_t^c and S_0^c differ significantly. If the two samples were drawn from distributions with different means with 95% or greater probability (p-value $<= 0.05$), then we reject the null hypothesis and consider the class to be disparately affected by t-sparsity pruning relative to the baseline.

After finding the subset of classes for a given t-sparsity that show a statistically significant change relative to the baseline, we can quantify the degree of deviation, which we refer to as *normalized recall difference*, by comparing the average t-sparse and baseline class accuracies after normalizing for their respective average model accuracies:

$$\frac{1}{K_t}\sum_{k=1}^{K_t}\left(\beta_{t,k}^c - \beta_{t,k}^{\mathcal{M}}\right) - \frac{1}{K_0}\sum_{k=1}^{K_0}\left(\beta_{0,k}^c - \beta_{0,k}^{\mathcal{M}}\right) \tag{1}$$

2.2 Pruning Identified Exemplars

How does pruning impact model performance on individual images? A natural extension of the hypothesis testing in the prior section is to consider whether to reject or retain the null hypothesis that the output probability for a given image for a dense and sparse models is equal. However, recent work has highlighted that deep neural networks produce output probabilities that are uncalibrated (Guo et al., 2017; Kendall & Gal, 2017; Lakshminarayanan et al., 2017) and thus cannot be interpreted as a measure of certainty. Deep neural networks do not know what they do not know, and often ascribe high probabilities to out-of-distribution data points or are overly sensitive to adversarially perturbed inputs (Hendrycks & Gimpel, 2016; Nguyen et al., 2014).

We are interested in how model predictive behavior changes through the pruning process. Given the limitations of uncalibrated probabilities in deep neural networks, we focus on the level of disagreement between the predictions of sparse and non-sparse networks on a given image. Let $y_{i,k,t}^*$ be the prediction of the kth t-sparse model of its population for image i where $t = 0$ denotes an unpruned model, and let $Y_{i,t}^* = \{y_{i,k,t}^*\}_{k=1}^K$ be the set of predictions for the t-sparse model population on exemplar i. For set $Y_{i,t}^*$ we find the *modal label*, i.e. the class predicted most frequently by the t-sparse model population for exemplar i, which we denote $y_{i,t}^M$. Exemplar i is classified as PIE_t if and only if the modal label is different between the t-sparse model and the unpruned (non-sparse) model:

$$PIE_{i,t} = \begin{cases} 1 & \text{if } y_{i,0}^M \neq y_{i,t}^M \\ 0 & \text{otherwise} \end{cases}$$

We note that there is no constraint that the non-sparse predictions for PIEs match the true label, thus the detection of PIEs is an unsupervised protocol that could in principal be performed at test time.

Sparsity (t)	Model accuracy diff.	Significant		Largest increase		Largest decrease	
		# incr.	# decr.	class	norm. diff.	class	norm. diff.
0.1	-0.0002	29	22	toaster	0.025	am. chameleon	-0.026
0.3	-0.002	35	34	bathtub	0.036	cleaver	-0.045
0.5	-0.008	91	54	petri dish	0.034	frying pan	-0.047
0.7	-0.017	189	128	cd player	0.050	tow truck	-0.069
0.9	-0.041	337	245	cd player	0.088	muzzle	-0.128

Table 1: Summary of class-level results for ImageNet. Only classes passing the significance test are included. The model accuracy difference column reports mean $\beta_t^{\mathcal{M}} - \beta_0^{\mathcal{M}}$ as the percentage point difference between the pruned and baseline model accuracies; a negative value means the pruned model's average accuracy is lower than the baseline model's.

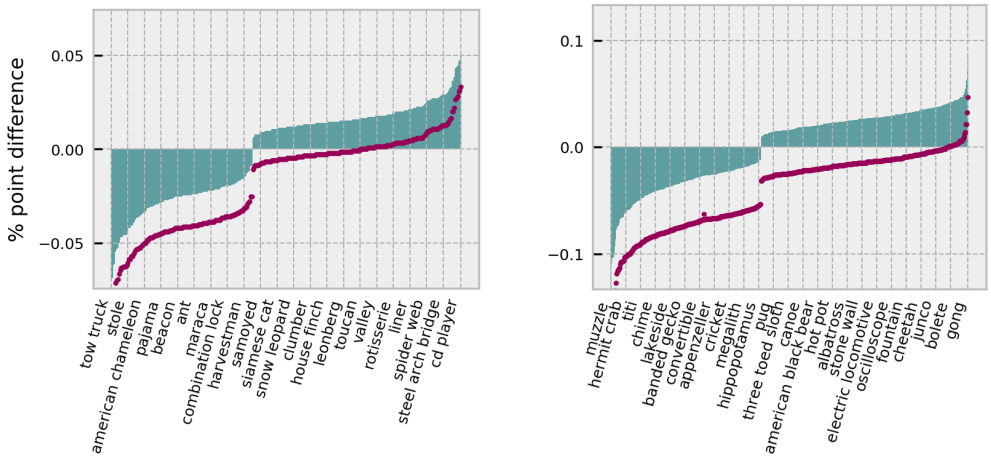

Figure 3: Normalized recall difference (bars) and absolute recall difference (points) per class for 70% sparsity (left) and 90% sparsity (right). The class labels are sampled for readability; there are 317 significant classes for 70% sparsity and 582 significant classes for 90%. Note the difference in scale on the y-axis.

3 Results and Discussion

3.1 Impact of Sparsity on Class Level Performance

The impact of magnitude pruning on ImageNet classification is disparate across classes and amplified as sparsity increases. For example, at 10% sparsity only 51 of 1,000 classes in the ImageNet test set exhibit statistically significant change in class accuracy, however at 90% sparsity, accuracy is impacted for 582 classes in a statistically significant way.

The directionality and magnitude of the impact is nuanced and surprising. Our results show that certain classes are relatively robust to the overall degradation experienced by the model whereas others degrade in performance far more than the model itself. This amounts to selective "brain damage" with performance on certain classes evidencing far more sensitivity to the removal of model capacity. Table 1 shows that more classes show a significant *relative* increase in accuracy than a decrease at every level, though the overall model accuracy decreases at every pruning level, indicating that the magnitude of class decreases must be larger in order to pull the model accuracy lower. Figure 3 visualizes the magnitude of the normalized recall differences for 70% and 90% pruning and highlights the degree to which the classes spread from model average.

We performed the same analysis on the CIFAR-10 models and found that while pruning presents a non-uniform impact there are fewer classes that are statistically significant. One class out of ten was significantly impacted at 10%, 30%, and 50%, and two classes were impacted at 90%. We suspect

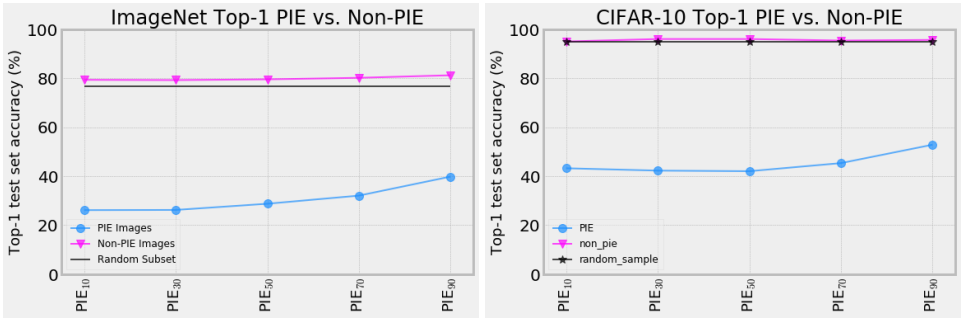

Figure 4: Excluding pruning identified exemplars (PIE) improves test-set top-1 accuracy for both ImageNet and CIFAR-10. This holds for PIE images identified at all levels of sparsity considered. Sparse model find generalizing to PIE more challenging, and inference on PIE images alone substantially degrades generalization performance. **Left:** Average top-1 test-set accuracy across 30 non-sparse ResNet-50 models when inference is restricted to PIE images (blue), non-PIE images (dark purple) and a random sample of the test set (black line) which is a constant independent of PIE sparsity level. **Right:** Average top-1 test-set accuracy across 30 non-sparse wide ResNets trained on CIFAR-10.

that we found less disparate impact for CIFAR-10 because, while the model has less capacity, the number of weights is still sufficient to model the limited number of classes and lower dimension dataset.

3.2 Impact of Sparsity on Individual Exemplars

At every level of sparsity, for both CIFAR-10 and ImageNet we identify a subset of PIE images. At 10% and 90% sparsity, we classify 3.34% and 10.27% of all ImageNet test-set images respectively as PIEs. For CIFAR-10, PIEs constitute 0.97% and 2.16% of the test set at 10% and 90% sparsity.

PIEs are more difficult for both sparse and non-sparse models to classify. In Fig. 4, we compare the test-set performance of a fully parameterized non-sparse model on a fixed number of randomly selected (1) PIE images, (2) non-PIE images and (3) a random sample of the test set. The results are consistent across both CIFAR-10 and ImageNet datasets; removing PIE images from the test-set improves top-1 accuracy for *both* sparse and non-sparse models relative to a random sample. Inference restricted to only PIE images greatly impairs the generalization performance of a model. In the appendix, we include additional plots that show that while all models perform far worse on PIE images, the degradation to performance is amplified as model sparsity increases.

The most challenging PIEs are identified at low levels of sparsity. In Figure 4, the lowest test-set accuracy for both sparse and non-sparse models occurs when inference is restricted to PIEs identified at 10% sparsity. Test-set accuracy steadily increases for PIEs identified at higher levels of sparsity. This suggests that sparsity first erodes performance on the images that the model finds the most challenging.

Why are PIEs impacted more by pruning? A qualitative inspection of PIEs (Figure 6) suggests that these hard-to-generalize-to images tend to be of lower image quality, mislabelled, entail abstract representations, require fine-grained classification or depict atypical class examples. We conducted a limited human study[1] to inspect a random sample of 400 PIE and non-PIE ImageNet images. We broadly group the properties we codify as indicative of 1) the exemplar being challenging or 2) the task being ill-specified. We introduce these groupings below (after each bucket we report the percentage of PIEs and non-PIEs in each category as a fraction of total PIEs and non-PIE codified):

1. Poorly specified task

[1]The authors acted as labelers ahead of the deadline, though the PIE label was shuffled and hidden during labeling. We will reproduce the study using other parties before publication.

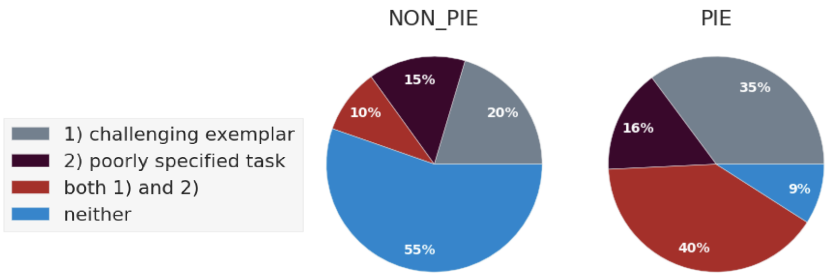

Figure 5: Limited human study of the relative distribution of PIE and non-PIE properties. **Challenging exemplars:** images positively codified as showing *common image corruptions* such as blur or overlaid text, or images where the object is in the form of an *abstract representation* or where the exemplar requires *fine grained classification*. **Poorly specified task:** images where *multiple classes* are visible in the same image, or images with *incorrect or insufficient ground truth.*

- **incorrect ground truth** or **insufficient information –** images where there is not sufficient information for a human to arrive at the correct ground truth label. For example in Fig. 6, the image of the plate of food with the label `restaurant` is cropped such that it is impossible to tell whether the food is in a restaurant or in a different setting. [2.9% of non-PIEs, 9.5% of PIEs]
- **multiple labelled classes are clearly visible in the image** – images depicting multiple objects where a human may consider several labels to be appropriate (e.g., an image which depicts both a `paddle` and `canoe`, `desktop computer` consisting of a `screen`, `mouse` and `monitor`, a `barber chair` in a `barber shop`). [22.3% of non-PIE, 50.8% of PIEs]

2. Challenging Exemplars

- **fine grained classification task –** involves classifying an object that is semantically close to various other classes present the data set (e.g., `rock crab` and `fiddler crab`, `bassinet` or `cradle`, `cuirass` and `breastplate`). [8.3% of non-PIEs, 38.2% of PIEs]
- **exemplar corruptions –** images exhibit common corruptions such as motion blur, contrast, pixelation. We also include in this category images with super-imposed text, an artificial frame and images that are black and white rather than the typical RBG color images in ImageNet. [8.7% of non-PIE, 12.1% of PIE]
- **exemplar abstractions –** the surfaced exemplar depicts a class object in an abstract form such a cartoon, painting, or sculptured incarnation of the object. [1.5% of non-PIE, 3.5% of PIE]

We find that PIEs overindex relative to non-PIEs on certain properties, such as having an *incorrect ground truth label* or *multiple objects*. This suggests that the task itself is often incorrectly specified. Both ImageNet and CIFAR-10 are single image classification tasks, however just over half of the PIEs codified by humans were identified as multi-object images where multiple labels could be considered reasonable (vs. 22.33% of non-PIEs). The over-indexing of incorrectly structured data in PIE hints that the explosive growth in number of parameters in deep neural networks may be solving a problem better addressed in the data cleaning pipeline. However, task mis-specification alone does not fully explain the images where predictive behavior diverges as model capacity is changed. We find that PIE also overindexes on *low quality images* that present corruptions like blur or overlaid text and exemplars that require *fine grained classification*. This suggests that there may be a relationship between number of model parameters and generalization to these types of images. We explore the relationship between capacity and robustness to these types of corruptions more thoroughly in Section 3.3.

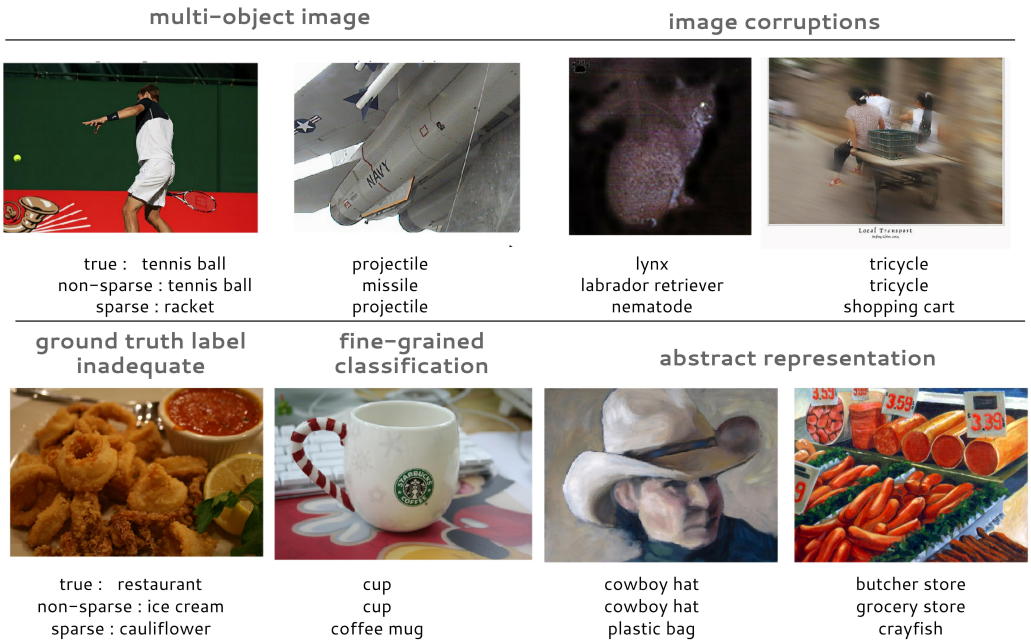

Figure 6: Visualization of PIE_{90}: images where modal label differs between 30 non-sparse ResNet-50 models and 30 models trained to 90% sparsity on ImageNet. A qualitative inspection of PIEs suggests that there are certain shared properties between the images most impacted by pruning; these images tend to be of lower image quality and frequently exhibit image corruptions (overlaid text, contrast, motion blur, zoom), mislabelled or unintuitive label, depict class in abstract form, require fine-grained classification or depict atypical class examples.

3.3 The Role of Additional Capacity

The PIE procedure surfaces exemplars that are harder for both sparse and non-sparse models to classify. Given that PIE surfaces data points where there is the greatest divergence in behavior between sparse and non-sparse models, it is useful to understand the directionality of some of the properties described in the previous section. For example, we noted that PIE frequently surfaces images with corruptions such as motion blur, contrast and pixelation. *Are sparse models more brittle to these type of corruptions?* Many PIEs are often atypical or unusual class examples. We have already noted that model degradation when restricted to inference on PIEs is amplified as sparsity increases. *Does this measure of model brittleness mirror other open source robustness benchmarks?*

ImageNet-C ImageNet-C (Hendrycks & Dietterich, 2019) is an open source data set that consists of algorithmically generated corruptions (blur, noise) applied to the ImageNet test-set. We compare top-1 accuracy given inputs with corruptions of different severity. As described by the methodology of Hendrycks & Dietterich (2019), we compute the corruption error for each type of corruption by measuring model performance rate across five corruption severity levels (in our implementation, we normalize the per-corruption error by the performance of the sparse model on the clean ImageNet dataset). ImageNet-C corruption substantially degrades mean top-1 accuracy of non-sparse models Fig. 7. This sensitivity is amplified at high levels of sparsity, where there is a further steep decline in top-1 accuracy. Sensitivity to different corruptions is remarkably varied, with certain corruptions such as gaussian, shot an impulse noise consistently causing more degradation.

ImageNet-A ImageNet-A is a curated test set of $7,500$ natural adversarial images designed to produce drastically low test accuracy. We find that the sensitivity of sparse models to ImageNet-A mirrors the patterns of degradation to ImageNet-C and sets of PIEs. As sparsity increase, top-1 and top-5 accuracy further erode, suggesting that sparse models are more brittle to adversarial examples.

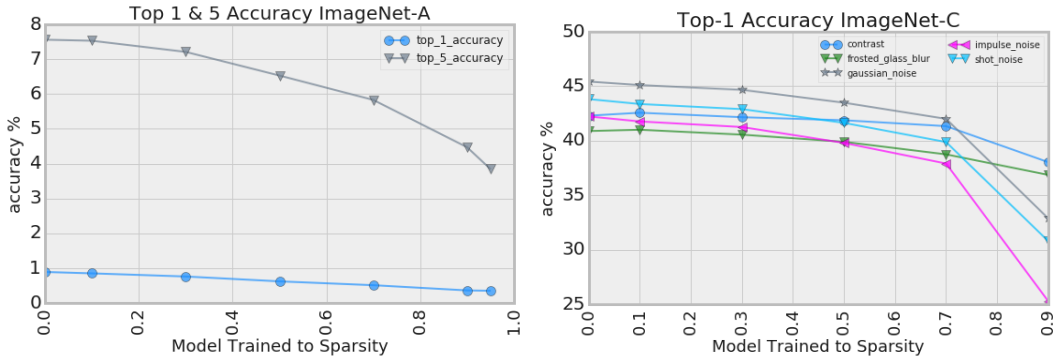

Figure 7: Sparse models are less robust to natural adversarial examples. At high levels of sparsity, models are also more brittle to common image corruptions. We measure the **relative** Top-1 and Top-5 test set ResNet-50 accuracy normalized by average sparse model performance on an uncorrupted ImageNet test set. **Left:** Mean test-set accuracy on ImageNet-A (across 30 models). **Right:** Test-set performance on a subset of ImageNet-C corruptions. An extended list of all corruptions considered is included in the appendix.

4 Related Work

Model compression is diverse and includes research directions such as reducing the precision or bit size per model weight (quantization) (Courbariaux et al., 2014; Hubara et al., 2016; Gupta et al., 2015), efforts to start with a network that is more compact with fewer parameters, layers or computations (architecture design) (Howard et al., 2017; Iandola et al., 2016; Kumar et al., 2017) and student networks with fewer parameters that learn from a larger teacher model (model distillation) (Hinton et al., 2015) and finally pruning by setting a subset of weights or filters to zero (Louizos et al., 2017; Wen et al., 2016; Cun et al., 1990; Hassibi et al., 1993a; Strm, 1997; Hassibi et al., 1993b; Zhu & Gupta, 2017; See et al., 2016; Narang et al., 2017). Articulating the trade-offs of compression has overwhelming centered on change to overall accuracy. Our contribution, while limited in scope to model compression techniques that prune deep neural networks, is the first work to our knowledge to consider the limitations of these aggregate performance measures and demonstrate that the impact of pruning in deep neural networks at a class and exemplar level is non-uniform.

We also consider how pruning impacts robustness to natural adversarial examples and image corruptions. We note that recent work by (Hendrycks & Dietterich, 2019; Hendrycks et al., 2019) considers complimentary variant of this question by benchmark ImageNet-A and ImageNet-B robustness across a limited set of dense non-sparse architectures with different numbers of parameters (for example ResNet-50 vs ResNet-101). While our work is focused on understanding the impact of sparsity on an exemplar and class level, one of our key findings is that PIE is far more challenging to classify for both sparse and non-sparse models. Leveraging this subset of data points for interpretability purposes or to cleanup the dataset fits into a broader and non-overlapping body of literature that aims to classify input data points as prototypes – "most typical" examples of a class – ((Carlini et al., 2018; Stock & Cisse, 2017)) or outside of the training distribution (OOD) (Hendrycks & Gimpel, 2016; Lee et al., 2018a; Liang et al., 2018; Lee et al., 2018a; Masana et al., 2018) and work on calibrating deep neural network predictions (Lakshminarayanan et al., 2017; Guo et al., 2017; Kendall & Gal, 2017).

5 Conclusion

We show that deep neural networks pruned to different levels of sparsity "forget" certain classes and examples more than others. While a subset of classes are systematically impacted, the direction of this impact is surprising and nuanced. Our results show certain classes are relatively impervious to the reduction in model capacity while others bear the brunt of degradation in performance. Pruning identified exemplars are a subset of exemplars where there is a high level of disagreement between sparse and non-sparse models. We show that this subset is universally challenging for models levels at all levels of sparsity to classify and mirrors the sensitivity of sparse models to open source robustness benchmarks ImageNet-C and ImageNet-A.

REFERENCES

T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. ISSN 01621459. URL http://www.jstor.org/stable/2281537.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, June 2008. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1390681.1442792.

Nicholas Carlini, Úlfar Erlingsson, and Nicolas Papernot. Prototypical examples in deep learning: Metrics, characteristics, and utility. 2018.

Rich Caruana. Case-based explanation for artificial neural nets. In Helge Malmgren, Magnus Borga, and Lars Niklasson (eds.), *Artificial Neural Networks in Medicine and Biology*, pp. 303–308, London, 2000. Springer London. ISBN 978-1-4471-0513-8.

B.J. Casey, Jay N. Giedd, and Kathleen M. Thomas. Structural and functional brain development and its relation to cognitive development. *Biological Psychology*, 54(1):241 – 257, 2000. ISSN 0301-0511. doi: https://doi.org/10.1016/S0301-0511(00)00058-2. URL http://www.sciencedirect.com/science/article/pii/S0301051100000582.

Y. Chen, J. Emer, and V. Sze. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 367–379, June 2016. doi: 10.1109/ISCA.2016.40.

M. D. Collins and P. Kohli. Memory Bounded Deep Convolutional Networks. *ArXiv e-prints*, December 2014.

Maxwell D. Collins and Pushmeet Kohli. Memory bounded deep convolutional networks. *CoRR*, abs/1412.1442, 2014. URL http://arxiv.org/abs/1412.1442.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. 2016a. URL https://cs.nyu.edu/~mohri/pub/rej.pdf.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1660–1668. Curran Associates, Inc., 2016b. URL http://papers.nips.cc/paper/6336-boosting-with-abstention.pdf.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online Learning with Abstention. *arXiv e-prints*, art. arXiv:1703.03478, Mar 2017.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Training deep neural networks with low precision multiplications. *arXiv e-prints*, art. arXiv:1412.7024, Dec 2014.

Yann Le Cun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pp. 598–605. Morgan Kaufmann, 1990.

R. B. D'Agostino and M. A. Stephens. *Goodness-of-fit techniques*. 1986.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Andre Esteva, Brett Kuprel, Roberto Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 01 2017. doi: 10.1038/nature21056.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019. URL http://arxiv.org/abs/1902.09574.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *arXiv e-prints*, art. arXiv:1706.04599, Jun 2017.

Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. *CoRR*, abs/1608.04493, 2016. URL http://arxiv.org/abs/1608.04493.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. *CoRR*, abs/1502.02551, 2015. URL http://arxiv.org/abs/1502.02551.

Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient Data Representation by Selecting Prototypes with Importance Weights. *arXiv e-prints*, art. arXiv:1707.01212, Jul 2017.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both Weights and Connections for Efficient Neural Network. In *NIPS*, pp. 1135–1143, 2015.

B. Hassibi, D. G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pp. 293–299 vol.1, March 1993a. doi: 10.1109/ICNN.1993.298572.

Babak Hassibi, David G. Stork, and Stork Crc. Ricoh. Com. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*, pp. 164–171. Morgan Kaufmann, 1993b.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, December 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv e-prints*, art. arXiv:1610.02136, Oct 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. *arXiv e-prints*, art. arXiv:1907.07174, Jul 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, art. arXiv:1503.02531, Mar 2015.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *CoRR*, abs/1806.10758, 2018. URL http://arxiv.org/abs/1806.10758.

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv e-prints*, April 2017.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *CoRR*, abs/1609.07061, 2016. URL http://arxiv.org/abs/1609.07061.

C Huber-Carol, N Balakrishnan, M S. Nikulin, and Mounir Mesbah. *Goodness-of-Fit Tests and Model Validity*. 01 2002. ISBN 978-1-4612-6613-6. doi: 10.1007/978-1-4612-0103-8.

F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5MB model size. *ArXiv e-prints*, February 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL http://arxiv.org/abs/1502.03167.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient Neural Audio Synthesis. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 2415–2424, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5574–5584. Curran Associates, Inc., 2017.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2280–2288. Curran Associates, Inc., 2016.

B. Kolb and I.Q. Whishaw. *Fundamentals of Human Neuropsychology*. A series of books in psychology. Worth Publishers, 2009. ISBN 9780716795865. URL https://books.google.com/books?id=z0DThNQqdL4C.

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

Ashish Kumar, Saurabh Goyal, and Manik Varma. Resource-efficient machine learning in 2 KB RAM for the internet of things. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1935–1944, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/kumar17a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL http://dl.acm.org/citation.cfm?id=3295222.3295387.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=ryiAv2xAZ.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: single-shot network pruning based on connection sensitivity. *CoRR*, abs/1810.02340, 2018b. URL http://arxiv.org/abs/1810.02340.

Christian Leibig, Vaneeda Allken, Murat Seckin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-17876-z.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang. Learning Efficient Convolutional Networks through Network Slimming. *ArXiv e-prints*, August 2017.

C. Louizos, M. Welling, and D. P. Kingma. Learning Sparse Neural Networks through L_0 Regularization. *ArXiv e-prints*, December 2017.

Marc Masana, Idoia Ruiz, Joan Serrat, Joost van de Weijer, and Antonio M. Lopez. Metric Learning for Novelty and Anomaly Detection. *arXiv e-prints*, art. arXiv:1808.05492, Aug 2018.

Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable Training of Artificial Neural Networks with Adaptive Sparse Connectivity Inspired by Network Science. *Nature Communications*, 2018.

Sharan Narang, Erich Elsen, Gregory Diamos, and Shubho Sengupta. Exploring Sparsity in Recurrent Neural Networks. *arXiv e-prints*, art. arXiv:1704.05119, Apr 2017.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv e-prints*, art. arXiv:1412.1897, Dec 2014.

Steven J. Nowlan and Geoffrey E. Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4):473–493, 1992. doi: 10.1162/neco.1992.4.4.473. URL https://doi.org/10.1162/neco.1992.4.4.473.

Pasko Rakic, Jean-Pierre Bourgeois, and Patricia S. Goldman-Rakic. Synaptic development of the cerebral cortex: implications for learning, memory, and mental illness. In J. Van Pelt, M.A. Corner, H.B.M. Uylings, and F.H. Lopes Da Silva (eds.), *The Self-Organizing Brain: From Growth Cones to Functional Networks*, volume 102 of *Progress in Brain Research*, pp. 227 – 243. Elsevier, 1994. doi: https://doi.org/10.1016/S0079-6123(08)60543-9. URL http://www.sciencedirect.com/science/article/pii/S0079612308605439.

B. Reagen, P. Whatmough, R. Adolf, S. Rama, H. Lee, S. K. Lee, J. M. Hernndez-Lobato, G. Wei, and D. Brooks. Minerva: Enabling low-power, highly-accurate deep neural network accelerators. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 267–278, June 2016. doi: 10.1109/ISCA.2016.32.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. Compression of Neural Machine Translation Models via Pruning. *arXiv e-prints*, art. arXiv:1606.09274, Jun 2016.

Elizabeth R. Sowell, Paul M. Thompson, Christiana M. Leonard, Suzanne E. Welcome, Eric Kan, and Arthur W. Toga. Longitudinal mapping of cortical thickness and brain growth in normal children. *Journal of Neuroscience*, 24(38):8223–8231, 2004. doi: 10.1523/JNEUROSCI.1798-04.2004. URL https://www.jneurosci.org/content/24/38/8223.

Pierre Stock and Moustapha Cisse. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. *arXiv e-prints*, art. arXiv:1711.11443, Nov 2017.

Nikko Strm. Sparse connection and pruning in large dynamic artificial neural networks, 1997.

Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and Fisher pruning. *CoRR*, abs/1801.05787, 2018. URL http://arxiv.org/abs/1801.05787.

Karen Ullrich, Edward Meeds, and Max Welling. Soft Weight-Sharing for Neural Network Compression. *CoRR*, abs/1702.04008, 2017.

Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving Neural Speech Synthesis Through Linear Prediction. *CoRR*, abs/1810.11846, 2018. URL http://arxiv.org/abs/1810.11846.

Andreas S. Weigend, David E. Rumelhart, and Bernardo A. Huberman. Generalization by weight-elimination with application to forecasting. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky (eds.), *Advances in Neural Information Processing Systems 3*, pp. 875–882. Morgan-Kaufmann, 1991.

B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947. ISSN 0006-3444. doi: 10.2307/2332510. URL https://doi.org/10.2307/2332510.

W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning Structured Sparsity in Deep Neural Networks. *ArXiv e-prints*, August 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. URL http://arxiv.org/abs/1605.07146.

M. Zhu and S. Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *ArXiv e-prints*, October 2017.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *CoRR*, abs/1710.01878, 2017. URL http://arxiv.org/abs/1710.01878.

6 APPENDIX

6.1 MAGNITUDE PRUNING

There are various pruning methodologies that use the absolute value of the weight as way to rank importance and remove from the network weights that are below a user specified threshold. This is often over the course of training; training is punctuated at certain pruning steps and a fraction of weights are set to zero. Many different magnitude pruning methods have been proposed (Collins & Kohli, 2014; Guo et al., 2016; Zhu & Gupta, 2017) that largely differ in whether the weights are removed permanently or can "recover" by still receiving subsequent gradient updates. This would allow certain weights to become non-zero again if pruned incorrectly. While magnitude pruning is often used as a criteria to remove individual weights, it can be adapted to remove entire neurons or filters by extending the ranking criteria to a set of weights and setting the threshold appropriately. Recent work on evolutionary strategies has also leveraged an interative version of magnitude pruning (Mocanu et al., 2018).

In this work, we use the magnitude pruning methodology proposed by Zhu & Gupta (2017). Pruning is introduced over the course of training and removed weights continue to receive gradient updates after being pruned. For ImageNet, each model trains for a total of 32,000 steps. We prune every 500 steps between 1,000 and 9,000 steps. For CIFAR-10, we train the model for 80,000 steps. We prune every 2,000 steps between 1,000 and 20,000 steps. These hyperparameter choices were based upon a limited grid search which suggested that these particular settings minimized degradation to test-set accuracy across all sparsity levels. At the end of training, the final pruned mask is fixed and during inference only the remaining weights contribute to the model prediction.

6.2 ADDITIONAL CLASS-LEVEL RESULTS

Tables 2 and 3 provide top-line metrics for ImageNet and CIFAR-10, respectively.

The class-level analysis summary for CIFAR-10 is in Table 6.2. Relative to ImageNet, the percentage of classes significantly impacted at 90% pruning is small: 20% for CIFAR-10 versus 58% for ImageNet. As discussed in the main body, we suspect this is due to CIFAR-10 being a simpler task and the network we started from having much more capacity than necessary for the task.

Fraction Pruned	Top 1	Top 5	# Signif classes	# PIEs
0	76.68	93.25	-	-
0.10	76.66	93.25	51	1,694
0.30	76.46	93.17	69	1,819
0.50	75.87	92.86	145	2,193
0.70	75.02	92.43	317	3,073
0.90	72.60	91.10	582	5,136

Table 2: ImageNet top-1 and top-5 accuracy at all levels of sparsity, averaged over all runs. The fourth column is the number of classes significantly impacted by pruning.

Fraction Pruned	Top 1	# Signif classes	# PIEs
0	94.53	-	-
0.1	94.51	1	97
0.3	94.47	1	114
0.5	94.39	1	144
0.7	94.30	0	137
0.9	94.14	2	216

Table 3: CIFAR-10 top-1 accuracy at all levels of sparsity, averaged over runs. Top-5 accuracy for CIFAR-10 was 99.8% for all levels of sparsity. The fourth column is the number of classes significantly impacted by pruning.

Sparsity (t)	Model accuracy diff.	Significant		Largest increase		Largest decrease	
		# incr.	# decr.	class	norm. diff.	class	norm. diff.
0.1	-0.0002	0	1	-	-	automobile	-0.0016
0.3	-0.0006	1	0	frog	0.0022	-	-
0.5	-0.0014	1	0	truck	0.0022	-	-
0.7	-0.0023	0	0	-	-	-	-
0.9	-0.0039	2	0	truck	0.0030	-	-

Table 4: Summary of class-level results for CIFAR-10. Only classes passing the significance test are included. The model accuracy difference column reports mean $\beta_t^{\mathcal{M}} - \beta_0^{\mathcal{M}}$ as the percentage point difference between the pruned and baseline model accuracies; a negative value means the pruned model's average accuracy is lower than the baseline model's. The normalized difference is calculated using Equation 1.

Figure 3 in the main body of the paper visualizes the relative increases and decreases in performance across classes at 70% and 90% pruning on ImageNet. Those figures were shrunk for space; Figure 8 shows the full chart for 70% sparsity for clarity.

6.3 Additional PIE Results

In the body of the paper we show the performance of the unpruned model on PIE images found at varying levels of sparsity and showed that performance is worst for PIE_{10} images which we suspect are the most difficult and performance is still poor but better for PIE_t for larger values of t. In Figure 10 we plot the performance of the pruned models on PIEs identified at different levels of sparsity and show that the behavior of the pruned models track the behavior of the unpruned model in this regard.

6.4 Additional Corruption and Adversarial Results

Sparse models are less robust to natural adversarial examples. At high levels of sparsity, models are also more brittle to common image corruptions. We include the raw ImageNet-C results in Figure 5.

6.5 Human Study

A balanced sampled PIE and non-PIE were selected at random and shuffled. The classification as PIE or non-PIE was not known or available in the sample. Questions codified for every image considered:

Does label 1 accurately label an object in the image? (0/1)

Does this image depict a single object? (0/1)

Would you consider labels 1,2 and 3 to be semantically very close to each other? (does this image require fine grained classification) (0/1)

Do you consider the object in the image to be a typical exemplar for the class indicated by label 1? (0/1)

Is the image quality corrupted (some common image corruptions – overlaid text, brightness, contrast, filter, defocus blur, fog, jpeg compression, pixelate, shot noise, zoom blur, black and white vs. rbg)? (0/1)

Is the object in the image an abstract representation of the class indicated by label 1? [[an abstract representation is an object in an abstract form, such as a painting, drawing or rendering using a different material.]] (0/1)

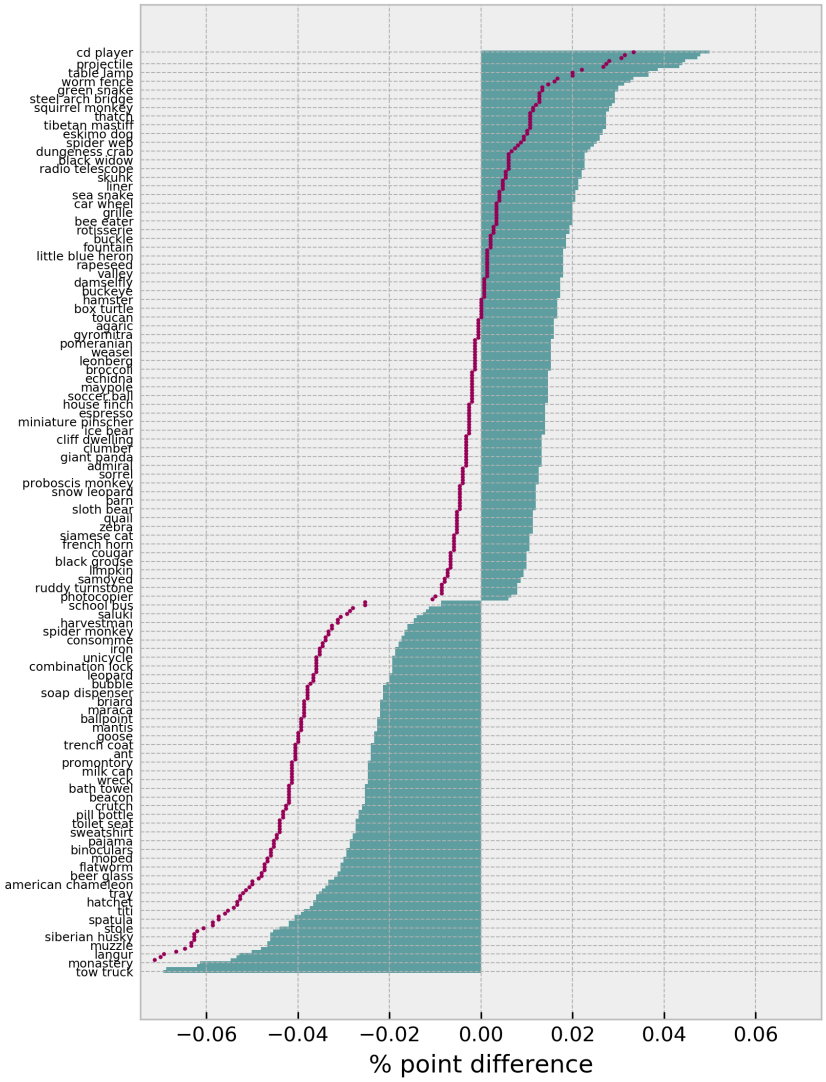

Figure 8: Expanded version of the 70% sparsity chart from Figure 3. Normalized recall difference (bars) and absolute recall difference (points) per class. Every third class label is shown.

Figure 9: Images surfaced by PIE evidence common corruptions such as motion blur, defocus or post-processing with overlaid text. Many PIE images depict objects in an abstract form, such as a painting, drawing or rendering using a different material. PIEs displayed were identified by comparing the modal label of a set of 90% sparse and non-sparse ResNet-50 models.

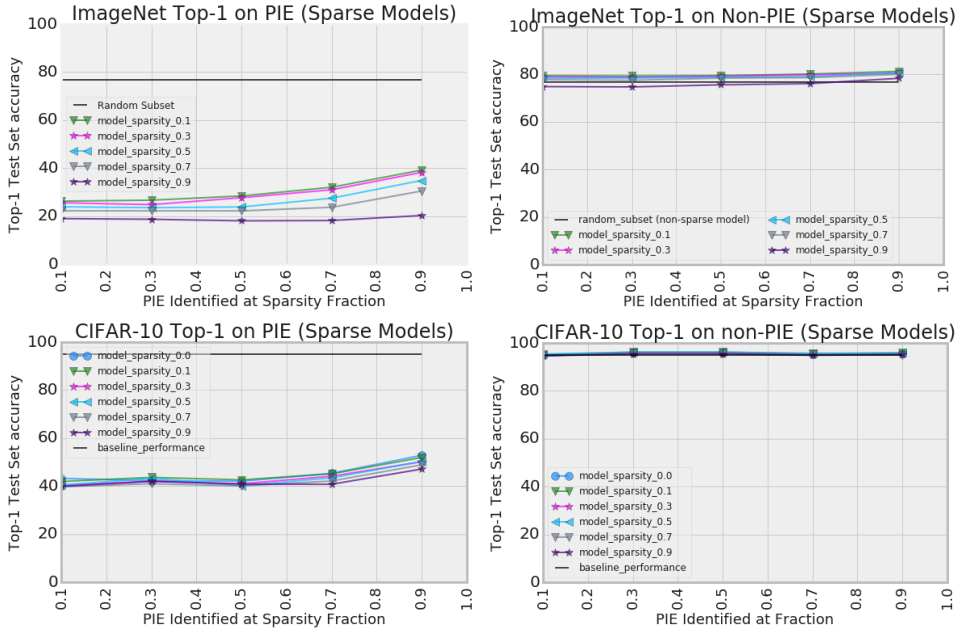

Figure 10: Excluding pruning identified exemplars (PIE) improves test-set top-1 accuracy for both ImageNet and CIFAR-10. The sensitivity to PIE images is amplified at higher levels of sparsity.

ImageNet Robustness to ImageNet-C Corruptions (By Level of Sparsity)					
Corruption Type	**Pruning Fraction**	**Top-1**	**Top-5**	**Top-1 Relative**	**Top-5 Relative**
brightness	0.0	0.69	0.89	90.90	95.53
brightness	0.3	0.69	0.89	90.51	95.35
brightness	0.7	0.67	0.88	90.01	95.08
brightness	0.9	0.64	0.86	88.39	94.02
contrast	0.0	0.42	0.62	55.32	66.35
contrast	0.3	0.42	0.62	55.15	66.27
contrast	0.7	0.41	0.62	55.13	66.64
contrast	0.9	0.38	0.58	52.44	64.16
defocus blur	0.0	0.50	0.72	65.10	77.79
defocus blur	0.3	0.49	0.72	64.65	77.52
defocus blur	0.7	0.47	0.71	63.33	76.50
defocus blur	0.9	0.45	0.68	61.60	74.95
elastic	0.0	0.57	0.77	74.68	82.36
elastic	0.3	0.57	0.77	74.33	82.18
elastic	0.7	0.55	0.75	73.46	81.48
elastic	0.9	0.53	0.74	72.80	80.84
fog	0.0	0.56	0.79	73.52	85.08
fog	0.3	0.56	0.79	73.31	85.04
fog	0.7	0.54	0.78	72.62	84.68
fog	0.9	0.50	0.75	69.42	82.46
gaussian noise	0.0	0.45	0.66	59.42	70.50
gaussian noise	0.3	0.45	0.65	58.44	69.63
gaussian noise	0.7	0.42	0.62	56.03	67.52
gaussian noise	0.9	0.33	0.51	45.32	56.53
impulse noise	0.0	0.42	0.63	55.24	67.81
impulse noise	0.3	0.41	0.62	53.97	66.74
impulse noise	0.7	0.38	0.59	50.55	63.65
impulse noise	0.9	0.25	0.43	34.86	47.36
jpeg compression	0.0	0.66	0.86	86.00	92.61
jpeg compression	0.3	0.65	0.86	85.35	92.24
jpeg compression	0.7	0.63	0.85	84.64	91.78
jpeg compression	0.9	0.61	0.83	83.50	90.89
pixelate	0.0	0.57	0.78	75.00	83.80
pixelate	0.3	0.57	0.78	74.47	83.46
pixelate	0.7	0.55	0.76	73.25	82.43
pixelate	0.9	0.51	0.73	70.73	80.13
shot noise	0.0	0.44	0.64	57.32	68.78
shot noise	0.3	0.43	0.63	56.12	67.73
shot noise	0.7	0.40	0.60	53.19	64.97
shot noise	0.9	0.31	0.49	42.46	53.65

Table 5: Sparse models are more sensitive to image corruptions that are meaningless to a human. We measure the average Top-1 and Top-5 test set accuracy of models trained to varying levels of sparsity on the ImageNet-C test-set (the models were trained on uncorrupted ImageNet). For each corruption we consider and the relative measures see appendix (Table. 5)) we compute the average accuracy of 50 trained models across all 5 levels of corruption severity.