

---

# SELECTIVE BRAIN DAMAGE: MEASURING THE DISPARATE IMPACT OF MODEL PRUNING

**Sara Hooker** \*  
Google Brain

**Aaron Courville**  
MILA

**Yann Dauphin**  
Google Brain

**Andrea Frome**  
Google Brain

## ABSTRACT

Neural network pruning techniques have demonstrated it is possible to remove the majority of weights in a network with surprisingly little degradation to 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* pruned and non-pruned models. These hard-to-generalize-to images tend to be mislabelled, of lower image quality, depict multiple objects or require fine-grained classification. These findings shed light on previously unknown trade-offs, and suggest that a high degree of caution should be exercised before pruning is used in sensitive domains.

## 1 INTRODUCTION

Between infancy and adulthood, the number of synapses in our brain first multiply and then fall. Synaptic pruning improves efficiency by removing redundant neurons and strengthening synaptic connections that are most useful for the environment [58]. Despite losing 50% of all synapses between age two and ten, the brain continues to function [42, 62]. The phrase "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 [5].

In this work, we ask what is *lost* when we prune a deep neural network. In 1990, a popular paper was published titled "Optimal Brain Damage" [13]. The paper was among the first [27, 57, 68] to propose that deep neural networks could be pruned of "excess capacity" in a similar fashion to synaptic pruning. At face value, pruning appears to promise you can (almost) have it all. Deep neural networks are remarkably tolerant of high levels of pruning with an almost negligible loss to top-1 accuracy [24, 66, 50, 51, 7, 47]. For example, Gale et al. [18] show that removing 90% of all weights in a ResNet-50 network [28] trained on ImageNet [16] results in less than a 3% absolute decrease in top-1 test set accuracy. These more compact networks are frequently favored in resource constrained settings; pruned models require less memory, energy consumption and have lower inference latency [59, 6, 65, 39, 67].

The ability to prune networks with seemingly so little degradation to generalization performance is puzzling. The cost to top-1 accuracy appears minimal if it is spread uniformly across all classes, but what if the cost is concentrated in only a few classes? *Are certain types of examples or classes disproportionately impacted by pruning?* An understanding of these trade-offs is critical for sensitive tasks such as hiring [15, 25], health care diagnostics [71, 19], self-driving cars [56], where the introduction of pruning may be at odds with fairness objectives to treat protected attributes uniformly and/or the need to guarantee a certain level of recall for certain classes. Pruning is already commonly used in these domains, often driven by the resource constraints of deploying models to mobile phone or embedded devices [17, 60].

In this work we propose a formal methodology to evaluate the impact of pruning on a class and exemplar level (Sections 3.2 and 2.3). The measures we propose identify classes and images where

---

\*Correspondence can be sent to shooker@google.com, aaron.courville@umontreal.ca, ynd@google.com and onepinkfairyarmadillo@gmail.com.

Code associated with this paper is available at: <https://bit.ly/2C8GrID>.



Figure 1: Visualization of a subset of pruning identified images (PIE) for the ImageNet dataset. Below each image: the ground truth label, the most frequent non-pruned prediction (across 30 models) and most frequent pruned prediction (across 30 models trained to 50% sparsity). In Section 2.3, we conduct a small scale human study and find that PIEs heavily overindex on images with an *incorrect ground truth label*, involve *fine grained* classification tasks or depict *multiple objects*. Removing PIEs from the test-set greatly improves test-set accuracy for both pruned and non-pruned models.

there is a high level of disagreement or difference in generalization performance between pruned and non-pruned models. Our results are surprising and suggest that a reliance on top-line metrics such as top-1 or top-5 test-set accuracy hides critical details in the ways that pruning impacts model generalization. The primary findings of our work can be summarized as follows:

1. *Pruning in deep neural networks is better described as “selective brain damage.” 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 pruned and non-pruned models to classify.*
3. *We conduct a small scale human study and find that PIEs tend to overindex on images with an incorrect ground truth label, images that involve fine grained classification tasks or depict multiple objects.*
4. *Pruning significantly reduces robustness to image corruptions and adversarial attacks.*

For (1) and (2), we establish consistent findings for different standard architectures on CIFAR-10 [43] and ImageNet. Toward finding (4), we measure changes to model sensitivity to both common

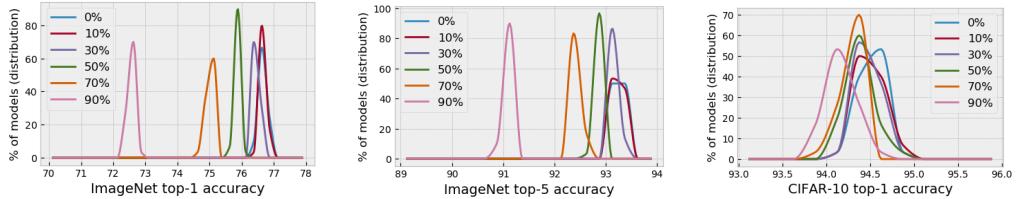


Figure 2: Distributions of top-1 and top-5 model accuracy for populations of independently trained pruned and non-pruned models on ImageNet and CIFAR-10. The distributions for CIFAR-10 top-5 accuracy (not shown) are tightly clustered and overlapping in [99.65, 100.0]. The distributions are fairly tight with one outlier for the ImageNet baseline model.

image corruptions and natural adversarial examples using two open source robustness benchmarks: ImageNet-C [29] and ImageNet-A [31].

The over-indexing of poorly structured data (multi-object or incorrectly labelled data) in PIE hints that the explosion of growth in number of parameters in deep neural networks may be solving a problem that is better addressed in the data cleaning pipeline. More broadly, 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 pruned and non-pruned models. PIE could be used to surface atypical examples for further human inspection [48], choose not to classify certain examples when the model is uncertain [2, 10, 9, 11], or to aid interpretability as a case based reasoning tool to explain model behavior [41, 23, 4, 33].

## 2 METHODOLOGY AND EXPERIMENT FRAMEWORK

### 2.1 PRELIMINARIES

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  $C$  categories or classes, such that  $y_i = [1, \dots, C]$ .

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 yet yielded competitive test-set performance. Instead, current research centers on training strategies where models are initialized with “excess capacity” which is then subsequently removed through pruning. A pruning method  $\mathcal{P}$  identifies the subset of weights to remove (i.e. set to zero). A pruned model function,  $\hat{F}_t$ , is one where a fraction  $t \in [0.0, 1.0]$  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-pruned model function,  $\hat{F}_0$ , is one where all weights are trainable ( $t = 0$ ). At times, we interchangeably refer to  $\hat{F}_t$  and  $\hat{F}_0$  as sparse and non-sparse model functions (where the level of sparsity is indicated by  $t$ ).

### 2.2 CLASS LEVEL MEASURE OF IMPACT

Comparing only top-1 model accuracy between a baseline and a pruned model amounts to assuming that class accuracy is expected to maintain its relative relationship 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  $\beta_t^c$  for class  $c \in \mathcal{C}$  and sparsity  $t \in [0, 1]$ . We compute overall model accuracy from the set of class metrics:

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

---



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

Figure 3: 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-pruned wide ResNet models and 30 models trained to 30% sparsity. Below each image is three labels: 1) true label, 2) the modal (most frequent) prediction from the set of non-pruned models, 3) the modal prediction from the set of 30% pruned models.

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 pruned and non-pruned 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:

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

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

Evaluating whether the difference between a sample of mean-shifted class accuracy from pruned and non-pruned 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 [14, 1, 36]. Neural network training is most often done in an independent 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  $S_t^c = \{\beta_{t,k}^c - \beta_{t,k}^{\mathcal{M}}\}_{k=1}^K$ . In this work, we use a two-sample, two-tailed, independent Welch’s t-test [69] 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  $\leq 0.05$ ), 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 shows 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  $t$ -pruned and baseline class accuracies after normalizing for their respective average model accuracies:

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

### 2.3 IMAGE LEVEL MEASURE OF IMPACT

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 pruned models is equal. However, recent work has highlighted that deep neural networks produce output probabilities that are uncalibrated [20, 40, 45] 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 [30, 55].

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

Sparsity ( $t$ )	Model acc diff.	Significant		Largest increase			Largest decrease		
		# incr.	# decr.	class	norm	abs	class	norm	abs
0.1	-0.02	29	22	toaster	2.55	2.53	chameleon	-2.64	-2.66
0.3	-0.2	35	34	bathtub	3.55	3.33	cleaver	-4.51	-4.73
0.5	-0.8	91	54	petri dish	3.41	2.6	frying pan	-4.66	-5.46
0.7	-1.7	189	128	cd player	4.99	3.33	tow truck	-6.94	-8.6
0.9	-4.1	337	245	cd player	8.78	4.67	muzzle	-12.82	-16.93

Table 1: Summary of class-level results for ImageNet. Only classes passing the significance test are included. The model accuracy difference column reports mean  $\beta_t^M - \beta_0^M$  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 (norm) is calculated using Equation 1. The absolute difference (abs) is the difference between average per-class accuracy at  $t_0$  (no pruning) compared to  $t$  (indicated by sparsity column).

disagreement between the predictions of pruned and non-pruned networks on a given image. Let  $y_{i,k,t}^*$  be the prediction of the  $k$ th  $t$ -pruned model of its population for image  $i$  where  $t = 0$  denotes an non-pruned model, and let  $Y_{i,t}^* = \{y_{i,k,t}^*\}_{k=1}^K$  be the set of predictions for the  $t$ -pruned model population on exemplar  $i$ . For set  $Y_{i,t}^*$  we find the *modal label*, i.e. the class predicted most frequently by the  $t$ -pruned model population for exemplar  $i$ , which we denote  $y_{i,t}^M$ . Exemplar  $i$  is classified as a pruning identified exemplar  $PIE_t$  if and only if the modal label is different between the set of  $t$ -pruned models and the non-pruned models:

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

We note that there is no constraint that the non-pruned 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.

### 3 EXPERIMENT SETUP AND RESULTS

#### 3.1 EXPERIMENT SETUP

We consider two classification tasks and models; a wide ResNet model [72] trained on CIFAR-10 and a ResNet-50 model [28] trained on ImageNet. Both networks are trained with batch normalization [38]. 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 for 80,000 steps on CIFAR-10 with a batch size of 128. For ImageNet, the baseline non-pruned 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 2 shows the distributions of model accuracy across model populations for the non-pruned and pruned models for ImageNet and CIFAR-10.

Across all experiments, we use magnitude pruning as proposed by Zhu & Gupta [74] 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 [18]. 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 [50, 51, 7, 70, 68, 57]. Although the ability to precisely vary sparsity is required for this experimental framework, we note that our methodology can be extended to other methods. In order to encourage replication of our results using additional pruning methods, we have open sourced our code for all experiments.

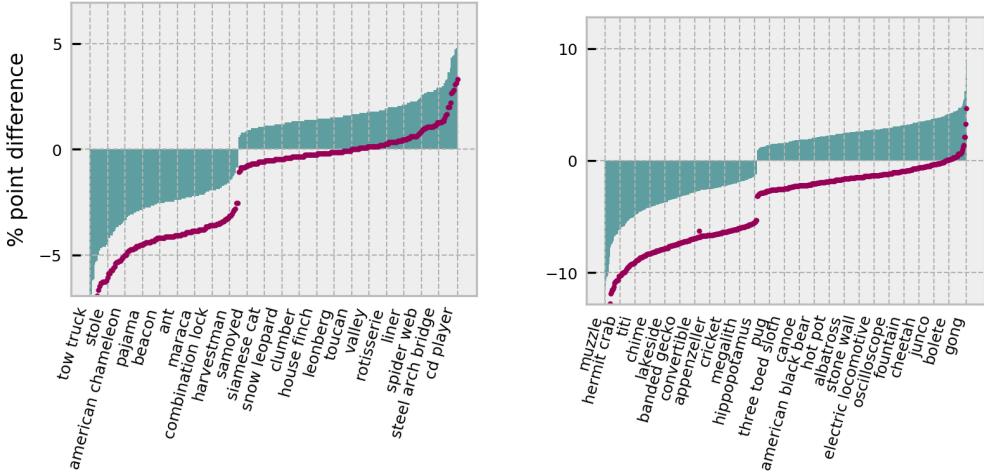


Figure 4: Normalized recall difference (green bars) and absolute recall difference (plum points) per class for 70% sparsity (left) and 90% sparsity (right). The class labels are sampled for readability; there are 317 significant classes for 70% sparsity and 582 significant classes for 90%. Note the difference in scale on the y-axis. The normalized difference (norm) is described in section 2.2. The absolute difference (abs) is the difference between average per-class accuracy at  $t_0$  (no pruning) compared to  $t$  (indicated by sparsity column)

### 3.2 IMPACT OF SPARSITY ON CLASS LEVEL PERFORMANCE

We now return to our initial question about class level impact – *Is relative performance unaltered by pruning or are some classes impacted more than others?*. We compute the the *normalized recall class difference* (introduced in 2.2) for each class in image. We find that 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 a 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. The model appears to cannibalize performance on a small subset of classes in order to preserve performance (and even improve relative performance on a small number of classes). Figure 4 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 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.

### 3.3 IMPACT OF SPARSITY ON INDIVIDUAL EXEMPLARS

We now turn to the impact of pruning at an exemplar level. We use the PIE methodology introduced in 2.3 and identify a subset of PIE images at every level of sparsity (for both CIFAR-10 and ImageNet).

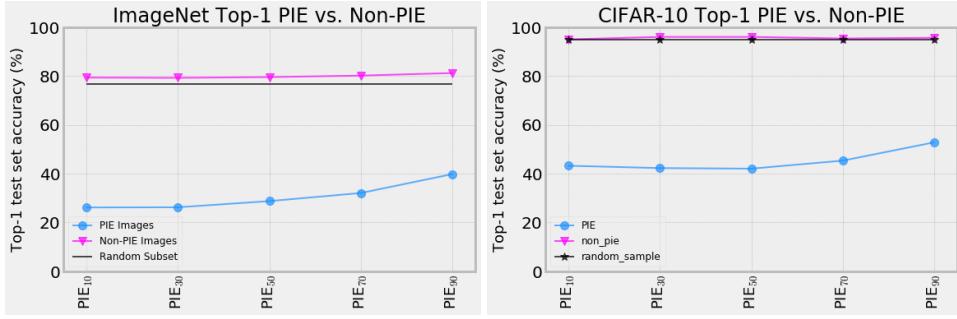


Figure 5: Excluding pruning identified exemplars (PIE) improves test-set top-1 accuracy for both ImageNet and CIFAR-10. This holds for PIE images identified at all levels of sparsity considered. Inference on PIE images alone substantially degrades generalization performance of pruned models. **Left:** Average top-1 test-set accuracy across 30 non-pruned 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-pruned wide ResNets trained on CIFAR-10.

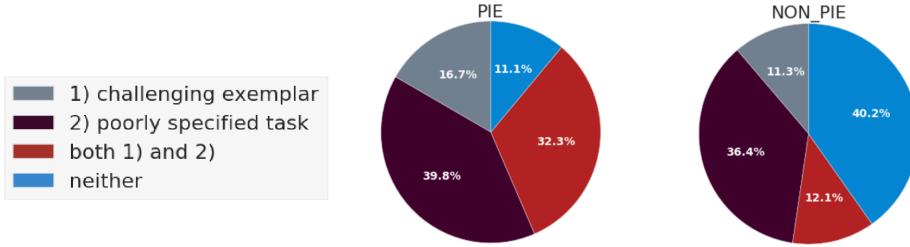


Figure 6: Limited human study of the relative distribution of PIE and non-PIE properties. **Challenging exemplars:** images positively codified as showing *common image corruptions* such as blur or overlaid text, or images where the object is in the form of an *abstract representation* or where the exemplar 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*.

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. Why does pruning introduce a high level of prediction disagreement for some images but not others? We turn our attention now to gaining an understanding of *what makes PIEs different from non-PIEs?*

**PIEs are more difficult for both pruned and non-pruned models to classify.** In Fig. 5, we compare the test-set performance of a fully parameterized non-pruned 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* pruned and non-pruned models relative to a random sample. Inference restricted to only PIE images significantly degrades top-1 accuracy. 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 5, the lowest test-set accuracy for both pruned and non-pruned 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 the introduction of sparsity first erodes performance on the images that the model finds the most challenging.

**Why are PIEs harder to classify?** A qualitative inspection of PIEs (Figure 7) 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 (involving 85 volunteers who work at an industry research lab) to label a random sample

multi-object image	image corruptions	
		
true : tennis ball non-sparse : tennis ball sparse : racket	projectile missile projectile	lynx labrador retriever nematode
ground truth label inadequate	fine-grained classification	abstract representation
		
true : restaurant non-sparse : ice cream sparse : cauliflower	cup cup coffee mug	cowboy hat cowboy hat plastic bag
		butcher store grocery store crayfish

Figure 7: PIE images often exhibit shared characteristics. We conduct a limited human study to measure the relative representation of these properties and visualize an example prototypical of each grouping here. The PIE images visualized here are computed by comparing a set of 30 non-pruned ResNet-50 models to 30 models trained to 90% sparsity on ImageNet.

of 1230 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

- **ground truth label incorrect or inadequate** – images where there is not sufficient information for a human to arrive at the correct ground truth label. For example in Fig. 7, 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. [12% of non-PIEs, 22% of PIEs]
- **multiple-object 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`). [40% of non-PIE, 62% of PIEs]

## 2. Challenging Exemplars

- **fine grained classification** – involves classifying an object that is semantically close to various other class categories present the data set (e.g., `rock crab` and `fiddler crab`, `bassinet` and `cradle`, `cuirass` and `breastplate`). [7% of non-PIEs, 41% of PIEs]
- **image 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 RGB color images in ImageNet. [13% of non-PIE, 11% of PIE]
- **abstract representations** – the surfaced exemplar depicts a class object in an abstract form such a cartoon, painting, or sculptured incarnation of the object. [4% of non-PIE, 4% of PIE]

We find that the number of image corruptions and abstract representations surfaced by PIE appears similar to the overall representation in the ImageNet dataset. However, we find that PIEs appear to heavily overindex relative to non-PIEs on certain properties, such as having an *incorrect ground truth label*, involving a *fine-grained classification task* or *multiple objects*. This suggests that the task itself is often incorrectly specified. Both ImageNet and CIFAR-10 are single image classification tasks, however 62% of PIEs codified by humans were identified as multi-object images where multiple labels could be considered reasonable (vs. 41% 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.

### 3.4 THE ROLE OF ADDITIONAL CAPACITY

The PIE procedure surfaces exemplars that are harder for both pruned and non-pruned models to classify. Given that PIE surfaces data points where there is the greatest divergence in behavior between pruned and non-pruned models, it is useful to understand the directionality of some of the properties described in the previous section. For example, 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 [29] 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 [29], 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 pruned model on the clean ImageNet dataset). ImageNet-C corruption substantially degrades mean top-1 accuracy of non-pruned models Fig. 8. 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 pruned 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 pruned models are more brittle to adversarial examples.

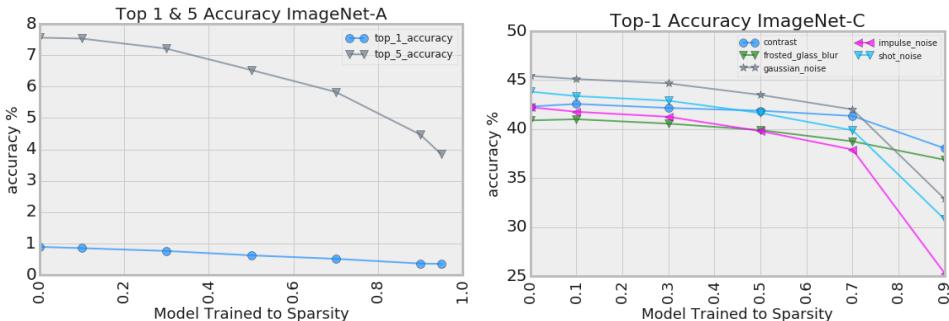


Figure 8: Pruned models are less robust to natural adversarial examples. High levels of sparsity amplifies sensitivity to image corruptions. We measure the **relative** Top-1 and Top-5 test set ResNet-50 accuracy normalized by average sparse model performance on an uncorrupted ImageNet test set. **Left:** 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) [12, 35, 22], efforts to start with a network that is more

---

compact with fewer parameters, layers or computations (architecture design) [34, 37, 44] and student networks with fewer parameters that learn from a larger teacher model (model distillation) [32] and finally pruning by setting a subset of weights or filters to zero [51, 70, 13, 26, 64, 27, 73, 61, 54]. 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 propose a formal methodology to evaluate 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 [29, 31] considers complimentary variant of this question by benchmark ImageNet-A and ImageNet-B robustness across a limited set of dense non-pruned 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 pruned and non-pruned 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 – ([3, 63]) or outside of the training distribution (OOD) [30, 46, 49, 46, 52] and work on calibrating deep neural network predictions [45, 20, 40].

## 5 CONCLUSION

We propose a formal methodology to evaluate the impact of pruning at a class and exemplar level. 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 pruned and non-pruned models. We show that this subset is universally challenging for models at all levels of sparsity to classify. Our results shed light on previously unknown trade-offs, and suggest that caution should be used before using pruned models in sensitive domains where human welfare can be adversely impacted.

## ACKNOWLEDGEMENTS

We thank the generosity of our peers for valuable input on earlier versions of this work. In particular, we would like to acknowledge the input of Jonas Kemp, Simon Kornblith, Julius Adebayo, Hugo Larochelle, Dumitru Erhan, Nicolas Papernot, Catherine Olsson, Cliff Young, Martin Wattenberg, Utku Evci, James Wexler, Trevor Gale, Melissa Fabros, Prajit Ramachandran, Pieter Kindermans, Erich Elsen and Moustapha Cisse. We thank the institutional support and encouragement of Dan Nanan, Rita Ruiz, Sally Jesmonth and Alexander Popper.

## REFERENCES

- [1] 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>.
- [2] 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>.
- [3] Nicholas Carlini, Ulf Erlingsson, and Nicolas Papernot. Prototypical examples in deep learning: Metrics, characteristics, and utility, 2019. URL <https://openreview.net/forum?id=r1xyx3R9tQ>.
- [4] 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.
- [5] 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](https://doi.org/10.1016/S0301-0511(00)00058-2). URL <http://www.sciencedirect.com/science/article/pii/S0301051100000582>.

- [6] 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.
- [7] M. D. Collins and P. Kohli. Memory Bounded Deep Convolutional Networks. *ArXiv e-prints*, December 2014.
- [8] Maxwell D. Collins and Pushmeet Kohli. Memory bounded deep convolutional networks. *CoRR*, abs/1412.1442, 2014. URL <http://arxiv.org/abs/1412.1442>.
- [9] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *ALT*, 2016.
- [10] 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., 2016. URL <http://papers.nips.cc/paper/6336-boosting-with-abstention.pdf>.
- [11] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online Learning with Abstention. *arXiv e-prints*, art. arXiv:1703.03478, Mar 2017.
- [12] 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.
- [13] 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.
- [14] Ralph B D’Agostino and Michael A Stephens (eds.). *Goodness-of-fit Techniques*. Marcel Dekker, Inc., New York, NY, USA, 1986. ISBN 0-824-77487-6.
- [15] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL <https://reut.rs/2p0ZWqe>.
- [16] 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.
- [17] 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.
- [18] 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>.
- [19] Ross Gruetzmacher, Ashish Gupta, and David B. Paradice. 3d deep learning for detecting pulmonary nodules in ct scans. *Journal of the American Medical Informatics Association : JAMIA*, 25 10:1301–1310, 2018.
- [20] 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.
- [21] 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>.
- [22] 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>.
- [23] 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.
- [24] 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.
- [25] Drew Harwell. A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post*, 2019. URL <https://wapo.st/2X3bupO>.
- [26] 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 1993. doi: 10.1109/ICNN.1993.298572.

- 
- [27] 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, 1993.
- [28] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, December 2015.
- [29] 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>.
- [30] 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.
- [31] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. *arXiv e-prints*, art. arXiv:1907.07174, Jul 2019.
- [32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, art. arXiv:1503.02531, Mar 2015.
- [33] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS 2019*, 2019.
- [34] 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.
- [35] 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>.
- [36] C. Huber-Carol, N. Balakrishnan, M. Nikulin, and M. Mesbah. *Goodness-of-Fit Tests and Model Validity*. Goodness-of-fit Tests and Model Validity. Birkhäuser Boston, 2002. ISBN 9780817642099. URL [https://books.google.com/books?id=gUMcv2\\_NrhkC](https://books.google.com/books?id=gUMcv2_NrhkC).
- [37] 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.
- [38] 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>.
- [39] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Nouy, Norman Casagrande, Edward Lockhart, Florian Stimberg, Äaron 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.
- [40] 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.
- [41] 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.
- [42] 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>.
- [43] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.
- [44] 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>.

- 
- [45] 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>.
- [46] 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*, 2018. URL <https://openreview.net/forum?id=ryiAv2xAZ>.
- [47] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: single-shot network pruning based on connection sensitivity. *CoRR*, abs/1810.02340, 2018. URL <http://arxiv.org/abs/1810.02340>.
- [48] 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.
- [49] 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>.
- [50] 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.
- [51] C. Louizos, M. Welling, and D. P. Kingma. Learning Sparse Neural Networks through  $L_0$  Regularization. *ArXiv e-prints*, December 2017.
- [52] 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.
- [53] 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.
- [54] Sharan Narang, Erich Elsen, Gregory Diamos, and Shubho Sengupta. Exploring Sparsity in Recurrent Neural Networks. *arXiv e-prints*, art. arXiv:1704.05119, Apr 2017.
- [55] 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.
- [56] NHTSA. Technical report, U.S. Department of Transportation, National Highway Traffic, Tesla Crash Preliminary Evaluation Report Safety Administration. *PE 16-007*, Jan 2017.
- [57] 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>.
- [58] 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](https://doi.org/10.1016/S0079-6123(08)60543-9). URL <http://www.sciencedirect.com/science/article/pii/S0079612308605439>.
- [59] B. Reagen, P. Whatmough, R. Adolf, S. Rama, H. Lee, S. K. Lee, J. M. Hernandez-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.
- [60] Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine & Biology*, 63(9):095005, may 2018. doi: 10.1088/1361-6560/aabb5b.
- [61] 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.

- 
- [62] 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>.
  - [63] Pierre Stock and Moustapha Cisse. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. *arXiv e-prints*, art. arXiv:1711.11443, Nov 2017.
  - [64] Nikko Strm. Sparse connection and pruning in large dynamic artificial neural networks, 1997.
  - [65] 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>.
  - [66] Karen Ullrich, Edward Meeds, and Max Welling. Soft Weight-Sharing for Neural Network Compression. *CoRR*, abs/1702.04008, 2017.
  - [67] 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>.
  - [68] 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.
  - [69] 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>.
  - [70] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning Structured Sparsity in Deep Neural Networks. *ArXiv e-prints*, August 2016.
  - [71] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, 85:109 – 119, 2019. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2018.07.031>. URL <http://www.sciencedirect.com/science/article/pii/S0031320318302711>.
  - [72] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. URL <http://arxiv.org/abs/1605.07146>.
  - [73] M. Zhu and S. Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *ArXiv e-prints*, October 2017.
  - [74] 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 [8, 21, 74] 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 iterative version of magnitude pruning [53].

In this work, we use the magnitude pruning methodology proposed by [74]. 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 4. 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.

Figure 4 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 9 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 11 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 non-pruned 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.

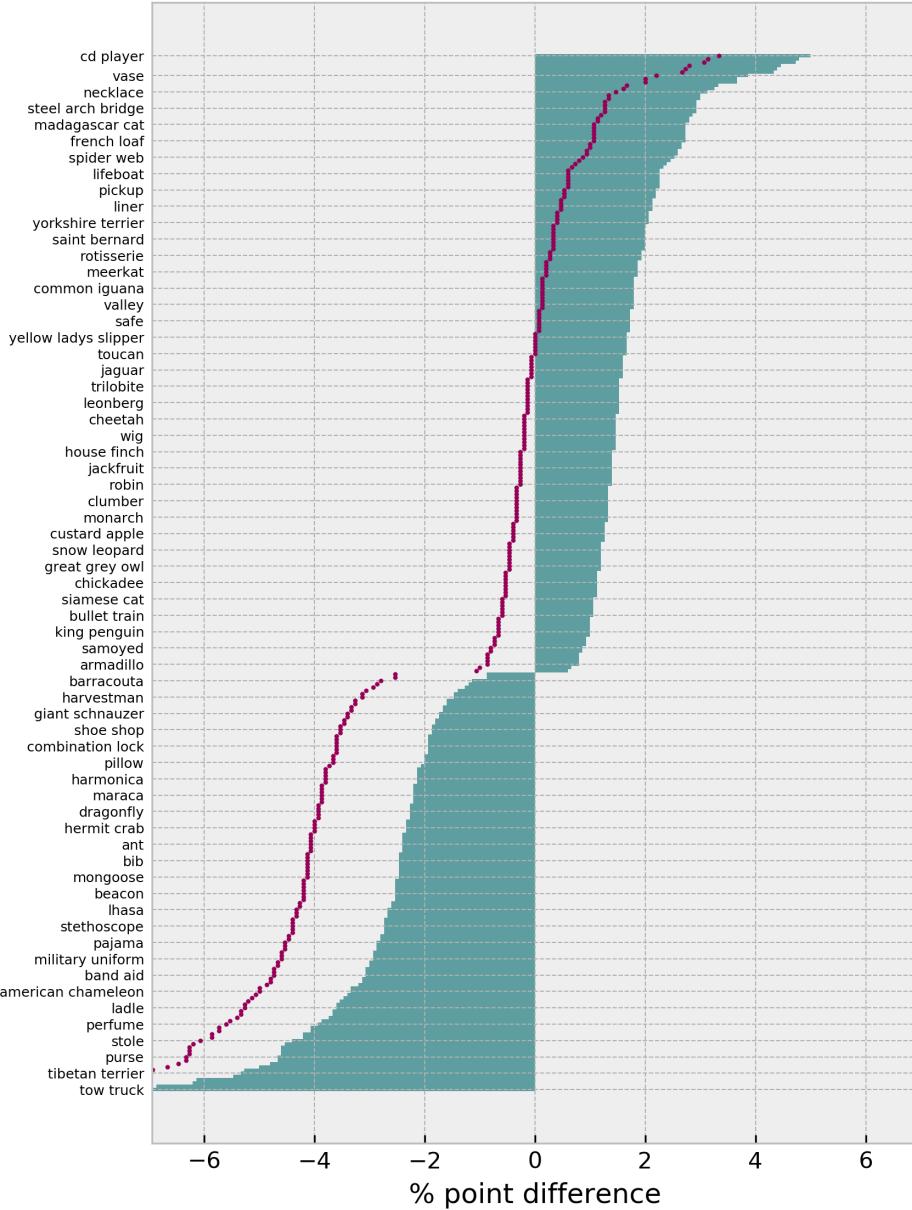


Figure 9: Larger version of the 70% sparsity. Normalized recall difference (green bars) and absolute recall difference (plum points) per class. Every third class label is sampled.

## 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 to the human labels. We include an image of the data collection interface in Figure 12. 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)*

Sparsity ( $t$ )	Model accuracy diff.	Significant		Largest increase			Largest decrease		
		# incr.	# decr.	class	norm	abs	class	norm	abs
0.1	-0.02	0	1	-	-	-	automobile	-0.15	-0.18
0.3	-0.06	1	0	frog	0.22	0.26	-	-	-
0.5	-0.14	1	0	truck	0.22	0.07	-	-	-
0.7	-0.23	0	0	-	-	-	-	-	-
0.9	-0.39	2	0	truck	0.30	0.08	-	-	-

Table 4: Summary of class-level results for CIFAR-10. Only classes passing the significance test are included. The model accuracy difference column reports the percentage difference between the average pruned and baseline model accuracy. The normalized difference (norm) is calculated as described in section 2.2. The absolute difference (abs) is the difference between average per-class accuracy at  $t_0$  (no pruning) compared to models trained to sparsity  $t$ .

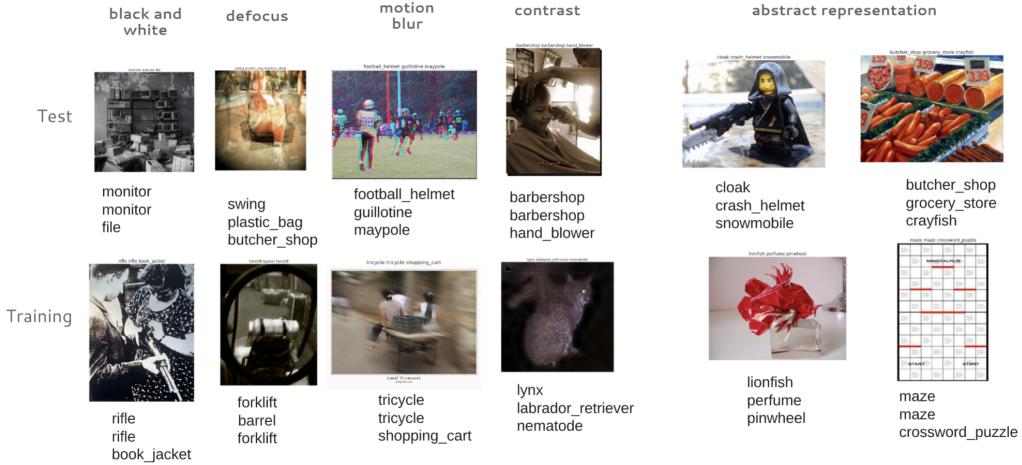


Figure 10: Images surfaced by PIE evidence common corruptions such as motion blur, defocus or post-processing with overlaid text. Many PIE images depict objects in an abstract form, such as a painting, drawing or rendering using a different material. PIEs displayed were identified by comparing the modal label of a set of 90% pruned and non-pruned ResNet-50 models.

*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)*

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: Pruned 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.

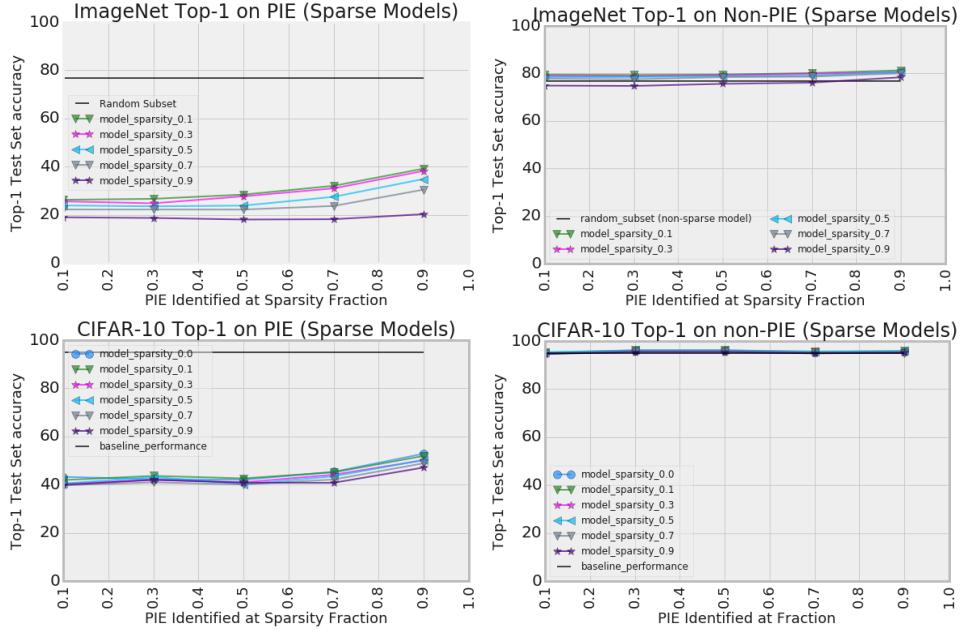


Figure 11: Excluding pruning identified exemplars (PIE) improves test-set top-1 accuracy for both ImageNet and CIFAR-10. The sensitivity to PIE images is amplified at higher levels of sparsity.



All multiple-choice questions must be answered for submit to succeed.

Does "knot" describe an object in this image?  Yes  No  I don't know [Show me pictures for "knot"](#)  
 Does "sax" describe an object in this image?  Yes  No  I don't know [Show me pictures for "sax"](#)  
 Does "cauldron" describe an object in this image?  Yes  No  I don't know [Show me pictures for "cauldron"](#)

Would you consider these labels to be semantically very close to each other, i.e. a **fine-grained classification task**?  Yes  No  I don't know  
 Does this image depict a **single object** class?  Yes  No  I don't know  
 Is the image quality **corrupted**? (see [NOTES](#) below)  Yes  No  I don't know  
 Is the object in the image an **abstract representation** of the "knot" class? (see [NOTES](#))  Yes  No  I don't know

Any comments about labeling this image?  
 (Optional)

Figure 12: Data collection interface for classifying the attributes of PIE and non-PIE images (depicts a single example). Humans in the study were shown a balanced sample of PIE and non-PIE images that were selected at random and shuffled. The classification as PIE or non-PIE was not known or available to the human.