WHAT DOES A PRUNED DEEP NEURAL NETWORK "FORGET"?

Sara Hooker Google Brain
Aaron Courville MILA
Yann Dauphin Google Brain
Andrea Frome

ABSTRACT

The interplay between human learning and synaptic pruning has long been of great interest to cognitive scientists. The term "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. In this work, we ask how varying the number of network weights alters the generalization behavior of a network at a class and exemplar level. Models with radically different numbers of weights have comparable top-line performance metrics, but diverge considerably in behavior on a narrow subset of the input distribution. We explore why varying the number of weights disproportionately and systematically impacts a small subset of classes and examples, which we term *pruning identified exemplars (PIEs)*. Our work provides insights about the role of capacity in deep neural networks and is a valuable human-in-the-loop tool for understanding the trade-offs incurred when deploying compressed models to the real world.

1 INTRODUCTION

Between infancy and adulthood, the number of synapses in our brain first increase and then fall. Synaptic pruning improves efficiency by removing redundant neurons and strengthening synaptic connections that are most 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). 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 (Casey et al., 2000).

In 1990, a popular paper was published titled "Optimal Brain Damage" (Cun et al., 1990). The paper was among the first (Hassibi et al., 1993b; Nowlan & Hinton, 1992; Weigend et al., 1991; Mozer & Smolensky, 1989) to propose that deep neural networks could be pruned of "excess capacity" in a similar fashion to our biological synaptic pruning. In deep neural networks, weights are pruned or removed from the network by setting the value to zero. At face value, deep neural network pruning also appears to promise you can (almost) have it all. State of art pruning techniques remove the majority of weights with an almost negligible loss to top-1 accuracy (Gale et al., 2019). These newly slimmed down networks require less memory, energy consumption and have lower inference latency. All these attributes make pruned models ideal for deploying deep neural networks to resource constrained environments (Lane & Warden, 2018).

However, the ability to prune networks with seemingly so little degradation to generalization performance is puzzling. How can networks with radically different structures and number of parameters have comparable top-level metrics? One possibility may be that test-set accuracy is not a precise enough measure to capture how pruning impacts the generalization properties of the model.

We explore this hypothesis in this work by going beyond test-set accuracy and proposing a formal methodology to evaluate the impact of pruning at a class and exemplar level (Section 1.1). The measures we propose identifies classes and images where there is a high level of disagreement or difference in generalization performance between pruned and non-pruned models.

We find that:

1. *Top-line metrics such as top-1 or top-5 test-set accuracy hides critical details in the ways that pruning impacts model generalization. Pruning in deep neural networks is better*

true: cloak | ladle | espresso | mashed potato
baseline model: gasmask | ladle | espresso | mashed potato
pruned model: breastplate | perfume | red wine | ice cream

true: stretcher | sewing machine | bathtub | crutch
baseline model: folding chair | sewing machine | bathtub | crutch
pruned model: barrow | polaroid camera | cucumber | apron

Figure 1: The exemplars most sensitive to varying capacity are more challenging for both fully parameterized models and humans to classify. This figure visualizes a selection of ImageNet exemplars most impacted by pruning (PIEs). Below each image: the ground truth label, the most frequent non-pruned prediction and most frequent pruned prediction.

described as "selective brain damage" where a small subset of classes and exemplars are disproportionately and systematically impacted by the introduction of pruning.

2. *Why are certain exemplars more sensitive to varying capacity? For everyday object classification datasets like ImageNet (Deng et al., 2009), we conduct a human study and find that PIEs tend to overindex on images in the long-tail of the natural image distribution. Our human study shows that PIEs are more challenging for both human and model.*
3. *For datasets like CelebA (Liu et al., 2015) PIE surfaces known spurious correlations between protected demographic attributes. PIEs over-index on the protected attribute, which suggests that compression methods like pruning may pose unexpected trade-offs with other properties we may care about such as fairness.*

We establish consistent results across multiple datasets—ImageNet, CelebA and Cifar-10 (Krizhevsky, 2012) —and model architectures— a wide ResNet model (Zagoruyko & Komodakis, 2016) and a ResNet-50 model (He et al., 2015).

1.1 Methodology And Experiment Setup

We start by asking a simple question: is the impact of pruning uniform or are certain classes disproportionately impacted? If the impact of pruning was uniform across all classes, we would expect the model accuracy on each class 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 impacted the class's accuracy β_t^c in either a positive or negative direction:

$$H_0 : \beta_0^c - \beta_0^{\mathcal{M}} = \beta_t^c - \beta_t^{\mathcal{M}} \tag{1}$$

$$H_1 : \beta_0^c - \beta_0^{\mathcal{M}} \neq \beta_t^c - \beta_t^{\mathcal{M}} \tag{2}$$

We independently train a population of $K = 30$ models for each level of pruning, dataset and model that we consider. Thus, for each level of pruning $t \in 0.0, 0.1, 0.3, 0.5, 0.7, 0.9$ we have a sample of 30 functions from the underlying population of possible models trained to that capacity. The pruning

true:	blonde hair	blonde hair	blonde hair	blonde hair	blonde hair
baseline model:	not blonde	blonde hair	blonde hair	not blonde	blonde hair
pruned model:	blonde hair	not blonde	not blonde	blonde hair	not blonde

Figure 2: A selection of CelebA exemplars most impacted by pruning (PIEs). PIEs over-index on spurious correlations between demographic attributes like gender and the true label. This suggests deploying compressed models may incur trade-offs with other properties we may care about such as fairness. Below each image: the ground truth label, the most frequent non-pruned prediction and most frequent pruned prediction (across models trained to 50% pruning).

fraction t indicates the percentage of weights removed. For example, $t = 0.9$ indicates that 90% of model weights are removed over the course of training, leaving a maximum of 10% non-zero weights. Our experimental setup is computationally intensive but necessary to arrive at insights beyond anecdotal observations.

For each class c, we we use a two-sample, two-tailed, independent Welch's t-test (Welch, 1947) to determine whether the mean-shifted class accuracy $S_t^c = \{\beta_{t,k}^c - \beta_{t,k}^{\mathcal{M}}\}_{k=1}^K$ of the samples S_t^c and S_0^c differ significantly. If the p-value $<= 0.05$, we reject the null hypothesis and consider the class to be disparately impacted by t-pruning relative to the baseline. After finding the subset of classes for a given t-pruning 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:

$$\frac{1}{K_t}\sum_{k=1}^{K_t}\left(\beta_{t,k}^c - \beta_{t,k}^{\mathcal{M}}\right) - \frac{1}{K_0}\sum_{k=1}^{K_0}\left(\beta_{0,k}^c - \beta_{0,k}^{\mathcal{M}}\right) \tag{3}$$

The *normalized recall difference* ensures that we are controlling for any overall difference in test-set accuracy between the samples. We ask instead whether the class performed better or worse than expected for a given level of pruning. We include the details of training hyperparameters and the pruning methodology for each dataset in the appendix.

1.2 What do we give up when we prune?

We find that test-set accuracy provides insufficient insight into the trade-offs incurred by pruning. While pruned models are comparable to non-pruned using top-1 accuracy, performance diverges considerably on a small subset of classes which are disproportionately impacted in a statistically significant way (As seen in appendix Fig 5). Certain parts of the data distribution are far more sensitive to varying the number of weights in a network, and bear the brunt cost of altering the network structure. The directionality and magnitude of the impact is nuanced and surprising. Our results show that within the small subset of classes significantly impacted, more classes are relatively robust to the overall degradation experienced than the small subset that degrades in performance far more than the model itself. However, the magnitude of class decreases is larger than the gains (which pulls overall accuracy downwards).

Our findings at a class level prompt the natural question of why certain classes are impacted more than others? To explore this question, we consider the impact of pruning on individual images. Given the limitations of uncalibrated probabilities in deep neural networks (Guo et al., 2017; Kendall & Gal, 2017; Lakshminarayanan et al., 2017), we focus on the level of disagreement between the predictions of pruned and non-pruned networks on a given image. We classify *Pruning Identified Exemplars (PIEs)* as the images where the most frequent prediction differs between a population of independently trained pruned and non-pruned models.

We find that PIEs, the images most sensitive to pruning, overindex on the long-tail of the image distribution. We conduct a human study (85 participants (Appendix Fig. 4) and find that for everyday objects dataset like ImageNet PIEs heavily overindex relative to non-PIEs on certain properties, such

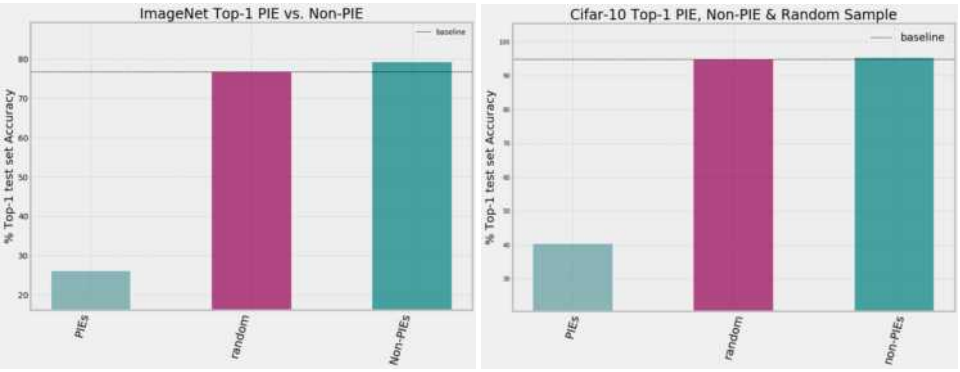

Figure 3: A comparison of model performance on a random sample of 1024 PIE, non-PIE and random images drawn independently from the test-set. Inference on the non-PIE sample improves test-set top-1 accuracy relative to the baseline for both ImageNet and Cifar-10. Inference on PIE images alone substantially degrades generalization performance of pruned models for all datasets considered.

as having an *incorrect ground truth label*, involving a *fine-grained classification task* or *multiple objects*. Over half of all PIE images were classified by human participants as either having an incorrect ground truth label or depicting multiple objects. PIEs are also more challenging for non-pruned models to generalize to. In Fig. 3, we compare the test-set performance of a non-pruned model on a fixed number of randomly selected ImageNet (1) PIE images, (2) non-PIE images and (3) a random sample of the test set. Top-line metrics greatly degrade when inference is restricted to PIE. Removing PIE images from the test-set improves top-1 accuracy for *both* pruned and non-pruned models relative to a random sample.

1.3 PIE AS A HUMAN-IN-THE-LOOP VISUALIZATION TOOL

Our PIE methodology identifies a tractable subset of images which are more challenging for pruned and non-pruned models. On every-day object datasets like ImageNet, PIEs over-index on noisy data which is mislabelled or incorrectly structured for the task. However, sometimes the stakes are higher than correctly classifying `guacamole` or `canoe`. Pruned models are widely used by many real world machine learning applications which often occur in sensitive domains like health care or self-driving cars. To understand how pruning trades off against properties like fairness which are of great concern in these sensitive domains, we consider how varying the number of weights causes performance to diverge on a dataset with a known spurrious correlation between protected demographic attributes.

CelebA has a known spurious association between a target label (blonde vs non-blonde hair color) and two protected demographic attributes which are codified in the dataset ($gender = male, female$ and $age = young, old$). There are 19962 test examples, with 62 in the smallest group (blond-haired old males). If reducing capacity amplified the existing bias of the models towards the underrepresented demographic group *blonde old males*, we would expect to also see PIEs over-index on *blonde old males*. In Table 1 we show that PIEs do indeed do so, which suggests varying the number of weights amplifies existing biases towards underrepresented protected attributes.

Conclusion and future work Our results suggest that while overall accuracy is comparable between pruned and non-pruned models, performance between models with radically different numbers of weights diverges in a significant way on the long-tail of the input distribution. This can introduce unintended trade-offs such as compromising model performance on underrepresented classes or vantage points. We suggest that visualizing PIEs can be a valuable human-in-the-loop machine learning tool which could be used to surface a tractable subset of atypical examples for further human inspection (Leibig et al., 2017; Zhang, 1992), choose not to classify certain examples when the model is uncertain (Bartlett & Wegkamp, 2008; Cortes et al., 2016b;a), or 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., 2019; Bien & Tibshirani, 2011). Further work is needed to understand how PIEs can be leveraged to propose more robust and equitable pruning methods.

REFERENCES

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.

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, Dec 2011. ISSN 1932-6157. doi: 10.1214/11-aoas495. URL http://dx.doi.org/10.1214/11-AOAS495.

Nicholas Carlini, Ulfar Erlingsson, and Nicolas Papernot. Prototypical examples in deep learning: Metrics, characteristics, and utility, 2019. URL https://openreview.net/forum?id=r1xyx3R9tQ.

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.

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. In *ALT*, 2016a.

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.

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.

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.

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.

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

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. A benchmark for interpretability methods in deep neural networks. In *NeurIPS 2019*, 2019.

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.

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.

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. *University of Toronto*, 05 2012.

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.

N. D. Lane and P. Warden. The deep (learning) transformation of mobile and embedded computing. *Computer*, 51(5):12–16, May 2018. ISSN 1558-0814. doi: 10.1109/MC.2018.2381129.

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.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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.

Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, pp. 107–115. Morgan-Kaufmann, 1989.

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

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.

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

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.

Jianping Zhang. Selecting typical instances in instance-based learning. In Derek Sleeman and Peter Edwards (eds.), *Machine Learning Proceedings 1992*, pp. 470 – 479. Morgan Kaufmann, San Francisco (CA), 1992. ISBN 978-1-55860-247-2. doi: https://doi.org/10.1016/B978-1-55860-247-2.50066-8. URL http://www.sciencedirect.com/science/article/pii/B9781558602472500668.

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.

2 APPENDIX

3 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); 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; Ström, 1997; Hassibi et al., 1993b; Zhu & Gupta, 2017; See et al., 2016; Narang et al., 2017). Articulating the trade-offs of compression has overwhelmingly 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 find 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 – ((Carlini et al., 2019; Stock & Cisse, 2017)) or outside of the training distribution (OOD) (Hendrycks & Gimpel, 2016; Lee et al., 2018; Liang et al., 2018; Lee et al., 2018; Masana et al., 2018) and work on calibrating deep neural network predictions (Lakshminarayanan et al., 2017; Guo et al., 2017; Kendall & Gal, 2017).

3.1 METHODOLOGY (INTRODUCING NOTATION)

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 $y_i = [1, ..., C]$.

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 pruning is indicated by t).

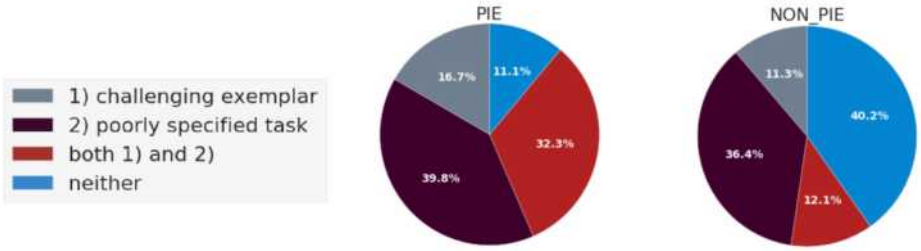

Figure 4: We conducted a human study of the relative distribution of PIE and non-PIE properties. PIE over-indexes significantly on images where multiple classes are visible in the same image and images with incorrect ground truth. **Challenging exemplars:** images positively codified as showing *common 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.*

3.2 Experiment Setup

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%.

3.3 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 9000 steps. For CIFAR-10, we train the model for 80000 steps. We prune every 2000 steps between 1000 and 20000 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 pruning levels. At the end of training, the final pruned mask is fixed and during inference only the remaining weights contribute to the model prediction.

3.4 Human Study

A qualitative inspection of PIEs 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) to label a random sample of 1230 PIE and non-PIE ImageNet images. 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 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):

class label	gender label	young label	# non pie	# pie	non pie pct	pie pct	pct growth
blonde	male	old	57	5	0.29	6.58	2168.97
blonde	male	young	115	3	0.58	3.95	581.03
blonde	female	young	1932	29	9.72	38.16	292.59
blonde	female	old	513	6	2.58	7.89	205.81
non_blonde	female	young	8344	26	41.96	34.21	-18.47
non_blonde	female	old	1395	2	7.01	2.63	-62.48
non_blonde	male	old	2868	2	14.42	2.63	-81.76
non_blonde	male	young	4662	3	23.44	3.95	-83.15

Table 1: We consider the CelebA dataset where there is a correlation between protected demographic attributes codified in the data (young and gender) and the target label (blonde, non-blonde). If reducing capacity amplified the dependence of the model on spurious correlations, we would expect to see PIEs over-index on the spurious correlation. We show that PIEs do indeed do with, with far more *blonde old males* than *non-blonde young males*. We measure this as the relative percent growth in share of each attribute combination in PIE vs non-PIE.

Fraction Pruned	Top 1	# Signif classes	# PIEs
0	94.53	-	-
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 2: CIFAR-10 top-1 accuracy at all levels of pruning, averaged over runs. Top-5 accuracy for CIFAR-10 was 99.8% for all levels of pruning. The fourth column is the number of classes significantly impacted by pruning.

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. [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 RBG 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]

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)

Fraction Pruned	Top 1	Top 5	# Signif classes	# PIEs
0	76.68	93.25	-	-
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 3: ImageNet top-1 and top-5 accuracy at all levels of pruning, averaged over all runs. The fourth column is the number of classes significantly impacted by pruning.

Figure 5: 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% pruning. 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.

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)

					
true:	blonde hair	blonde hair	not blonde	blonde hair	blonde hair
baseline model:	not blonde	not blonde	blonde hair	not blonde	blonde hair
pruned model:	blonde hair	blonde hair	blonde hair	not blonde	not blonde

Table 4: A selection of CelebA PIEs. Below each image: the ground truth label, the most frequent non-pruned prediction and most frequent pruned prediction (across models trained to 50% pruning).

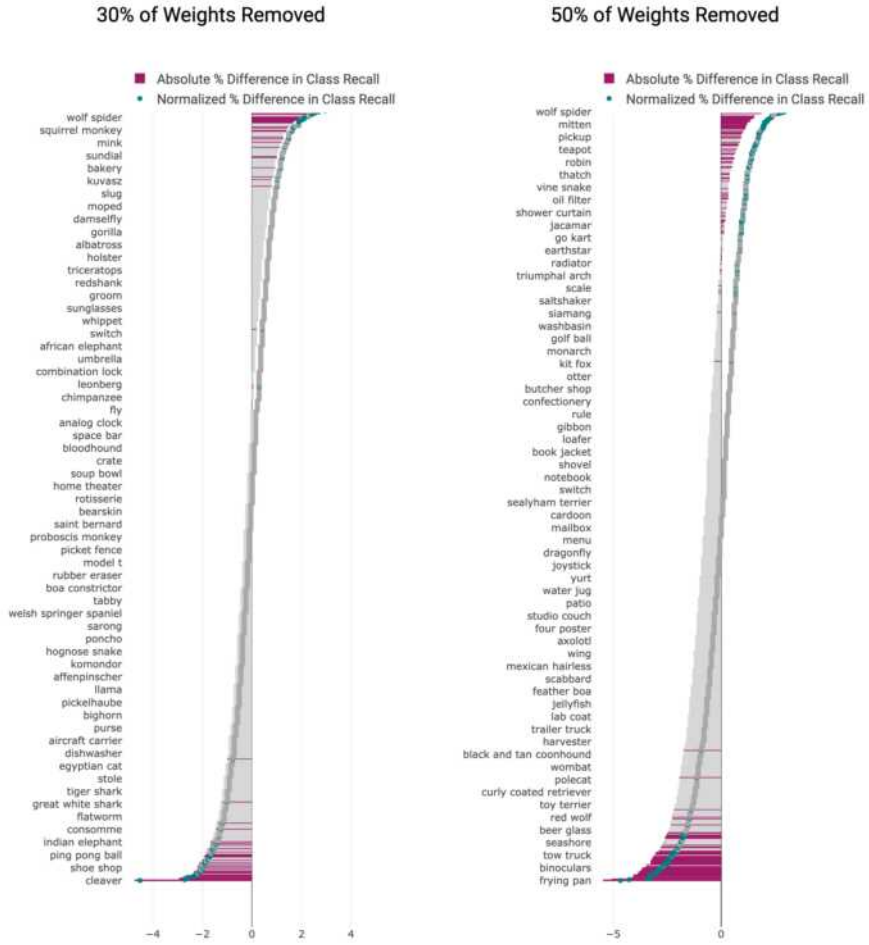

Table 5: We independently train a population of pruned and non-pruned models and apply the t-test to determine whether the means of the samples differ significantly. At all levels of pruning, some classes are impacted far more than others (classes that are statistically significant indicated by pink vs. the classes in grey where the relative change in performance is not statistically significant). We plot both the absolute $\%$ change in class recall (grey and pink bars) and the normalized accuracy relative to change in overall top-1 accuracy caused by pruning (grey and green markers).