DISSECTING PRUNED NEURAL NETWORKS

Jonathan Frankle
MIT CSAIL
jfrankle@csail.mit.edu

David Bau
MIT CSAIL
davidbau@csail.mit.edu

ABSTRACT

Pruning is a standard technique for removing unnecessary structure from a neural network to reduce its storage footprint, computational demands, or energy consumption. Pruning can reduce the parameter-counts of many state-of-the-art neural networks by an order of magnitude without compromising accuracy, meaning these networks contain a vast amount of unnecessary structure.

In this paper, we study the relationship between pruning and interpretability. Namely, we consider the effect of removing unnecessary structure on the number of hidden units that learn disentangled representations of human-recognizable concepts as identified by network dissection. We aim to evaluate how the interpretability of pruned neural networks changes as they are compressed.

We find that pruning has no detrimental effect on this measure of interpretability until so few parameters remain that accuracy beings to drop. Resnet-50 models trained on ImageNet maintain the same number of interpretable concepts and units until more than 90% of parameters have been pruned.

1 INTRODUCTION

Neural network pruning (e.g., LeCun et al. (1990); Han et al. (2015); Li et al. (2016)) is a standard set of techniques for removing unnecessary structure from networks in order to reduce storage requirements, improve computational performance, or diminish energy demands. In practice, techniques for pruning individual connections from neural networks can reduce parameter-counts of state-of-the-art models by an order of magnitude (Han et al., 2015; Gale et al., 2019) without reducing accuracy. In other words, only a small portion of the model is necessary to represent the function that it eventually learns, meaning that—at the end of training—the vast majority of parameters are superfluous. In this paper, we seek to understand the relationship between these superfluous parameters and the interpretability of the underlying model. To do so, we study the effect of pruning a neural network on its interpretability. We consider three possible hypotheses about this relationship:

Hypothesis A: No relationship. Pruning does not substantially alter the interpretability of a neural network model (until the model has been pruned to the extent that it loses accuracy).

Hypothesis B: Pruning improves interpretability. Unnecessary parameters only obscure the underlying, simpler function learned by the network. By removing unnecessary parameters, we focus attention on the most important components of the neural network, thereby improving interpretability.

Hypothesis C: Pruning reduces interpretability. A large neural network has the capacity to represent many human-recognizable concepts in a detectable fashion. As the network loses parameters, it must learn compressed representations that obscure these concepts, reducing interpretability.

Interpretability methodology. We measure the interpretability of pruned neural networks using the *network dissection* technique (Bau et al., 2017). Network dissection aims to identify convolutional units that recognize particular human-interpretable concepts. It does so by measuring the extent to which each unit serves as binary segmenter for that concept on a series of input images. The particular images considered are from a dataset called Broden assembled by Bau et al.; this dataset contains pixel-level labels for a wide range of hierarchical concepts, including colors, textures, objects, and scenes. For each image in Broden, network dissection computes a convolutional unit's activation map, interpolates to expand it to the size of the input image, and segments the image based on the

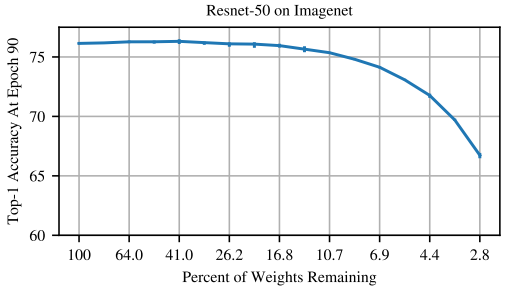

Figure 1: The top-1 accuracy of Resnet-50 on ImageNet when pruned to the specified size.

pixels for which the unit has a high activation according to its typical distribution of activations.[1] Network dissection then computes the size of the intersection of the mask pixels and pixels labeled for particular concepts and divides this quantity by the size of the union; the technique considers units for which this ratio is larger than 0.05 to be interpretable, having learned a disentangled representation of this concept.

Pruning methodology. The neural networks that we dissect are Resnet-50 (He et al., 2016) models trained on the ImageNet (Russakovsky et al., 2015) dataset. We apply a sparse pruning technique, removing weights with the highest magnitudes at the end of training (as in Han et al. (2015), Gale et al. (2019), and Frankle & Carbin (2019)). Doing so produces pruned networks that have fewer parameters but the same number of neurons, meaning these pruned networks retain the capability to contain as many interpretable neurons as the original network.

Immediately after pruning, a neural network's accuracy decreases because part of the model has been removed; pruned networks are typically *fine-tuned* for a small number of training steps to recover accuracy. We use the lottery ticket fine-tuning procedure (Frankle & Carbin, 2019), where the weights of a network are reset back to their values at an iteration early in training. Frankle & Carbin show (and we confirm for our models) that networks trained in this way can learn to match the accuracy of the original network. We choose this fine-tuning approach to allow pruned networks to learn from an early stage of training, potentially learning different functions better adapted to the smaller model. The Resnet-50 networks for ImageNet studied in this paper were uncovered using a modified version of this technique described by Frankle et al. (2019).

We prune Resnet-50 *iteratively*, removing 20% of weights, fine-tuning, and then pruning again. This process produces pruned networks at increments of 20%, making it possible to evaluate the effect of pruning on interpretability as a network is gradually reduced in size. Figure 1 shows the top-1 accuracy of this network as a function of the number of parameters remaining. When 16.8% of parameters or more remain, accuracy matches that of the original network. When 10% of parameters remain, accuracy drops by a percentage point, followed by a steeper decline under further pruning.

Findings. We find that sparse pruning does not reduce the interpretability of Resnet-50 until so many parameters are pruned that accuracy declines, supporting Hypothesis A. We conclude that the parameters that pruning considers to be superfluous for accuracy are also superfluous for interpretability.

2 Results

Network dissection considers both the number of units that learn disentangled concepts and the overall number of Broden concepts learned by any unit. We study these quantities for the final four layers of Resnet-50, comprising 2048 units in total. Based on the analysis of Bau et al., we expect these layers to learn higher-level concepts like objects and scenes.

[1] A high activation is determined by a threshold "T_k such that $P(a_k > T_k) = 0.005$ over every spatial location of the activation map in the data set." (Bau et al., 2017)

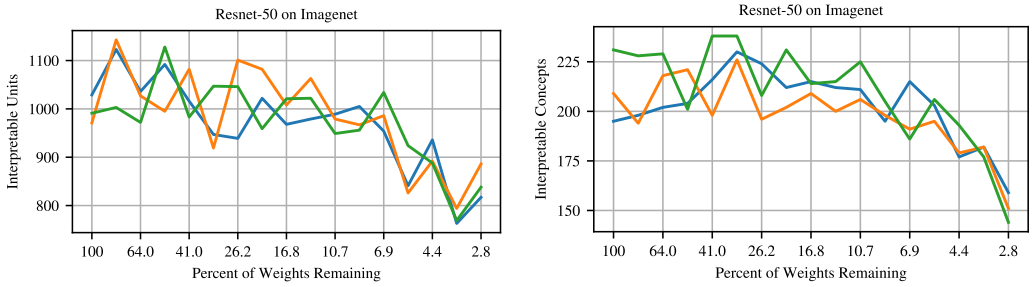

Figure 2: Out of 2048 convolutional units in the fourth group of layers in Resnet-50, the number that learn disentangled concepts (left) and the overall number of concepts learned by any unit (right). Each line represents a separate trial of a model trained with a different initialization. Pruning does not reduce interpretability until it also reduces accuracy (compare to Figure 1).

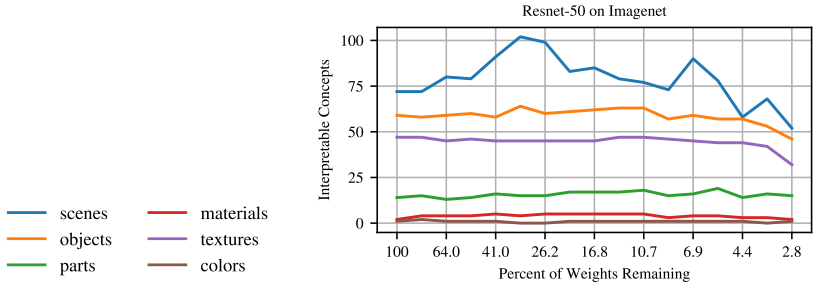

Figure 3: The number of disentangled concepts learned by any unit. The categories are sorted into higher-level Broden categories representing the granularity of each concept. This graph breaks down these concepts for a single trial from Figures 1 and 2.

Interpretability. Figure 2 plots the number of units that learn disentangled concepts (left) and the overall number of concepts learned (right). Each line represents a separate trained Resnet-50 model starting with a different random initialization. All three trials show similar behavior: until 16.8% of parameters remain, the network remains as interpretable as it was before it was pruned; after this point, interpretability begins to gradually decline. This pattern indicates that sparse pruning has little relationship with interpretability (Hypothesis A)—interpretability barely suffers until more than 90% of parameters have been pruned. Instead, this pattern seems to follow the trend of network accuracy (Figure 1). Interpretability begins to decline at the same parameter-count that the network becomes less accurate as a product of over-pruning.

Figure 3 separates a single trial from the right plot of Figure 2 into a taxonomy of concepts according to their level of granularity. Higher-level concepts like scenes and objects seem to be more volatile in the face of pruning. Higher-level concepts are also more likely to disappear as interpretability and accuracy drop at extreme levels of pruning. It is possible that the network's failure to learn as many disentangled, higher-level concepts diminishes its overall accuracy.

Consistency. We use the lottery ticket procedure (Frankle & Carbin, 2019) to fine-tune after pruning, meaning that the pruned networks are re-trained nearly from initialization. In comparison to other fine-tuning strategies, we believe this configuration offers pruned networks more leeway to learn new representations. We are therefore interested in understanding the extent to which the same units are interpretable and learn to recognize the same concepts in the original and pruned networks.

The left graph in Figure 4 shows the percentage of interpretable units in the original network that were also interpretable in the pruned network. Although this figure declines as the networks are pruned, nearly 80% of the originally interpretable units remain interpretable even after 89% of parameters have been pruned. Of these units that were interpretable in both the original and pruned networks, the right graph in Figure 4 explores the consistency of the concepts these units learn. For each pruned

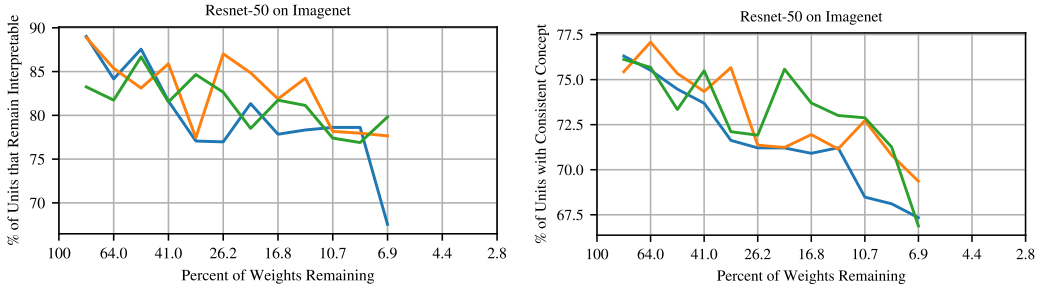

Figure 4: Of the units that are interpretable in the original network, (left) the percent of interpretable units that were interpretable in the original network and (right) the percent of interpretable units that recognize the same concept as they did when they were in the original network. Each line represents a model trained with a different initialization.

network, it plots the percentage of interpretable units that recognize the same concept as they did when they were in the original network, considering only those units that were interpretable both in the original network and in the particular pruned network. As the network is pruned, the fraction of such *consistent* units declines. However, it remains relatively high: about 70% of such units learn to recognize the same concept even after 89% of parameters are pruned.

3 Discussion and Future Work

In this short paper, we only consider a sparse pruning technique that preserves the number of units in the network. It is possible that, if entire convolutional filters were pruned as in (Li et al., 2016), a completely different set of behaviors might result. For example, the network might become less interpretable with pruning as it has less capacity with which to develop intermediate representations. Or alternatively, if—as Liu et al. (2019) argue—pruned convolutional filters were never necessary to begin with, then the network would remain equally interpretable but with a higher percentage of interpretable units (convolutional filters are units with respect to network dissection).

It is also possible that another fine-tuning strategy might produce different results. The lottery ticket strategy allows the network to retrain nearly from the start after each round of pruning, meaning that the network has the opportunity to learn entirely new representations. In contrast, standard pruning techniques retain the trained weights of unpruned connections and fine-tune for a small number of iterations at a low learning rate, severely limiting the network's ability to learn new representations. It would be interesting to compare the interpretability of the networks produced by each approach. It is possible that lottery ticket fine-tuning makes it possible to learn new, disentangled representations for the smaller network size, or, alternatively, that limited fine-tuning more effectively sustains the interpretability of the unpruned networks.

For this workshop paper, we only consider Resnet-50. It would be valuable to study the extent to which the behavior we observe extends to other networks as in Bau et al. (2017).

4 Conclusions

We study the interpretability of pruned networks. Specifically, we use network dissection (Bau et al., 2017) to examine the number of units that learn to recognize disentangled, human-identifiable concepts in networks whose weights have been removed using lottery ticket pruning (Frankle & Carbin, 2019; Frankle et al., 2019). We find that this sparse pruning has no impact on the interpretability of the Resnet-50 model (as trained on ImageNet) until so many parameters are pruned that accuracy beings to decline. We conclude that parameters considered unnecessary by magnitude pruning are also unnecessary to maintain the level of interpretability of the unpruned model. However, pruning does not cause interpretability to improve either.

REFERENCES

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Int. Conf. Represent. Learn.*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611*, 2019.

Trevor Gale, Erich Elsen, and Sarah Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pp. 1135–1143, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pp. 598–605, 1990.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJlnB3C5Ym`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.