Jonathan Frankle Research Statement

I develop new experimental methods for identifying empirical properties of the neural networks we use in practice. In my research, I have adapted neural network *pruning* (a technique for removing superfluous structure from neural networks) to determine the capacity that a network needs to learn a particular task. I have shown that practical neural networks are capable of learning just as effectively with an order of magnitude fewer parameters than we typically use in practice, shedding new light on the nature of neural network optimization and creating new opportunities to reduce the cost of training.

Motivation & Overview

In the rapidly growing area of deep learning, we face fundamental challenges that make it especially difficult to show that properties hold for the neural networks that we create. We do not understand the role that each component of a trained neural network has come to serve, denying us the scaffolding that typically facilitates reasoning about systems. Many useful properties may only hold in practice for specific tasks and data distributions (e.g., natural image classification) on specific architectures (e.g., ResNet) optimized in a specific way (e.g., gradient descent); at practical scales, each of these objects defies formalization with existing tools. The properties of a network are tightly connected to those of the training process—the sophisticated interplay between architecture, dataset, and optimizer that determined the network's function and representation—yet there remain many questions about how training behaves in practice and why it reliably produces useful networks.

However, the same complexity that poses these challenges also gives rise to important, nonobvious properties of neural networks that, for example, threaten reliability (e.g., adversarial examples [Goodfellow et al., 2015]) or offer new opportunities to reduce the cost of training [Frankle & Carbin, 2019]. My research goal is to identify such properties in practical neural networks, which I believe is essential if we wish to understand how they learn in real-world settings, enhance the efficiency of training and deploying them, and improve our confidence in systems based on deep learning as they become increasingly ubiquitous.

My Contribution: The Lottery Ticket Hypothesis

My research provides a new understanding of the capacity that neural networks need to learn. We have long known that practical networks are larger than necessary to represent the functions they learn: we can *prune* a large fraction (typically 80-90%) of their weights after training without affecting accuracy [Le Cun et al., 1990; Han et al., 2015]. I study whether optimization—the process of learning these functions—needs these additional weights, even if the final representations do not. My proposed *lottery ticket hypothesis* [Frankle & Carbin, 2019] states that there exist similarly small *subnetworks* that—from early in training—can train on their own to the same accuracy as the full network in the same amount of time. In other words, networks can be an order of magnitude smaller than standard practice suggests while still reaching the same accuracy.

To evaluate this hypothesis, I designed an experiment to find such subnetworks retroactively and accumulated comprehensive evidence that they exist in standard settings for computer vision [Frankle & Carbin, 2019; Frankle et al. 2020b]. The results overturned years of received wisdom that networks of this size could not train effectively [Han et al., 2015; Li et al., 2017] and created a new area dedicated studying these networks. In recognition of this contribution, my work received a best paper award at ICLR 2019.

Empiricism for Identifying Properties of Practical Deep Learning Systems

More broadly, I identify properties of neural networks by studying them empirically as would an experimental biologist or physicist studying natural systems. This process involves posing hypotheses, developing experiments to assess whether they hold, and rigorously evaluating them on a range of settings. The advantage of this approach is that we can directly analyze the settings we use in practice—settings that remain beyond the reach of our current theory. Once well-supported hypotheses are established, we are in a stronger position to formalize these ideas. And, since we are examining practical settings, our insights have a clearer path for improving the practical state of the art.

Technical Work

Problem Statement: How much capacity do we need to train a neural network?

Since the 1980s, we have known that it is typically possible to *prune* most of the parameters from trained neural networks without affecting accuracy [Le Cun et al., 1990; Reed, 1993; Han et al., 2015; Blalock et al., 2020]. The fact that it is possible to do so means that the networks do not need their full capacity to represent the functions that they learn. However, until my work, it was believed that the training process requires more capacity than does this final representation [Han et al., 2015; Li et al., 2017; Ziv & Tishby, 2017]; that is, it is necessary to train the full network before pruning. This difference has important practical implications: pruning after training reduces the cost of using the network (*inference*), while pruning before or early in training could reduce the cost of creating the network in the first place.

The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

I showed that these pruned networks can indeed be trained on their own to full accuracy. My insight was that weights that can be pruned after training could have been pruned before training; however, to allow these subnetworks to train effectively, each weight needs to be set to the same initial value as it received before training. We typically initialize neural networks by sampling their weights from distributions designed to maintain gradients of appropriate scales during backpropagation [Glorot & Bengio, 2010; He et al., 2015]. Others had tried to train pruned networks without success, and this distributional view of initialization was at the heart of those failures: before training, the networks were always reinitialized by sampling a new initialization from these distributions [Han et al., 2015; Li et al, 2017].

To evaluate my claims, I developed an experiment: after pruning the network, *rewind* each weight back to the initial value it had before training and then train just this subnetwork. If this subnetwork can train to full accuracy, then it is evidence that such small, trainable networks exist. And if reinitializing this subnetwork leads to lower accuracy, then it is evidence that the original initialization was important. On a range of small-scale computer vision networks, I showed that this was indeed the case. The takeaway is that, in these settings, there exist *winning ticket* subnetworks that have "won the initialization lottery," making it possible for them to train from the start to the same accuracy as the unpruned network.

Linear Mode Connectivity and the Lottery Ticket Hypothesis

In larger-scale settings, my procedure only worked when the networks were trained with learning rate *warmup* (where the learning rate is gradually increased from zero to the full value over the early part of training) [Frankle et al., 2020b]. Warmup has been established as a technique to mitigate the effects of noise in the optimization process early in training [e.g., Goyal et al., 2018], and this observation suggested that the early part of training might pose particular difficulties for training pruned networks.

To evaluate this hypothesis, I developed an experiment to assess a network's sensitivity to the noise of stochastic gradient descent (SGD - a standard optimization procedure used to train neural networks): train two copies of the pruned network with two different samples of noise (in this case, two different random data orders). If they reach the same linearly connected optimum, then noise must not have been large enough to affect the outcome of training, i.e., the network is *stable* to noise. Otherwise, noise affected the outcome of training, i.e., the network is *unstable* to noise. Pruned networks where the lottery ticket observations held were stable to noise in this way, and those where they did not hold were unstable to noise.

Not only did this distinguish successes and failures of my procedure, but it provided a crucial insight for finding small, trainable networks at larger scales without modifying the learning rate schedule: look for subnetworks that are stable to noise. The success of warmup in particular suggested that noise was only a challenge during the early part of training. To evaluate this hypothesis, I studied the stability of subnetworks created after initialization: prune after training (just like before) and rewind the weights to their values from iteration $k > 0$ (rather than their values at initialization). At an early point in training, all settings became stable to SGD noise; moreover, accuracy improved alongside stability, culminating in stable subnetworks that could train on their own from early in training to full accuracy.

Impact

The high-level takeaway from my research is a generalization of my original hypothesis: practical neural networks contain smaller subnetworks capable of training to full accuracy from the point in training when they are stable. For the smaller-scale settings in the original work, this is at initialization. For larger-scale settings, it is early (1% to 5%) in training. Today, this rewinding procedure is the standard way to study lottery tickets in the literature [e.g., Morcos et al., 2019], my results have been extended to a range of other settings [Yu et al., 2019; Chen et al., 2020; Kalibhat et al., 2020], and my experiment for analyzing stability has become a research topic in its own right [e.g., Mirzadeh et al., 2020].

By showing that training is possible without the full capacity of the network, this line of work has catalyzed an area of deep learning dedicated to studying winning tickets and, more generally, to training pruned networks. Dozens of papers have been published on the subject in recent conferences, for example showing that winning tickets can transfer to new tasks [Morcos et al., 2019], proving variants of the hypothesis [Malach et al., 2020; Pensia et al. 2020], proposing ways to find these subnetworks efficiently [Wang et al., 2020; Tanaka et al., 2020], and suggesting other strategies for training smaller networks [Evci et al., 2020].

Future Work

Practical Applications of the Lottery Ticket Hypothesis

Although my research shows that it would have been possible to use a smaller network for much or all of training, it has not yet delivered an efficient way to find these networks. The most important practical open problem in this research area is developing a procedure to find these subnetworks as early in training as possible. Doing so would create new opportunities to significantly reduce the cost of training (5x-10x if we can also accelerate sparsity–see below). I am actively pursuing this question, starting with a rigorous analysis of the shortcomings of existing proposals [Frankle et al., 2020c] and continuing by taking inspiration from my ideas about the stability of training to SGD noise [Frankle et al., 2020b]. To solve this problem, I believe we will need a to gain a better scientific understanding of how neural networks learn and use those insights to develop new pruning methods.

Even once we have effective methods for training smaller networks, we still need to ensure that these networks improve efficiency in practice. When pruning individual weights in an unstructured fashion, the resulting networks contain *sparse* tensors. Although current-generation CPUs and GPUs do not immediately benefit from sparsity, there is an active research area dedicated to writing libraries to accelerate sparse neural networks on these platforms [e.g., Elsen et al., 2020] and next generation hardware has native sparsity support (e.g., the NVIDIA A100, GraphCore IPU, and Cerebras Wafer-Scale Engine). To ensure that my research realizes its practical potential, I intend to both build on existing work to accelerate sparsity and to co-design pruning techniques conducive to efficient implementations on the increasingly diverse array of hardware accelerators for deep learning.

Sparsity as a Tool for Understanding Deep Learning

My research shows that *sparsity* (fixing parameters to zero) can emerge earlier than previous understood in the neural networks we train in practice: neural networks learn functions with sparse representations and, when the right parameters are removed, they can train in a sparse fashion. These observations shed new light on existing questions and suggest many new questions at the heart of our understanding of deep learning. Why does sparsity emerge in the first place? Is it intrinsic to neural networks in general, or is it a product of choices we make in practice (e.g., the nature of real-world data, the architectures we use in practice, or implicit biases of SGD [Neyshabur et al., 2015])? Does sparsity imply that neural network optimization occurs in a lower-dimensional subspace [Gur-Ari et al., 2018] and, if so, can we identify it?

Furthermore, why are sparse networks sensitive to initialization in ways that the dense neural networks we typically train are not? This ties to broader questions about the role of *overparameterization* (where a neural network has enough capacity to memorize the training set) in deep learning [Belkin et al., 2019; Liu

et al., 2020]. Can we only sparsify neural networks in this overparameterized regime? Is optimizing a sparse neural network intrinsically more difficult than the dense networks, or does sparsity simply require a different approach to initialization and optimization? With a better understanding of the emergence of sparsity and the nature of sparse optimization problems, we may eventually be able to identify and train sparse networks from scratch without needing to resort to pruning at all.

Empiricism for Identifying Properties of Practical Deep Learning Systems

More broadly, my research demonstrates the value of empiricism as a tool for identifying new properties of neural networks. As my research illustrates, empiricism makes it possible to investigate questions, behaviors, and settings that arise in practice but may be beyond the reach of existing theory. Such new empirical observations made in practical settings can lead to well-motivated practical advances while identifying new directions for theoretical work.

In my future research, I intend to continue to develop empirical approaches for studying practical neural networks. My research philosophy is that, in areas where the community is competing to chase state-of-the-art numbers, there is an opportunity to scientifically understand and systematize the properties underlying that success [e.g., Blalock et al, 2020; Renda et al., 2020; Frankle et al., 2020c]. In general, practical deep learning systems engineered to reach high accuracy on real-world datasets will inevitably be complex, and it is my conviction that empiricism will be an essential tool as we seek to understand and improve these systems amid this complexity.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal. Reconciling Modern Machine Learning Practice and the Classical Bias-Variance Trade-off. Proceedings of the National Academy of Sciences, 2019.

Davis Blalock, Jose Javier Gonzalez Ortiz, **Jonathan Frankle**, John Guttag. What is the State of Neural Network Pruning? Machine Learning and Systems, 2020.

Tianlong Chen, **Jonathan Frankle**, Shiyu Chang, SIjia Liu, Yang Zhang, Zhangyang Wang, Michael Carbin. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. Neural Information Processing Systems, 2020.

Erich Elsen, Marat Dukhan, Trevor Gale, Karen Simonyan. Fast Sparse ConvNets. Computer Vision and Pattern Recognition, 2020.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. International Conference on Learning Representations, 2019.

Jonathan Frankle, David J. Schwab, Ari S. Morcos. The Early Phase of Neural Network Training. International Conference on Learning Representations, 2020a.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, Michael Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. International Conference on Machine Learning, 2020b.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, Michael Carbin. Pruning Neural Networks at Initialization: Why are we missing the mark? ArXiv preprint, 2020c.

Trevor Gale, Erich Elsen, Sara Hooker. The State of Sparsity in Deep Neural Networks. ArXiv preprint, 2019.

Utku Evci, Erich Elsen, Pablo Castro, Trevor Gale. Rigging the Lottery: Making All Tickets Winners. International Conference on Machine Learning, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. Artificial Intelligence and Statistics, 2010.

Priya Goyal, Piotr Dollar, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He. Accurate Large Minibatch SGD: Training ImageNet in 1 Hour. CVPR, 2017.

Guy Gur-Ari, Daniel A. Roberts, Ethan Dyer. Gradient Descent Happens in a Tiny Subspace. Arxiv preprint, 2018.

Song Han, Jeff Pool, John Tran, William J. Dally. Learning both Weights and Connections for Efficient Neural Networks. Neural Information Processing Systems, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. IEEE International Conference on Computer Vision, 2015.

Neha Mukund Kalibhat, Yogesh Balaji, Soheil Feizi. Winning Lottery Tickets in Deep Generative Models. Arxiv, 2020.

Yann Le Cun, John S. Denker, Sara A. Solla. Optimal Brain Damage. Neural Information Processing Systems, 1990.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, Hans Peter Grap. Pruning Filters for Efficient ConvNets. International Conference on Learning Representations, 2017.

Chaoyue Liu, Libin Zhu, Mikhail Belkin. Toward a Theory of Optimization for Over-Parameterized Systems of Non-Linear Equations: The Lessons of Deep Learning. Arxiv, 2020.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, Trevor Darrell. Rethinking the Value of Network Pruning. International Conference on Learning Representations, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, Ohad Shamir. Proving the Lottery Ticket Hypothesis: Pruning is All You Need. International Conference on Machine Learning, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, Hassan Ghasemzadeh. Linear Mode Connectivity in Multitask and Continual Learning. ArXiv preprint, 2020.

Ari S. Morcos, Haonan Yu, Michela Paganini, Yuandong Tian. One Ticket to Win Them All: Generalizing Lottery Ticket Initializations Across Datasets and Optimizers. Neural Information Processing Systems, 2019.

Benham Neyshabur, Ryota Tomioka, Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. International Conference on Learning Representations, 2015.

Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, Dimitris Papailiopoulos. Optimal Lottery Tickets via SubsetSum: Logarithmic Over-Parameterization is Sufficient. Neural Information Processing Systems, 2020.

Russell Reed. Pruning Algorithms—A Survey. IEEE Transactions on Neural Networks, 1993.

Alex Renda, **Jonathan Frankle**, Michael Carbin. Comparing Rewinding and Fine-Tuning in Neural Network Training. ICLR, 2020.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. ArXiv preprint, 2017.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, Surya Ganguli. Pruning Neural Networks Without Any Data by Iteratively Conserving Synaptic Flow. Neural Information Processing Systems, 2020.

Chaoqi Wang, Guodong Zhang, Roger Grosse. Picking Winning Tickets Before Training by Preserving Gradient Flow. International Conference on Learning Representations, 2019.

Haonan Yu, Sergey Edunov, Yuandong Tian, Ari S. Morcos. Playing the Lottery with Rewards and Multiple Languages: Lottery Tickets in RL and NLP. International Conference on Learning Representations, 2020.