Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Doctoral Consortium Track

Learning Sparse Neural Networks for Better Generalization

Shiwei Liu
Eindhoven University of Technology, The Netherlands
s.liu3@tue.nl

Abstract

Deep neural networks perform well on test data when they are highly overparameterized, which, however, also leads to large cost to train and deploy them. As a leading approach to address this problem, sparse neural networks have been widely used to significantly reduce the size of networks, making them more efficient during training and deployment, without compromising performance. Recently, sparse neural networks, either compressed from a pre-trained model or obtained by training from scratch, have been observed to be able to generalize as well as or even better than their dense counterparts. However, conventional techniques to find well fitted sparse sub-networks are expensive and the mechanisms underlying this phenomenon are far from clear. To tackle these problems, this Ph.D. research aims to study the generalization of sparse neural networks, and to propose more efficient approaches that can yield sparse neural networks with generalization bounds.

1 Introduction

Occam's razor is a well-known principle of parsimony, anchored in scientific thinking in general and incorporated in practical statistical problems. Applied to deep learning, it implies that given two hypotheses explaining the data equally well, the simpler hypothesis is preferable. While modern deep neural networks generalize well with high overparameterization, the resources required to train and infer these networks can be prohibitive. It is difficult to deploy those models with hundreds of millions of parameters to resource-limited devices.

As an active area of current research, sparse neural networks [LeCun *et al.*, 1990; Han *et al.*, 2015] have been shown to be an effective approach to address these challenges. Sparse neural network structures are also more in line with the natural structure of the human brain where the connections between neurons are highly sparse. Motivated by the limited computational capacity and memory storage on mobile devices, various techniques including but not limited to pruning, L_0 and L_1 regularization and Bayesian methods, can effectively achieve inference efficiency by yielding a sparse neural network as an output of a pre-training phase which, however, add extra computational cost during the training phase. Recently, several approaches [Mocanu *et al.*, 2018; Mostafa and Wang, 2019] have been proposed to train sparse neural networks from scratch with a fixed parameter budget based on adaptive sparse connectivity instantiated by Sparse Evolutionary Training (SET) [Mocanu *et al.*, 2018]. By solving a combinatorial optimization problem (weights and sparse sub-networks), these techniques can find sparse neural networks reach test accuracy comparable to the dense network without retraining and fine-tuning. However, off-the-shelf methods are mainly for feedforward networks (e.g., convolutional neural networks, multi-layer perceptrons) on image recognition. Besides, due to the lack of hardware and libraries, the training efficiency provided by unstructured sparse neural networks can not be mapped to the parallel processors. Thus, novel and more efficient approaches are required.

Furthermore, recent research has shown that sparse neural networks with a small fraction of parameters can generalize better than the dense networks [Liu *et al.*, 2019b; Liu *et al.*, 2019c; Liu *et al.*, 2019a]. These observations can be treated as empirical evidence standing for the effectiveness of Occam's razor in the neural network regime. However, these insights have been found empirically, and a theory explaining these effects is still pending.

Based on the above-mentioned challenges, this research attempts to highlight the efficiency and effectiveness of sparse neural networks with three long term research goals: (1) to propose scalable sparsity inducing techniques that can yield sparse neural networks with state-of-the-art sparse performance on various neural network architectures; (2) to propose novel techniques inducing sparse neural networks that can be optimized by modern libraries and hardware; (3) to study the generalization of sparse neural networks from a theoretical perspective and study generalization bounds for sparse neural networks with adaptive sparse connectivity.

2 Contributions

For the first goal, we have proposed a novel class of sparsity inducing approaches. We develop intrinsically sparse recurrent neural networks (RNNs), which is more challenging to compress than CNNs and MLPs, due to the recurrent structure and the long-term dependencies over different

time steps. The results demonstrate that our method can discover a sparse sub-network with a single run, which usually achieves better performance than dense RNNs [Liu *et al.*, 2019b]. Furthermore, we invested a novel approach, termed Selfish-RNNs, that yields sparse RNNs significantly improving the language modeling performance for various RNN-based models on large scale datasets. The main contributions are two-fold. (1) We optionally allow redistributing weights across different RNN cells during training to better regulate information. (2) We proposed a new optimizer, sparse non-monotonically triggered averaged stochastic gradient descent (Sparse NT-ASGD), a variant of NT-ASGD [Merity *et al.*, 2017] that remedies the structural damage caused by the average operation of NT-ASGD.

For the second goal, we have devised the first truly sparse implementation for adaptive sparse connectivity [Liu *et al.*, 2019a][1]. Although adaptive sparse connectivity has shown its ability to reduce parameter-counts and to reach higher test accuracy, the sparsity is enforced by binary masks in the off-the-shelf work, since GPU-accelerated libraries have limited support for sparse operations. In this work, we address the above limitations of adaptive sparse connectivity and implemented SET from scratch using just Python, Scipy and Cython. With this efficient implementation, we are able to train truly sparse MLPs with over one million neurons on a typical laptop without GPU, which scales dramatically better than state-of-the-art techniques.

For the third goal, we have empirically shown that intrinsically sparse MLPs have better generalization capabilities than their fully-connected counterparts [Liu *et al.*, 2019c]. We have also developed a method to understand sparse neural network topologies from the perspective of graph theory. More specifically, we have introduced an approach to measure the topology similarity between different sparse neural networks based on Graph Edit Distance (GED). By visualizing the topological optimization process of adaptive sparse connectivity, we have shown that there are a lot of low-dimensional structures (sparse neural networks) that have the potential to substitute the highly overparameterized dense models via sparse topology optimization. Moreover, we have shown that sparse neural networks with high sparsity can significantly reduce sharpness of the minimizers. [Keskar *et al.*, 2016].

3 Conclusion and Future Research

In this research work, we have proposed some techniques to yield efficient and effective sparse neural networks and to provide some intuitions of sparse neural networks. However, the mechanisms underlying the superior generalization of sparse neural networks have not been fully explored yet. Our conjecture is that adaptive sparse connectivity helps the continuous optimization (for connection weights) escape the local optima or a local saddle point by changing the loss function landscape iteratively. In our future work, we intend to develop a theory on sparse neural networks and provide state-of-the-art generalization guarantees for them.

[1] https://github.com/Shiweiliuiiiiiii/SET-MLP-ONE-MILLION-NEURONS

Acknowledgements

I sincerely thank Dr. Robert Peharz for helpful discussions and feedback on drafts of this paper.

References

[Han *et al.*, 2015] 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*, pages 1135–1143, 2015.

[Keskar *et al.*, 2016] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

[Liu *et al.*, 2019a] Shiwei Liu, Decebal Constantin Mocanu, Amarsagar Reddy Ramapuram Matavalam, Yulong Pei, and Mykola Pechenizkiy. Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware. *arXiv preprint arXiv:1901.09181*, 2019.

[Liu *et al.*, 2019b] Shiwei Liu, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Intrinsically sparse long short-term memory networks. *arXiv preprint arXiv:1901.09208*, 2019.

[Liu *et al.*, 2019c] Shiwei Liu, Decebal Constantin Mocanu, and Mykola Pechenizkiy. On improving deep learning generalization with adaptive sparse connectivity. In *ICML Workshop on Understanding and Improving Generalization in Deep Learning*, 2019.

[Merity *et al.*, 2017] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

[Mocanu *et al.*, 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*, 9(1):1–12, 2018.

[Mostafa and Wang, 2019] Hesham Mostafa and Xin Wang. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *Proceedings of the 36th International Conference on Machine Learning-Volume 97*, pages 4646–4655. JMLR. org, 2019.