Cost-effective Artificial Neural Networks

Zahra Atashgahi
University of Twente, The Netherlands
z.atashgahi@utwente.nl

Abstract

Deep neural networks (DNNs) have gained huge attention over the last several years due to their promising results in various tasks. However, due to their large model size and over-parameterization, they are recognized as being computationally demanding. Therefore, deep learning models are not well-suited to applications with limited computational resources and battery life. Current solutions to reduce computation costs mainly focus on inference efficiency while being resource-intensive during training. This Ph.D. research aims to address these challenges by developing cost-effective neural networks that can achieve decent performance on various complex tasks using minimum computational resources during training and inference of the network.

1 Introduction

While DNNs have gained increasing attention in recent years, there are major concerns regarding the ever-increasing computational cost of training and deployment of these models, caused by large model sizes and over-parameterized layers. This can be severely worsened when DNNs are applied to high-dimensional data, leading to issues such as the curse of dimensionality and over-fitting. Furthermore, deploying DNNs on real-world applications often has demanding real-time requirements including learning from data streams, online learning, and adapting the model to abrupt variations in data distribution that eventually result in growing costs. As a result, training and inference of DNNs on low-resource devices, e.g., an edge device with limited computational resources and battery life, might not be economically viable. In addition, such an increase in computations can lead to a critical rise in the energy consumption in data centers and, consequently, results in extremely high carbon emissions and adverse environmental effects [Strubell *et al.*, 2020].

Sparse neural networks (SNNs) have been a popular approach toward addressing the over-parameterization of DNNs. By keeping only the most important connections of a DNN, they achieve a comparable result to their dense counterpart while having much fewer parameters. These networks can be achieved either by pruning DNNs (a.k.a *dense-to-sparse*) or training SNNs sparsely from scratch (a.k.a *sparse-to-sparse*). The latter approach is more efficient as it never uses dense matrices throughout training. By exploiting dynamic sparse connectivity during training, which started to be known in the literature as *Dynamic Sparse Training* (DST), this class of algorithms often matches or outperforms the dense equivalent networks in terms of accuracy while requiring significantly fewer training and inference FLOPs. In this Ph.D. research, we aim to leverage the power of SNNs and the DST framework in developing cost-effective artificial neural networks. The research questions are as follows:

- **RQ1:** How can we develop neural networks to learn unknown data distribution with minimum resources?
- **RQ2:** How can we address the challenges imposed by high-dimensional data using SNNs?
- **RQ3:** How can we design DNNs to learn time series efficiently?
- **RQ4:** How can we improve the generalization of DNNs on high-dimensional tabular datasets using SNNs?

All in one, we seek to answer *How we can reduce the training and deploying costs associated with today's deep learning models without compromising performance?*

2 Contributions

RQ1: To address this research question, we tried to first gain better insight into SNNs trained within the DST framework. For this purpose, in [Liu *et al.*, 2020], we analyzed the structure of SNNs trained with a DST algorithm. We proposed the first method to compare different SNN topologies, named Neural Network Sparse Topology Distance (NNSTD), based on graph theory. We demonstrated that there are many very different SNNs that can outperform their dense equivalent network in terms of accuracy, and DST is an effective technique to find them. Then, in [Atashgahi *et al.*, 2022b], we aim to improve the training of SNNs, which are trained sparsely from scratch with DST. By integrating the concept of Hebbian learning into the evolution process of weights during the training of SNNs, we obtain a new effective DST algorithm that outperforms several state-of-the-art DST algorithms in extremely sparsity regions by a large gap. Finally, in [Liu *et al.*, 2021] and [Liu *et al.*, 2022], we developed efficient learning algorithms for SNNs, demonstrating that they can match or even outperform their dense counterparts with much fewer parameters and computations.

RQ2: We sought to alleviate the burden of high computational costs and memory requirements imposed by the rise of big data. Feature selection, which identifies the most relevant and informative attributes of a dataset, addresses the problem of high dimensionality. However, most existing feature selection methods are computationally expensive. In [Atashgahi *et al.*, 2022c], we tackled this problem using SNNs; we proposed an energy-efficient method to select features from high-dimensional data. The proposed method, named "QuickSelection," introduces the strength of the neuron in SNNs as a criterion to measure the feature importance. Using a SNN to perform feature selection, our proposed method achieves the best trade-off between accuracy and computational efficiency when compared with several baselines. To further improve the feature selection performance on high-dimensional data, in [Atashgahi *et al.*, 2022d], we introduced dynamic input neuron evolution into the training of a SNN. We proposed an efficient supervised feature selection method that outperforms state-of-the-art supervised feature selection models on several real-world benchmark datasets. Finally, in [Sokar *et al.*, 2022], we propose an efficient unsupervised method that encourages the network to pay attention to important features quickly and speed up the training process.

RQ3: Within this research question, we aim to develop methods to learn from data streams efficiently. One of the significant issues when learning from data streams is Change Point Detection (CPD). Existing solutions have major issues, including large memory requirements, offline detection, dependence on the choice of hyperparameters, and expensive computations, making them inadequate for real-world applications. To address these issues, in [Atashgahi *et al.*, 2022a], we designed a novel LSTM Autoencoder-based network to perform memory-free online unsupervised CPD. In another work, we aim to learn from time series data efficiently [Atashgahi *et al.*, 2023]. Particularly, we focus on decreasing the computational and memory costs of training and deploying transformers for time series forecasting. We propose a novel method to obtain SNNs, that exploits loss heuristics to automatically find the best trade-off between loss and sparsity in one round of training. By performing experiments on six benchmark datasets and five SOTA transformer variants for time series forecasting, we show that PALS reduces the model's size (reducing 65% parameters on average) and computation (63% FLOPs reduction on average) substantially while maintaining comparable performance to the dense counterpart in terms of prediction loss.

3 Conclusion and Future Research

In this paper, we discussed challenges emerging with deploying DNNs on real-world applications and presented our solutions that address some of these challenges including efficient learning from data using SNNs, feature selection, and learning from data streams. Further in our research, we plan to investigate whether SNNs can improve the generalization on high-dimensional tabular data (RQ4). Despite the remarkable performance of deep learning in various fields of applications, it often falls short when compared to traditional machine learning methods on tabular datasets. Therefore, we aim to explore how we can obtain SNNs to improve the generalization of DNNs on tabular datasets. We expect this research to develop algorithms that can lower the high computational costs imposed by deep learning models during training and inference. In the end, we hope that this Ph.D. research will pave the way to designing cost-effective neural networks.

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Dr. Decebal Mocanu, Prof.dr. Raymond Veldhuis, and Prof.dr. Mykola Pechenizkiy for their guidance and mentorship. This research is partially funded by NWO EDIC project.

References

[Atashgahi *et al.*, 2022a] Zahra Atashgahi, Decebal Constantin Mocanu, Raymond Veldhuis, and Mykola Pechenizkiy. Memory-free online change-point detection: A novel neural network approach. *arXiv preprint arXiv:2207.03932*, 2022.

[Atashgahi *et al.*, 2022b] Zahra Atashgahi, Joost Pieterse, Shiwei Liu, Decebal Constantin Mocanu, Raymond Veldhuis, and Mykola Pechenizkiy. A brain-inspired algorithm for training highly sparse neural networks. *Machine Learning (ECML-PKDD Journal track)*, 2022.

[Atashgahi *et al.*, 2022c] Zahra Atashgahi, Ghada Sokar, Tim van der Lee, Elena Mocanu, Decebal Constantin Mocanu, Raymond Veldhuis, and Mykola Pechenizkiy. Quick and robust feature selection: the strength of energy-efficient sparse training for autoencoders. *Machine Learning (ECML-PKDD Journal track)*, 111(1), 2022.

[Atashgahi *et al.*, 2022d] Zahra Atashgahi, Xuhao Zhang, Neil Kichler, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Raymond Veldhuis, and Decebal Constantin Mocanu. Feature selection with neuron evolution in sparse neural networks. In *TMLR*, 2022.

[Atashgahi *et al.*, 2023] Zahra Atashgahi, Mykola Pechenizkiy, Raymond Veldhuis, and Decebal Constantin Mocanu. Adaptive sparsity level during training for efficient time series forecasting with transformers. *arXiv preprint arXiv:2305.18382*, 2023.

[Liu *et al.*, 2020] Shiwei Liu, Tim van der Lee, Anil Yaman, Zahra Atashgahi, Davide Ferraro, Ghada Sokar, Mykola Pechenizkiy, and Decebal C Mocanu. Topological insights into sparse neural networks. In *Proceedings of ECML-PKDD*, 2020.

[Liu *et al.*, 2021] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Zahra Atashgahi, Lu Yin, Huanyu Kou, Li Shen, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Sparse training via boosting pruning plasticity with neuroregeneration. In *NeurIPS*, 2021.

[Liu *et al.*, 2022] Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity. In *ICLR*, 2022.

[Sokar *et al.*, 2022] Ghada Sokar, Zahra Atashgahi, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Where to pay attention in sparse training for feature selection? In *NeurIPS*, 2022.

[Strubell *et al.*, 2020] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *AAAI*, 34, Apr. 2020.