AFRL-AFOSR-VA-TR-2023-0433

Cyber-resilient High-dimensional Data Analytics with Analytical Guarantees

Meng Wang
RENSSELAER POLYTECHNIC INST TROY NY
100 8TH STREET
TROY, NY,
US

08/22/2023
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20230822	Final	START DATE 20200601	END DATE 20230531

4. TITLE AND SUBTITLE
Cyber-resilient High-dimensional Data Analytics with Analytical Guarantees

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-20-1-0122	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Meng Wang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
RENSSELAER POLYTECHNIC INST TROY NY
100 8TH STREET
TROY, NY
US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203	AFRL/AFOSR RTA2	AFRL-AFOSR-VA-TR-2023-0433

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aims to extract useful information from large amounts of networked data obtained by the Air Force. It develops a framework of computationally efficient and cyber?resilient data acquisition, data recovery, and data classification methods from high?dimensional measurements. Specifically, this project develops algorithmic and theoretical foundation of data recovery from low?quality data using low?dimensional models, information extraction using neural networks with provable generalization guarantees, and computation reduction methods for neural networks while maintaining generalization.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
U	U	U	UU	33

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
DONALD WAGNER	000-0000

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

Final Report

Cyber-resilient High-dimensional Data Analytics with Analytical Guarantees
FA950-20-1-0122
Meng Wang, Associate Professor, Rensselaer Polytechnic Institute

August 17, 2023

1 Accomplishments

1.1 Research Objectives

This project aims to extract useful information from large amounts of networked data obtained by the Air Force. It develops a framework of computationally efficient and cyber-resilient data acquisition, data recovery, and data classification methods from high-dimensional measurements.

In years 1 and 2, this project has (i) developed data recovery methods from low-bit measurements, missing data, and errors using low-dimensional models, and (ii) developed a theoretical foundation of computationally efficient neural network learning with provable guarantees. In year 3, the main research objective is to develop efficient and reliable learners with theoretical guarantees in neural network learning, given the limited resources and training samples.

In this report, we mainly report the progress in year 3, while the progress reports for year 1 and year 2 are attached in the end for the record. In year 3, we develop the theoretical foundation of joint data-model sparsification and mixture-of-expert (MoE) in deep learning. Joint data-model sparsification focuses on sparsifying the data and neural network model simultaneously to reduce the computational cost. On the other hand, MoE executes training of different parts of the network (i.e. subnetworks) on different features in the data which allows the reduction of training compute and required training samples. In addition, we design the experiments on synthetic and real data to verify our theoretical findings.

1.2 Accomplishments

1.2.1 Major Activities

This project focuses on two major aspects in year 3:

Joint data-model sparsification. With data rapidly growing in size and deeper neural networks emerging, the training and inference of neural networks become increasingly expensive. The training and inference of neural networks present significant inefficiencies, creating obstacles for the scalability of deep learning-based AI systems in large-scale real-world applications. Therefore, various sparse learning techniques have been exploited to reduce memory and computational costs. Graph neural networks (GNNs) have demonstrated superior empirical performance in learning graph-structured data in applications such as object detection [31, 38], recommendation system [42, 47], rational learning [29], and machine translation [36, 37]. The approaches to accelerate GNN training can be categorized into two paradigms: (i) sparsifying the graph topology [6, 13, 25, 49], and (ii) sparsifying the network model [7, 43]. Sparsifying the graph topology means selecting a subgraph instead of the original graph to reduce the computation of neighborhood aggregation. One could either use a fixed subgraph (e.g., the graph typology [17], graph shift operator [1, 5], or the degree distribution [10, 20, 34] is preserved) or apply sampling algorithms, such as edge sparsification [13], or node sparsification [6, 49] to select a different subgraph in each iteration. Sparsifying the network model means reducing the complexity of the neural network model, including removing the non-linear activation [14, 35],

quantizing neuron weights [4, 33] and output of the intermediate layer [22], pruning network [12], or knowledge distillation [15, 18, 40, 41]. Both sparsification frameworks can be combined, such as joint edge sampling and network model pruning in [7, 43].

Despite many empirical successes in accelerating GNN training without sacrificing test accuracy, the theoretical evaluation of training GNNs with sparsification techniques remains largely unexplored. Most theoretical analyses are centered on the expressive power of sampled graphs [6,8,13,28,49] or pruned networks [9, 23, 46]. However, there is limited *generalization* analysis, i.e., whether the learned model performs well on testing data. Most existing generalization analyses are limited to two-layer cases, even for the simplest form of feed-forward neural networks (NNs), see, e.g., [16, 24, 32, 45] as examples. To the best of our knowledge, only [3, 21] go beyond two layers by considering three-layer GNNs and NNs, respectively. However, [21] requires a strong assumption, which cannot be justified empirically or theoretically, that the sampled graph indeed presents the mapping from data to labels. Moreover, [3, 21] focus on a linearized model around the initialization, and the learned weights only stay near the initialization [2]. The linearized model cannot justify the advantages of using multi-layer (G)NNs and network pruning. As far as we know, there is no finite-sample generalization analysis for joint sparsification, even for two-layer GNNs.

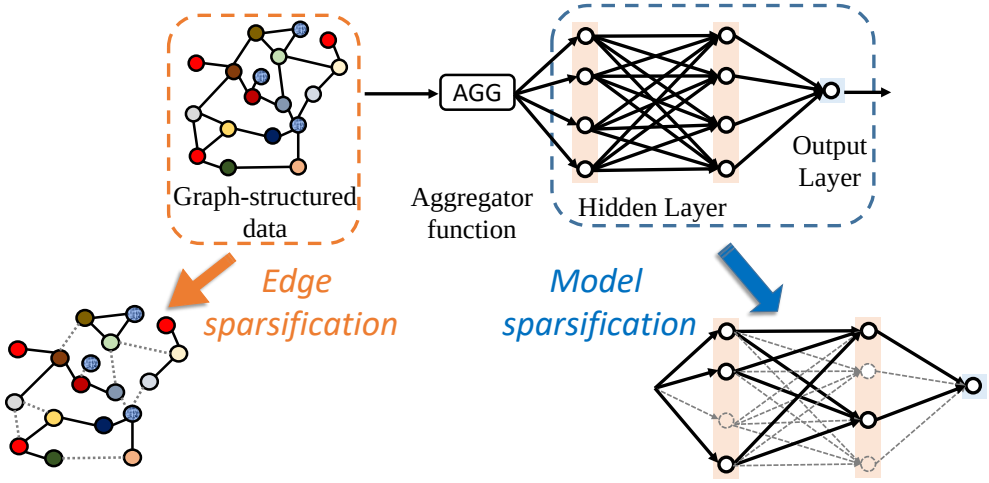


Figure 1: The illustration of our proposed joint data-model sparsification. Dash lines represent removed edges in the graph data and pruned neurons in the graph neural networks.

In this project, we explore a collaborative approach that combines data and model sparsification techniques, namely, joint data-model sparsification, within the context of learning graph-structured data. The proposed algorithm consists of two main components: (1) data sparsification, which involves sampling a subset of edges to form a subgraph for data sparsification, and (2) model sparsification, which entails pruning the trainable neurons within the neural network model. An illustration of our proposed algorithm can be found in Figure 1.

Mixture-of-experts. MoE activates different parts of the networks for different input and significantly reduces the training complexity without hurting the performance in many learning tasks [30, 39]. In a conventional MoE, in each layer, the neurons are divided into multiple groups referred to as experts and a learnable router routes each input sample to one or few of the experts [26]. However, in recent MoE architectures [11, 19, 27, 30, 48], the input samples are divided into patches and the routing decisions are made on patches as shown in Figure 2. MoE with patch-level routing (i.e. pMoE) showing unparallel empirical success. For example, it can reduce 20% training compute and 50% inference compute compared to dense architecture in vision task [27].

Despite the empirical success of pMoE, there is no theoretical underpinning behind the success. More specifically, there is no theoretical guarantee of the generalizability of pMoE with reduced training compute. Moreover, theoretical quantification of computational efficiency is necessary for future designs of pMoE. Furthermore, data efficiency of pMoE and implementation with Convolutional Neural Network (CNN) is

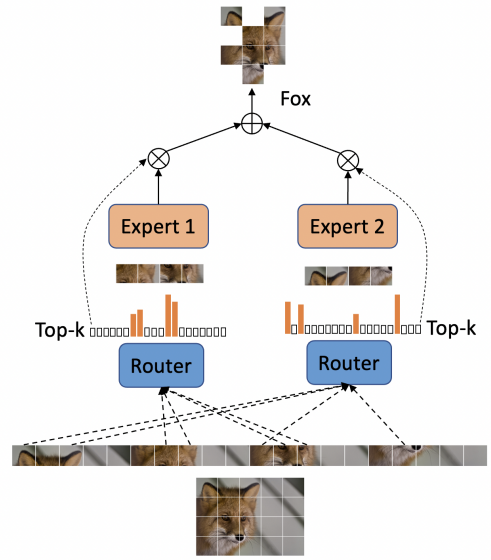


Figure 2: An illustration of pMoE. The image is divided into 20 patches while the router selects 4 of them for each expert.

unexplored.

In this project, we explored (both theoretically and empirically) the potential effectiveness of pMoE in CNN-based architecture, not only from the perspective of computational efficiency but also from the perspective of data efficiency.

1.2.2 Specific Objectives

The specific objectives of this project in year 3 include:

1. Investigation of the efficacy of the joint data-model sparsification approach in accelerating the convergence rate, reducing computational time, and enhancing the generalization performance.
2. Characterization of the gradient dynamics and the learning capacity of neurons when training graph neural networks to understand their ability to learn diverse data features.
3. Theoretical characterization of data and model sparsification are win-win strategies, effectively reducing the required number of training samples and accelerating the convergence rate.
4. Design of numerical experiments on synthetic and real data to verify the theoretical findings.
5. Theoretical investigation of the effectiveness of pMoE with CNN-based architecture for sample, time, and parameter efficiency.
6. Characterization of the respective roles of the router and experts in feature learning of a pMoE layer which facilitates the very success.
7. Experimental demonstration of the effectiveness of pMoE in deep CNN models.

1.2.3 Significant Results

Joint data-model sparsification. This project introduces a groundbreaking contribution by providing the first-ever theoretical analysis of joint topology-model sparsification in training Graph Neural Networks (GNNs). The analysis encompasses multiple crucial aspects, including: (1) The project establishes explicit bounds on the sample complexity, which refers to the required number of training samples, as well as the convergence rate of stochastic gradient descent (SGD). These bounds ensure the return of a highly accurate

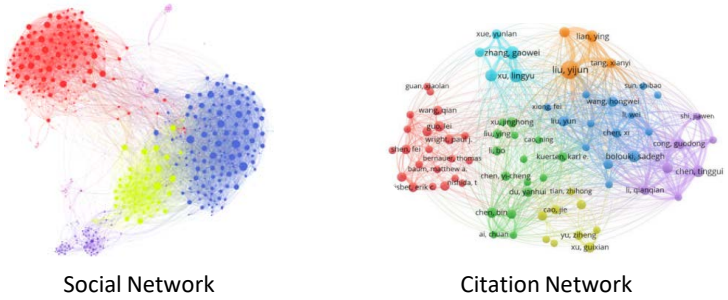


Figure 3: Visualizations of a social network and citation network.

model that successfully predicts unknown labels. (2) A quantitative proof is presented, demonstrating that joint data and model sparsification serve as a mutually beneficial strategy, leading to enhanced learning performance from both the sample complexity and convergence rate perspectives. This proof highlights the significant advantages gained by adopting this approach. (3) Furthermore, the project provides numerical validation of the joint data-model sparsification algorithm on various graph datasets, ranging from small-scale (Cora, Citeseer, and PubMed) to large-scale (Ogbn-ArXiv and Ogbn-Proteins) graphs. Remarkably, these numerical justifications showcase a remarkable reduction in computational complexity, ranging from a mere 5% to 20% when compared to training on dense data and models. This demonstrates the efficiency and effectiveness of the proposed algorithm. In summary, this project not only contributes theoretical advancements by establishing bounds and proofs but also provides concrete evidence through numerical evaluations, confirming the efficacy of joint data-model sparsification in GNN training.

The theoretical analysis is built upon two important assumptions observed from the graph datasets, namely, (1) nodes connected to each other tend to have the same label, and (2) some nodes have a stronger influence than the other nodes. The theoretical analysis is grounded on two crucial observations derived from the graph datasets, which form essential assumptions. The first assumption is that nodes that are connected to each other tend to possess the same label. This assumption acknowledges the tendency for connected nodes to share similar characteristics or attributes, thereby impacting their labels in a consistent manner. The second assumption recognizes that certain nodes exert a stronger influence compared to others within the graph. This assumption highlights the presence of influential nodes whose attributes or connections significantly impact the overall behavior and outcomes of the graph. See Figure 3 for an illustration.

Let r represent the number of sampled neighbor nodes, α denote the sampling rate of important nodes, β represent the pruning rate of neurons in the neural network model, and L denote the dimension of the data features. The following informal summary provides the context for the theorems. Theorem 1 concludes the sample complexity (**C1**) and convergence rate (**C2**) of our proposed algorithm in learning graph-structured data via graph sparsification. Specifically, the returned model achieves zero generalization error (from (3)) with enough samples (**C1**) after enough number of iterations (**C2**).

Theorem 1 (Informal version of Theorem 1 in [44]) *When the following conditions hold:*

*(**C1**) the number of training samples $|\mathcal{D}|$ is sufficiently large that it satisfies*

$$|\mathcal{D}| \geq \alpha^{-2} \cdot (1 + r^2) \cdot (1 - \beta)^2 \cdot L^2 \cdot \log q, \tag{1}$$

*(**C2**) the number of iterations T is sufficiently large that it satisfies*

$$T \geq (1 + |\mathcal{D}|^{-1/2}) \cdot (1 - \beta) \cdot \alpha^{-1} \cdot L, \tag{2}$$

the model returned $g(\hat{\theta})$ by our proposed joint data-model algorithm achieves zero generalization error. Namely, for any data with input $\boldsymbol{x}$ and label y in the graph datasets, we have

$$\mathbb{E}_{\boldsymbol{x},y} \|y - g(\hat{\theta}; \boldsymbol{x})\|_2 = 0. \tag{3}$$

Specifically, the highlights are summarized in the following aspects.

1. We provide the first convergence analysis with generalization guarantees for the joint data-model algorithms employed in training graph neural networks. This analysis marks a significant step forward in the understanding and assurance of the performance and effectiveness of these algorithms, enhancing the reliability and applicability of graph neural networks in practical settings.

2. We thoroughly analyze the neurons' capacity to learn various data features. Notably, we discover that neurons with small magnitudes tend to learn irrelevant features, such as background and noise. In contrast, neurons with large magnitudes tend to learn data features that align with important nodes. This finding serves as compelling evidence supporting the effectiveness of magnitude-based pruning. By removing neurons with small magnitudes, we can craft a highly proficient learning model. (See in Proposition 3 [44] and Figure 8 for details.).

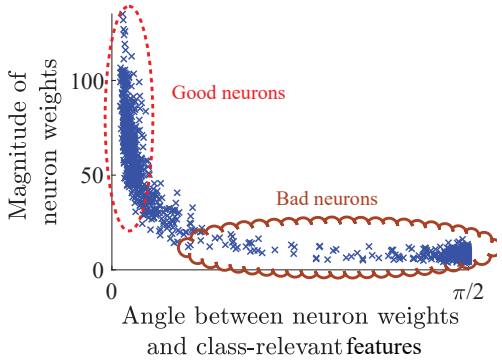


Figure 4: Distributions of Neuron Weights after Convergence. Neurons with small angles to class-relevant features and large magnitudes are considered 'Good,' indicating their proficiency in capturing crucial information for accurate classification. Conversely, 'Bad' neurons exhibit large angles to class-relevant features with small magnitudes, signifying their limited contribution to the classification task and association with less significant information.

3. Edge sparsification and magnitude-based model pruning synergistically enhance the learning performance by effectively reducing the required number of training samples and accelerating the convergence rate, resulting in a win-win strategy for GNN training. Our theorem offers a crucial theoretical validation for the success of joint edge-model sparsification. In particular, the sample complexity and the number of iterations exhibit quadratic and linear dependencies on $\frac{1-\beta}{\alpha}$, respectively. This implies that both techniques can be applied significantly to improve learning performance, demonstrating their combined effectiveness in training GNNs.

4. Numerical validation on a real graph dataset showcases the remarkable efficacy of our joint data-model sparsification, resulting in a significant boost in test accuracy when compared to random pruning. Additionally, this approach yields substantial computational cost savings compared to training on the dense model and data. These findings underscore the practical advantages and promising potential of our sparsification technique in enhancing both performance and efficiency in graph-based learning tasks.

Mixture-of-experts. Our work shows the theoretical guarantee of sample efficiency of pMoE for CNN compared to its dense counterpart which has been verified for standard deep CNN architecture. The major insights can be summarized as follows:

1. We provide the first theoretical generalization analysis for pMoE (with expert-choice routing). The analysis reveals that pMoE can reduce both sample complexity and computational complexity by a polynomial factor compared to the conventional CNN model. Our work also guarantees the polynomial reduction of model complexity.

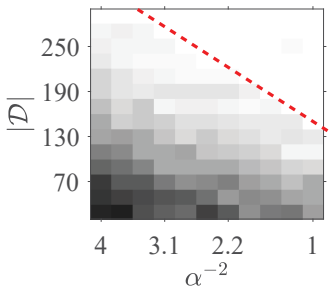


Figure 5: The number of training samples $|\mathcal{D}|$ against the importance sampling probability α

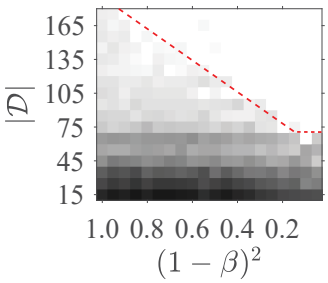


Figure 6: The number of training samples $|\mathcal{D}|$ against the pruning rate β

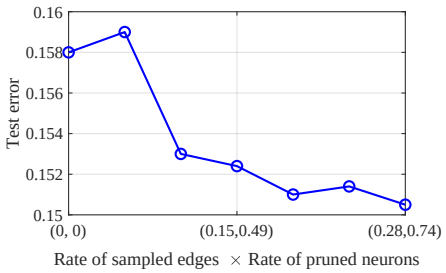


Figure 7: The test error on the Obgn-Protein dataset with different edge sparsity and model sparsity.

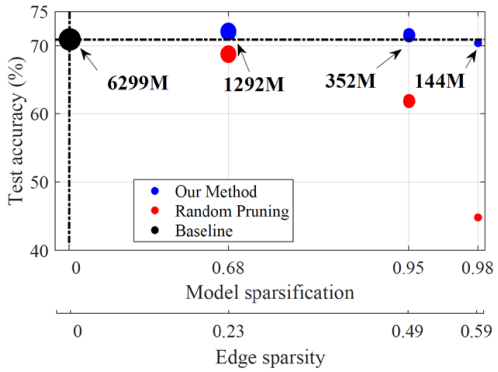


Figure 8: Effect of Joint Data-Model Sparsification on Citeseer Node Classification. Markers' size and number represent computational complexity during inference time at various levels of data and model sparsification.

2. We characterize the key property of the router behind the success of pMoE. It shows that a desired pMoE router effectively learns to dispatch similar important patches of input to the same experts and drop unimportant patches.

3. Furthermore, we provide the empirical demonstration of sample and computational efficiency of pMoE in deep CNN architecture such as WideResNet (WRN) on several benchmark vision datasets.

For the theoretical analysis, we consider a supervised binary classification problem. Given N i.i.d. training samples $\{(x_i, y_i)\}_{i=1}^N$ generated by an unknown distribution $\mathcal{D}$, the goal is to learn a neural network model that can map x to y for any (x, y) sampled from $\mathcal{D}$. Each input $x \in \mathbb{R}^{nd}$ is divided into n disjoint patches, where $y \in \{+1, -1\}$ denotes the corresponding label. For the analysis, we consider a pMoE architecture consisting of k experts (two-layer CNN) and corresponding k routers. Each of the routers selects l patches out of n patches from an input. An illustration of the analyzed architecture is given in Figure 9.

The learning problem solves the following empirical risk minimization problem with the logistic loss function,

$$\min_{\theta}: \quad L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i f_M(\theta, x_i)}\right) \tag{4}$$

Here, $f_M(\theta, x)$ denotes the analyzed pMoE model with θ representing all the trainable weights. We consider both *separate-training* and *joint-training* of the routers and experts.

Our work shows that to achieve ϵ generalization error:

- The *separate-training* pMoE requires $\Omega\left(l^8/\epsilon^{16}\right)$ training samples and $\Omega\left(l^{10}/\epsilon^{16}\right)$ neurons with $k = 2$. The number of iterations required for convergence is $\Omega\left(l^4/\epsilon^8\right)$.

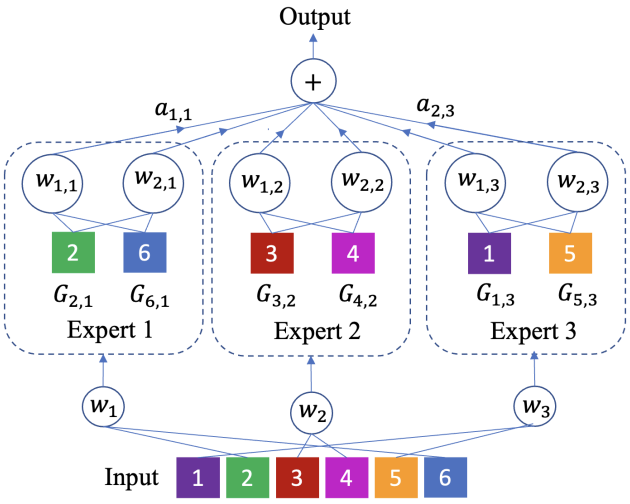


Figure 9: An illustration of the analyzed pMoE-CNN model

- The *joint-training* pMoE requires $\Omega\left(k^4l^6/\epsilon^{16}\right)$ training samples and $\Omega\left(k^3n^2l^6/\epsilon^{16}\right)$ neurons. The number of iterations required for convergence is $\Omega\left(k^2l^2/\epsilon^8\right)$.
- The conventional CNN requires $\Omega\left(n^8/\epsilon^{16}\right)$ training samples and $\Omega\left(n^{10}/\epsilon^{16}\right)$ neurons. The number of iterations required for convergence is $\Omega\left(n^4/\epsilon^8\right)$.

Therefore, the sample complexity is reduced by $O(n^8/l^8)$ and $O(n^8/k^4l^6)$ in *separate-training* and *joint-training* pMoE, respectively, compared to conventional CNN. Furthermore, our work shows that the computational complexity is reduced by $O(n^5/l^5)$ and $O(n^5/k^2l^3)$ in *separate-training* and *joint-training* pMoE, respectively compared to conventional CNN.

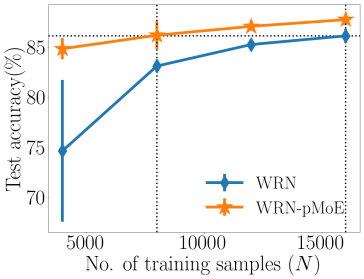


Figure 10: Classification accuracy of WRN-pMoE and WRN on multi-class classification in CelebA

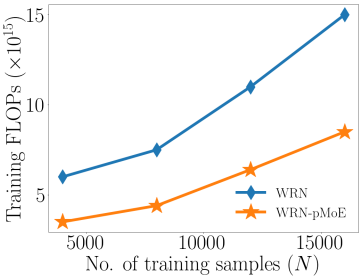


Figure 11: Training compute of WRN-pMoE and WRN on multi-class classification in CelebA

Besides verifying the theoretical findings empirically, we employ a 10-layer WideResNet (WRN) and the corresponding WRN-pMoE architecture on learning several benchmark vision datasets (e.g., CIFAR-10, CelebA) to demonstrate the effectiveness of pMoE in deep CNN model. Figure 10 and 11 compare the learning performance of WRN-pMoE with WRN in terms of test accuracy and training compute, respectively on "smiling" and "eyeglass" attribute detection of the CelebA dataset. The results suggest the significant advantage of pMoE over CNN both in terms of sample efficiency and training compute.

1.2.4 Publications

In total, this project has lead to 15 conference and journal publications. Among them, we have four journal papers, mainly in IEEE Transactions, as well as seven conference papers in top machine learning and AI conferences, i.e., Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (**ICML**), International Conference on Learning Representations (**ICLR**).

In year 3, we have the following published paper with full citations below. Among them, [4] is a journal publication, and [1], [2], [3], [5] are conference papers presented at ICML'22, ICLR'23, and ICML'23.

[1] Hongkang Li, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. "Generalization guarantee of training graph convolutional networks with graph topology sampling." In *International Conference on Machine Learning (**ICML**)*, pp. 13014-13051. PMLR, 2022.

[2] Shuai Zhang, Meng Wang, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Miao Liu. "Joint Edge-Model Sparse Learning is Provably Efficient for Graph Neural Networks." In The *Eleventh International Conference on Learning Representations (**ICLR**)*, 2023.

[3] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. "A Theoretical Understanding of Shallow Vision Transformers: Learning, Generalization, and Sample Complexity." In The *Eleventh International Conference on Learning Representations (**ICLR**)*, 2023.

[4] Ming Yi, Meng Wang, Tianqi Hong, and Dongbo Zhao. "Bayesian High-Rank Hankel Matrix Completion for Nonlinear Synchrophasor Data Recovery." *IEEE Transactions on Power Systems*, 2023.

[5] Nowaz Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. "Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convolutional Neural Networks." In *International Conference on Machine Learning (**ICML**)*, 2023. (Oral presentation).

1.3 Communities of Interest

This project leads to four papers in the top machine learning conferences (ICLR'22, ICLR'23, and ICML'23). Records of the presentations on these results are available to the participants in the virtual meetings hosted by ICLR and ICML. [5] is selected as an oral presentation in ICML. Moreover, we participated in the poster sessions to illustrate our results.

1.4 Future goals

After completing the project, we plan to investigate the following in the future,

1. Exploring structured sparsification algorithms with theoretical guarantees as hardware-friendly approaches.
2. Exploring emerging efficient network architectures in large language models (LLMs).

2 Impacts

2.1 Principal disciplines of the project

The major principal disciplines of this project include:

- **Theoretical Advancements in Generalization Guaranteed Deep Learning:** This research makes a significant contribution to the theoretical understanding of generalization guarantees in deep learning. By establishing robust generalization guarantees, it instills greater confidence in incorporating artificial intelligence (AI) technology across critical domains like self-driving cars and security systems. Enhanced generalization ensures that AI models can perform reliably and accurately on unseen data, bolstering their trustworthiness and real-world applicability.
- **Theoretical Foundation of Efficient Deep Learning**: Another significant contribution of this research lies in advancing the theoretical foundation of joint data and model sparsification approaches. By developing innovative learning methods capable of identifying well-compressed models and employing efficient data sampling techniques, substantial savings in computational resources are achieved. This enhanced efficiency holds profound benefits for various applications, particularly on mobile devices

and edge computing, where faster inference and reduced energy consumption are critical for delivering seamless user experiences. The theoretical advancements achieved through this work pave the way for more resource-efficient and scalable solutions, enabling the widespread application of AI technologies across diverse domains.

2.2 Impact on the development of human resources

This project supports two Ph.D. students and one graduated and became a postdoc at RPI, continuing on this project.

This project provides opportunities for underrepresented groups of students in engineering to engage in AI-related topics. For instance, this project involves one female undergraduate student in the Electrical, Computer, and System Engineering department at Rensselaer Polytechnic Institute. She was able to access the computation resources in the lab to test her developed algorithms.

2.3 Impact on teaching and educational experiences

The codes for numerical evaluations of the methods in this project are publicly available.

2.4 Impact on physical, institutional, and information resources

Not applicable.

2.5 Impact on society beyond science and technology

This project has delivered two significant impacts on society: a deeper understanding of the decision-making process of AI systems and the enhancement of their overall efficiency.

Through achieving complete explainability and transparency in AI, we equip ourselves to develop AI systems that are not only controllable but also align with legal regulations. Furthermore, this transparency aids governments in formulating effective policies and laws to promote and regulate AI, fostering increased public acceptance of these transformative technologies. Moreover, the widespread implementation of AI has the potential to revolutionize workplaces, vastly improving overall efficiency while simultaneously augmenting human capabilities. AI excels in handling repetitive or hazardous tasks, thereby freeing up the human workforce to concentrate on tasks that require creativity and empathy. As a result, this strategic reallocation of responsibilities can lead to increased overall happiness and job satisfaction among employees, as they are empowered to engage in work that aligns with their innate strengths and abilities.

Enhanced efficiency in AI offers a wide array of benefits, encompassing cost savings, increased productivity, and improved resource allocation. These gains serve as catalysts for economic growth, fostering innovation, attracting investments, and creating fresh job opportunities in the realm of AI-related fields. Consequently, this dynamic and competitive economy drives progress and prosperity for society as a whole. Furthermore, AI's efficiency plays an important role in promoting environmental sustainability. By optimizing energy consumption and minimizing waste during AI training, it actively contributes to mitigating climate change and safeguarding precious natural resources for the well-being of future generations. This responsible use of AI aligns with the broader goals of environmental conservation and fosters a sustainable future.

In summary, this project's pursuit of understanding AI decision-making and enhancing AI efficiency holds tremendous potential for positive societal transformation. It not only promotes responsible AI deployment but also sparks advancements in economic prosperity, environmental preservation, and the well-being of the workforce.

3 Changes

Not applicable.

References

[1] B. Adhikari, Y. Zhang, S. E. Amiri, A. Bharadwaj, and B. A. Prakash, "Propagation-based temporal network summarization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 4, pp. 729–742, 2017.

[2] Z. Allen-Zhu and Y. Li, "Feature purification: How adversarial training performs robust deep learning," in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2022, pp. 977–988.

[3] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," in *Advances in neural information processing systems*, 2019, pp. 6158–6169.

[4] M. Bahri, G. Bahl, and S. Zafeiriou, "Binary graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9492–9501.

[5] A. Chakeri, H. Farhidzadeh, and L. O. Hall, "Spectral sparsification in spectral clustering," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2301–2306.

[6] J. Chen, T. Ma, and C. Xiao, "Fastgcn: Fast learning with graph convolutional networks via importance sampling," in *International Conference on Learning Representations*, 2018.

[7] T. Chen, Y. Sui, X. Chen, A. Zhang, and Z. Wang, "A unified lottery ticket hypothesis for graph neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1695–1706.

[8] W. Cong, M. Ramezani, and M. Mahdavi, "On provable benefits of depth in training graph convolutional networks," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=r-oRRT-ElX

[9] A. da Cunha, E. Natale, and L. Viennot, "Proving the strong lottery ticket hypothesis for convolutional neural networks," in *International Conference on Learning Representations*, 2022.

[10] T. Eden, S. Jain, A. Pinar, D. Ron, and C. Seshadhri, "Provable and practical approximations for the degree distribution using sublinear graph samples," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 449–458.

[11] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[12] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJl-b3RcF7

[13] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[14] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 639–648.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *stat*, vol. 1050, p. 9, 2015.

[16] B. Huang, X. Li, Z. Song, and X. Yang, "Fl-ntk: A neural tangent kernel-based framework for federated learning analysis," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4423–4434.

[17] C. Hübler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani, "Metropolis algorithms for representative subgraph sampling," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 283–292.

[18] A. K. Jaiswal, H. Ma, T. Chen, Y. Ding, and Z. Wang, "Spending your winning lottery better after drawing it," *arXiv preprint arXiv:2101.03255*, 2021.

[19] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations*, 2020.

[20] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 631–636.

[21] H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Generalization guarantee of training graph convolutional networks with graph topology sampling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 014–13 051.

[22] Z. Liu, K. Zhou, F. Yang, L. Li, R. Chen, and X. Hu, "Exact: Scalable graph neural networks training via extreme activation compression," in *International Conference on Learning Representations*, 2021.

[23] E. Malach, G. Yehudai, S. Shalev-Shwartz, and O. Shamir, "Proving the lottery ticket hypothesis: Pruning is all you need," *arXiv preprint arXiv:2002.00585*, 2020.

[24] S. Oymak and M. Soltanolkotabi, "Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 84–105, 2020.

[25] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710.

[26] P. Ramachandran and Q. V. Le, "Diversity and depth in per-example routing models," in *International Conference on Learning Representations*, 2018.

[27] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.

[28] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," in *International Conference on Learning Representations*, 2019.

[29] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proceedings of European Semantic Web Conference*. Springer, 2018, pp. 593–607.

[30] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.

[31] W. Shi and R. Rajkumar, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711–1719.

[32] Z. Shi, J. Wei, and Y. Liang, "A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features," in *International Conference on Learning Representations*, 2022.

[33] S. A. Tailor, J. Fernandez-Marques, and N. D. Lane, "Degree-quant: Quantization-aware training for graph neural networks," in *International Conference on Learning Representations*, 2020.

[34] E. Voudigari, N. Salamanos, T. Papageorgiou, and E. J. Yannakoudakis, "Rank degree: An efficient algorithm for graph sampling," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2016, pp. 120–129.

[35] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International Conference on Machine Learning*, 2019, pp. 6861–6871.

[36] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[37] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[38] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI*, 2018.

[39] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[40] Y. Yang, J. Qiu, M. Song, D. Tao, and X. Wang, "Distilling knowledge from graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7074–7083.

[41] H. Yao, C. Zhang, Y. Wei, M. Jiang, S. Wang, J. Huang, N. Chawla, and Z. Li, "Graph few-shot learning via knowledge transfer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6656–6663.

[42] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 974–983.

[43] H. You, Z. Lu, Z. Zhou, Y. Fu, and Y. Lin, "Early-bird gcns: Graph-network co-optimization towards more efficient gcn training and inference via drawing early-bird lottery tickets," *AAAI Conference on Artificial Intelligence*, 2022.

[44] S. Zhang, M. Wang, P.-Y. Chen, S. Liu, S. Lu, and M. Liu, "Joint edge-model sparse learning is provably efficient for graph neural networks," *The Eleventh International Conference on Learning Representations*, 2023.

[45] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Fast learning of graph neural networks with guaranteed generalizability:one-hidden-layer case," in *2020 International Conference on Machine Learning (ICML)*, 2020.

[46] Z. Zhang, J. Jin, Z. Zhang, Y. Zhou, X. Zhao, J. Ren, J. Liu, L. Wu, R. Jin, and D. Dou, "Validating the lottery ticket hypothesis with inertial manifold theory," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[47] L. Zheng, Z. Zuo, W. Wang, C. Dong, M. Momma, and Y. Sun, "Heterogeneous graph neural networks with neighbor-sim attention mechanism for substitute product recommendation," *AAAI Conference on Artificial Intelligence*, 2021.

[48] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. M. Dai, Z. Chen, Q. V. Le, and J. Laudon, "Mixture-of-experts with expert choice routing," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=jdJo1HIVinI

[49] D. Zou, Z. Hu, Y. Wang, S. Jiang, Y. Sun, and Q. Gu, "Layer-dependent importance sampling for training deep and large graph convolutional networks," *Proceedings of Advances in Neural Information Processing Systems*, vol. 32, 2019.

Year 1 Project Report

Cyber-resilient High-dimensional Data Analytics with Analytical Guarantees

FA950-20-1-0122

Meng Wang, Associate Professor, Rensselaer Polytechnic Institute

May 12, 2021

This project aims to extract useful information from large amounts of networked data obtained by the Air Force. It will develop a framework of computationally efficient and cyber-resilient data acquisition, data recovery, and data classification methods from high-dimensional measurements.

The team has been focusing on the following two aspects in the first year.

1 Data recovery from quantized, errorniens and partial measurements by exploiting the low-rank tensor property

1.1 Motivation and major contributions

Maintaining high data quality is critically important to ensure the reliability of extracted information. The received measurements by the operator often contain missing data and data errors due to communication congestions or device errors, especially under the extreme military environment. A cyber intruder may delay or drop data packages to the operator or even fabricate malicious measurements to mislead the operator. Moreover, due to sensor issues or communication restrictions, images and videos may have very low resolution [24]. Quantization is also intentionally applied to enhance the data privacy in sensor networks [7, 10, 22]. Because all the sensor measurements are highly quantized, a cyber intruder with access to a few sensors cannot extract accurate information from the data. On the other hand, the system operator has data from all sensors and can leverage the correlations in the data to extract information collectively from large amounts of quantized data. The objective of this research thrust is to recover the actual measurements from all these data issues to enhance the accuracy of the subsequent information extraction tasks using these data.

The low-rank property exists for many practical datasets. For instance, the series measurements of multiple sensors in a sensor network can be represented by a low-rank matrix (the matrix rank is much less than its ambient dimension). A video can be represented by a matrix with each column representing a vectorized frame. Though matrix techniques like low-rank matrix completion and low-rank matrix recovery have been widely used to recover missing data and correct bad data, some data contain intrinsic correlations that cannot be simply captured by matrices. For example, users might give different ratings to the same object under different contexts [3]. A vector cannot well characterize the spatial correlations of objects in a video frame. In contrast, higher-order tensors have the capacity to capture the additional correlations, and are leveraged to improve the performances on recovery/completion tasks [19, 27, 28].

Existing works on low-rank tensor recovery mainly consider random noise or sparse noise [4, 23, 30], while only a few works [1, 12, 17] consider tensor recovery from one-bit measurements, i.e., all measurements are binary. Moreover, convex formulations are employed in [1, 12] and lead to a larger recovery error theoretically and empirically. In our papers [25, 26], we first study the recovery under the nonconvex formulation directly. In addition, we consider the more general multi-bit scenario.

Our contributions

• We for the first time solve the quantized tensor recovery problem with exact low-rank constraint in the multi-bit quantization framework.

• We consider both general tensors and a special tensor category named SVD-tensors. The theoretical recovery error bounds for both cases and the fundamental limitation are provided. We also provide the bounds of the fundamental limit of the recovery error by any method and show that our results are almost optimal.

• We propose a generalized quantized tensor recovery scheme when the quantization bin boundaries are unknown.

• With the aid of recent progress on proximal methods, we propose efficient algorithms that are proved to converge to critical points of a nonconvex optimization problem at a rate that is at least sublinear.
• Our method outperforms the existing works on synthetic data, real image data, and recommender data.

1.2 Problem formulation and theoretical results

$\mathcal{X}^*$ and $\mathcal{N} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_W}$ denote the W-order tensors containing actual data and noise data, respectively. $\mathcal{X}^*$ is a low-rank tensor and the values are bounded, i.e., $\mathcal{X}^* \leq r$ and $\|\mathcal{X}^*\|_\infty \leq \alpha$, where $r, \alpha > 0$. The values in noise tensor $\mathcal{N}$ are independent and identically distributed (i.i.d.) from a known cumulative distribution function Φ. Consider the quantization operator M that maps real values to K discrete values $[K]$ given the intervals formed by boundary pairs $(\omega_0^*, \omega_1^*]; (\omega_1^*, \omega_2^*]; \ldots (\omega_{K-1}^*, \omega_K^*]$. Formally,

$$\begin{aligned} &\mathcal{Q}_{i_1,i_2,\ldots,i_W} = M(\mathcal{X}^*_{i_1,i_2,\ldots,i_W} + \mathcal{N}_{i_1,i_2,\ldots,i_W}) = l \\ &\text{if } \omega^*_{l-1} < \mathcal{X}_{i_1,i_2,\ldots,i_W} + \mathcal{N}_{i_1,i_2,\ldots,i_W} \leq \omega^*_l,\ l \in [K], \end{aligned} \tag{1}$$

where $\omega_0^* = -\infty$ and $\omega_K^* = \infty$. Figure 1 provides a visualization of this quantization operation on a 3-order tensor and $K = 3$. $\mathcal{Q}$ is the quantized matrix. Some measurements in $\mathcal{Q}_\Omega$ are lost. Let Ω be the set of indices of observed entries.

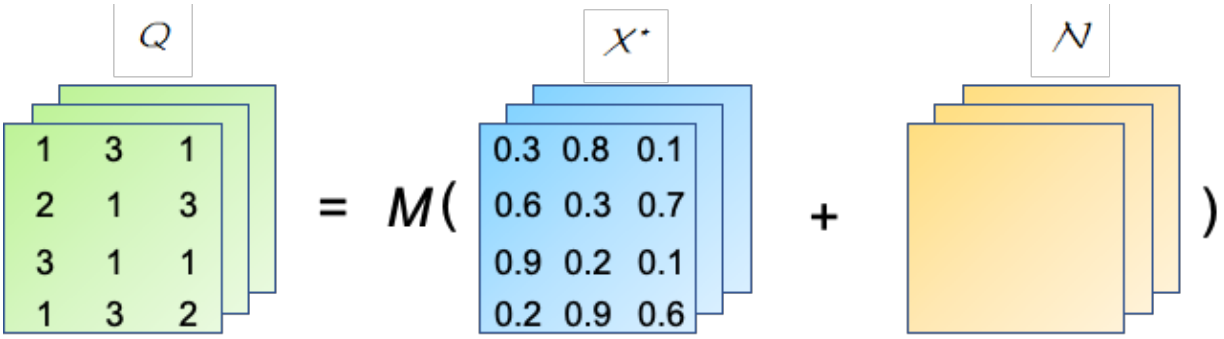


Fig. 1: Visualization of the quantization process on a 3-order tensor ($K = 3$).

The questions to address are : (1) How can we estimate the ground truth $\mathcal{X}^*$ given $\mathcal{Q}_\Omega$ and Φ? (P2): Is it possible to estimate $\mathcal{X}^*$ when the quantization rule is unknown?

We propose to estimate $\mathcal{X}^*$ by solving the following nonconvex optimization problem

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} F_\Omega(\mathcal{X}) \quad \text{s.t. } \mathcal{X} \in \mathcal{S}_f, \tag{2}$$

and the feasible set $\mathcal{S}_f$ contains the rank constraint and the infinity norm constraint.

$$\mathcal{S}_f := \{\mathcal{X} : \|\mathcal{X}\|_\infty \leq \alpha, \text{rank}(\mathcal{X}) \leq r\}. \tag{3}$$

Theorem 1. *If there is no missing data, with high probability, any global minimizer $\hat{\mathcal{X}}$ of (2) satisfies*

$$\|\hat{\mathcal{X}} - \mathcal{X}^*\|_F / \sqrt{n_1 n_2 \ldots n_W} \leq C_1 \left(\sqrt{\frac{r^{W-1} K \log W}{n^{W-1}}}\right) \tag{4}$$

when $n_1, n_2, \ldots, n_W$ are all in the order of n. If the tensor is a SVD-tensor, then the recovery error is reduced to

$$\|\hat{\mathcal{X}} - \mathcal{X}^*\|_F / \sqrt{n_1 n_2 \ldots n_W} \leq C_2 \left(\sqrt{\frac{rW \log W}{n^{W-1}}}\right) \tag{5}$$

We also show the fundamental limit of the recovery error by any method.

Theorem 2. *For any recovery method that always exists a rank-r tensor $\mathcal{X}^*$ such that the recovery error is at least*

$$\|\hat{\mathcal{X}} - \mathcal{X}^*\|_F / \sqrt{n_1 n_2 ... n_W} \geq C_3(\sqrt{\frac{r}{n^{W-1}}}), \tag{6}$$

where $\hat{\mathcal{X}}$ is the estimation of $\mathcal{X}^*$ by any recovery method.

Comparing Theorems 1 and 2, one can see that the recovery error by solving (2) is close to the minimum value that can be achieved by any recovery method.

We develop a fast algorithm named Tensor-based Alternating Proximal Gradient Descent (TAPGD) to solve the nonconvex problem (2) with the convergence guarantee.

Theorem 3. *The iterates returned by TAPGD converge to a critical point of* (2)*. The convergence rate is at least sublinear, and more specifically, in the order of $t^{\frac{\theta-1}{2\theta-1}}$, $\theta \in (\frac{1}{2}, 1)$, where t is the iteration number.*

1.3 Numerical evaluation

We test our method on the Extend Yale Face Dataset B [11, 16]. The dataset contains 192×168 pixel face images from 38 different people. Each person has 64 images with different poses and various illumination. We pick two objects to form a $192 \times 168 \times 128$ three-dimensional tensor. All entries are scaled to $[0, 1]$. We add $\mathcal{N}$ with i.i.d. entries generated from the Gaussian distribution with mean 0 and the standard deviation of 0.3. When $W = 2$, $\omega_0^* = -\infty$, $\omega_1^* = 0.4$, $\omega_2^* = \infty$. When $W = 3$, $\omega_0^* = -\infty$, $\omega_1^* = 0.2$, $\omega_2^* = 0.4$, $\omega_3^* = \infty$. Fig. 2 (a) compares TAPGD with MNC-1bit-TR, the quantized matrix recovery method, and a nonconvex low-rank tensor recovery method named Nonconvex Regularized Tensor (NORT) [28].Note that MNC-1bit-TR models the quantization process like our approach, while NORT does not model quantization and treats the data as general noisy measurements. We find that our method works well in a wide range of r, and the results are under the selection of $r = 50$. The tolerance rate is set as 0.001 for matrix recovery method. In the NORT method, we set the hyperparameters as $\lambda = 0.1, \theta = 5$ (The parameters have different meanings from the parameters in our work), and the tolerance rate as 0.0001. In the MNC-1bit-TR method, we use $r^{\frac{3}{2}}\alpha$ (here $\alpha = 1$) to bound the maximum row norm of the low-rank factors. It shows that the relative recovery error decreases when the percentage of the observation increases, and TAPGD obtains the best performance among all the methods. Fig. 2 (b) compares the recovery error when the bin boundaries are known and unknown to the recovery algorithm. When the boundaries are unknown, the initial point is uniformly chosen from $[0.1, 0.6]$ for ω_1 when $W = 2$, and $[0.1, 0.3], [0.2, 0.6]$ for ω_1, ω_2, respectively when $W = 3$. $\alpha_{\text{upper}}, \alpha_{\text{low}}$ are selected as $0.6, 0.1$. κ_l is set to 0.1 for $\forall l \in [W - 1]$.

In Fig. 3, we show a box-plot-diagram of relative recovery error with 100 runs obtained by TAPGD. All the setups are the same as the scenario $W = 3$ in Fig. 2 (a). The tops and bottoms of each "box" are the 25th and 75th percentiles of the samples, respectively. The maximum standard deviation happens when the observation rate is 0.3, which equals to 8.79×10^{-4}. The relative standard deviation, which is defined as the ratio of the standard deviation to the mean, reaches its maximum value 0.028 when the observation rate is 0.6.

Fig. 4 compares the time cost of TAPGD and MNC-1bit-TR [12] when the number of facial images changes. TAPGD is three magnitudes faster than MNC-1bit-TR. Fig. 5 visualizes the quantized and recovered images by TAPGD.

2 Fast learning of neural networks with provable generalizability

2.1 Background and motivation

Neural networks, especially convolutional neural networks (CNNs), have demonstrated superior performance in applications like image and video processing. Learning a neural network needs to find appropriate

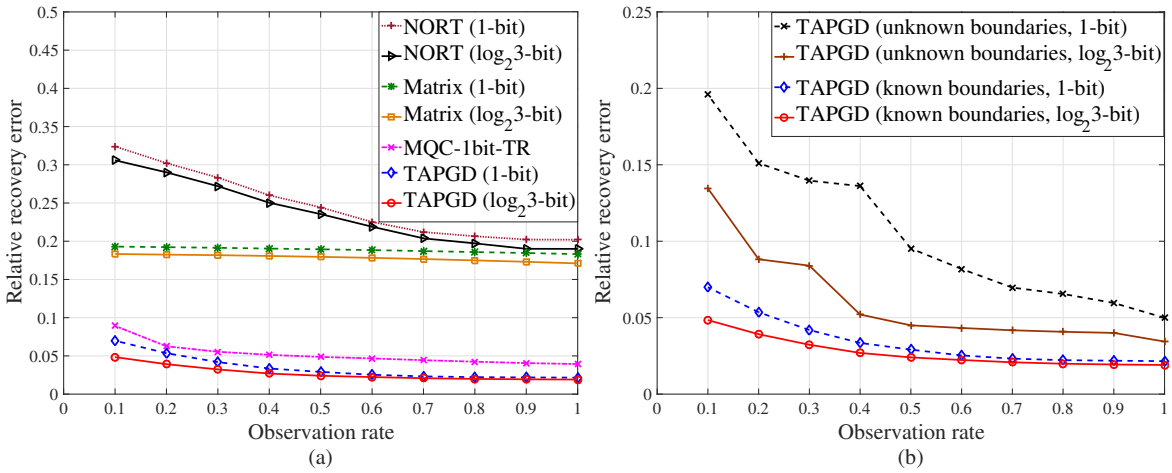


Fig. 2: (a) Relative recovery error when the observation rate changes (b) Relative recovery error of unknown boundaries.

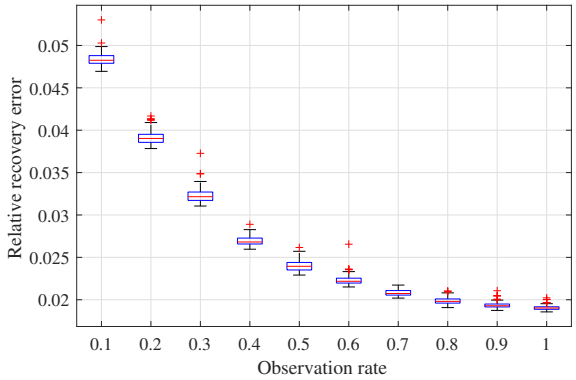


Fig. 3: Relative recovery error when the observation rate changes ($W = 3$, TAPGD).

parameters for the hidden layers using the training data and is achieved by minimizing a non-convex empirical loss function over the choices of the model parameters. The non-convex learning problem is usually solved by a first-order gradient descent (GD) algorithm. The convergence to the global optimal, however, is not guaranteed naturally due to the existence of spurious local minima. Another major hurdle to the widespread acceptance of deep learning is the lack of analytical performance guarantees about whether the parameters learned from the training data perform well on the testing data, i.e., the generalizability of the learned model. A learned model generalizes well to the testing data provided that it is a global minimizer of the population loss function, which takes the expectation over the distribution of testing samples. Since the distribution is unknown, one minimizes the empirical loss function of the training data assuming that the training data are drawn from the same distribution. Moreover, a large number of training samples are required to obtain a network model with powerful feature representation capability, while the method may perform poorly when the number of training samples is small. The theoretical characterization of the required size of the training data for a given network architecture is vastly unavailable.

One line of research employs the Mean Field approach to model the training process of neural networks by a differential equation [5, 20]. In this step, the network width needs to go to infinite, and the step size of stochastic gradient descent needs to be infinitesimal, both of which are not practical. Another line of research employs the neural tangent kernel (NTK) [15], which requires strong over-parameterization such

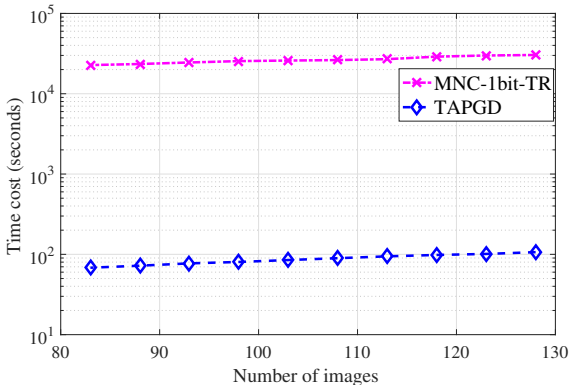


Fig. 4: Time cost of TAPGD and MNC-1bit-TR.

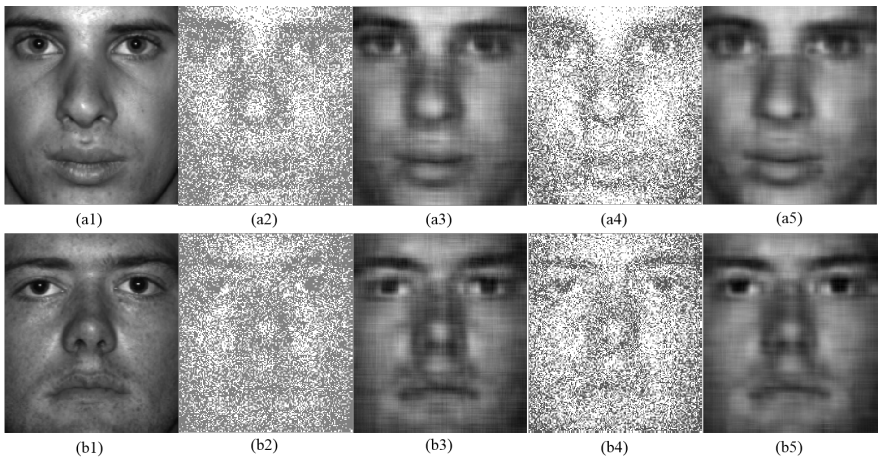


Fig. 5: (a1,b1) Original images (a2,b2) Quantized images ($W = 2$) (a3,b3) Recovered images ($W = 2$) (a4,b4) Quantized images ($W = 3$) (a5,b5) Recovered images ($W = 3$).

that the nonlinear neural network model behaves as its linearization around the initialization [2, 9, 31, 32]. However, it is unlikely that practical nerual entworks are operating in this so-called "lazy region" of strong parameterization [6].

We follow the line of works that study the generalization performance of neural networks in the "teacher-student" setup, where the training data are generated by a teacher neural network, and the learning is performed on a student network by minimizing the empirical risk of the training data. Assuming that the student network has the same architecture as the teacher network, the existing generalization analyses mostly focus on one-hidden-layer networks because the optimization problem is already nonconvex, and the analytical complexity increases tremendously when the number of hidden layers increases.

One critical assumption of most works in this line is that the input features follow the standard Gaussian distribution. Although other distributions are considered in [8,13,14,18,20,21,29], the generalization performance beyond the standard Gaussian input is less investigated. On the other hand, the learning performance clearly depends on the input data distribution.

2.2 Contributions

We provide a theoretical analysis of learning one-hidden-layer neural networks when the input distribution follows a Gaussian mixture model containing an arbitrary number of Gaussian distributions with arbitrary mean and variance. The Gaussian mixture model has been employed in many applications such as data clustering and unsupervised learning, image classification and segmentation, and few-shot learning.

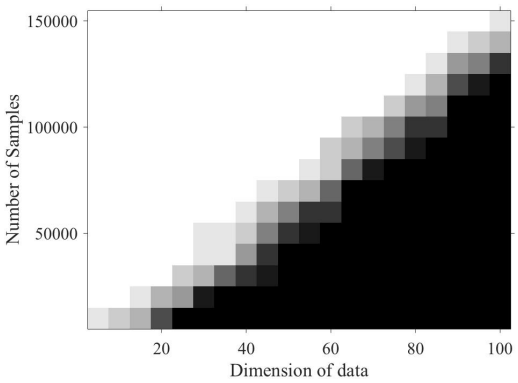


Fig. 6: The sample complexity against the feature dimension d

We propose a gradient descent algorithm with tensor initialization to estimate the weights of the one-hidden-layer fully-connected neural network. Our algorithm converges to a critical point linearly, and the returned critical point converges to the teacher model at a rate of $\sqrt{d \log n/n}$, where d is the dimension of the feature, and n is the number of samples. We also characterize the required number of samples for accurate estimation, referred to as the sample complexity, as a function of d, the number of neurons K, and the input distribution. To the best of our knowledge, *this is the first theoretical and explicit characterization about how the mean and variance of the input distribution affect the sample complexity and learning rate*.

The implications of our bounds include:

• When the absolute value of any mean in the Gaussian mixture model increases from zero, the sample complexity increases, and the algorithm converges slower, indicating that it will be more challenging to learn a model with a small test error.

• The same phenomenon happens when any variance in the mixture model increases to infinity from a certain positive value, or if all the variances in the mixture model approach zero.

• The training process converges faster and requires fewer samples if the input data are zero mean with a certain non-zero variance. This can be viewed as one theoretical explanation in one-hidden-layer for the success of Batch normalization.

2.3 Numerical validation

We evaluate our theoretical bounds of sample complexity and convergence rate on synthetic datasets. We vary d and the number of samples n. For each pair of d and n, 20 independent sets of $\mathbf{W}^*$ and the corresponding training samples are generated. Fig. 6 shows the success rate of these independent experiments. A black block means that all the experiments fail. A white block means that they all succeed. The sample complexity is indeed almost linear in d, as predicted by our bound.

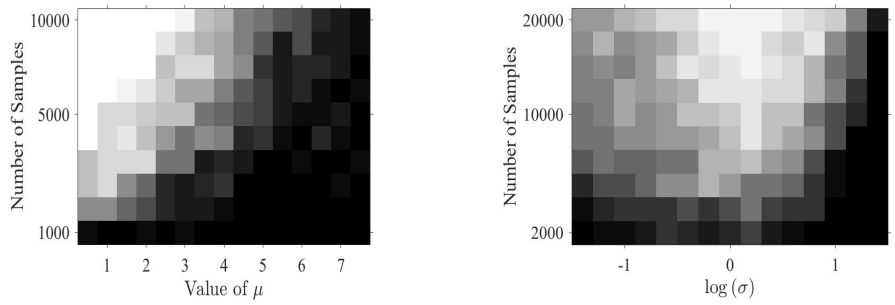


Fig. 7: The sample complexity (a) when one mean changes, (b) when one variance changes.

We then fix d and study the impact on the sample complexity when the mean and variance in the Gaussian mixture model change. Fig. 7(a) shows that when the mean increases, the sample complexity increases. This coincides with our theoretical analyses. In Fig. 7(b), the sample complexity increases both when σ increases and when σ approaches zero. These results match our theoretical predictions.

We next study the convergence rate of our gradient descent algorithm. d is fixed as 5. Fig. 8(a) shows the impact of the mean. One can see that the algorithm always converges linearly when $\tilde{\mu}$ changes. Moreover, as $\tilde{\mu}$ increases, the algorithm converges slower, as predicted by our theoretical analyses. Fig. 8(b) shows the impact of the variance of the Gaussian mixture model. Among different variance we test, the algorithm converges fastest when the variance is 1. The convergence rate slows down when it increases to 2 or when it decreases to 0.5. The result is consistent with our theoretical results. We then verify the dependence of the convergence rate on the number of neurons K. One can see from Fig. 9 that, as predicted, the convergence rate is almost linear in $1/K^2$.

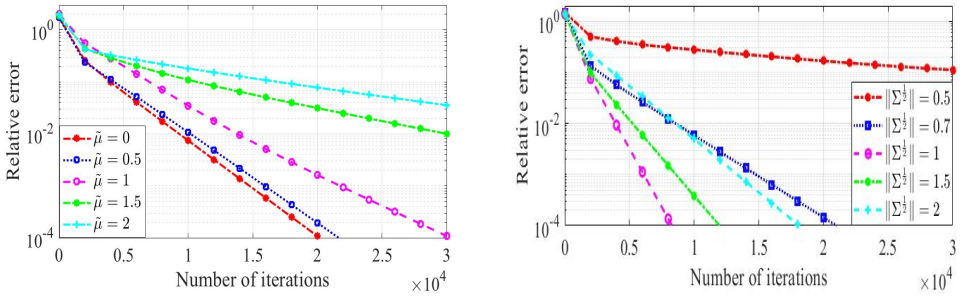


Fig. 8: (a) The convergence rate with different $\tilde{\mu}$ (b) The convergence rate with different σ.

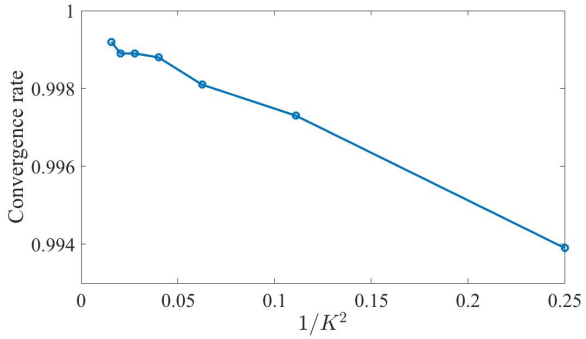


Fig. 9: Convergence rate when the number of neurons K changes

We then evaluate the distance between the model $\widehat{\mathbf{W}}_n$ returned by the algorithm and the ground-truth model weights $\mathbf{W}^*$, measured by $||\widehat{\mathbf{W}}_n - \mathbf{W}^*||_F$. The error in Fig. 10 is indeed linear in $\sqrt{\log(n)/n}$, as predicted by our theoretical bounds.

3 Next steps

In the second year, the team plans to continue the exploration of reliability of information extraction using neural networks. Specifically, we will focus on simplifying neural network architectures to enhance the learning performance and reduce the required number of training samples. This is extremely important for applications where the data collection is costly.

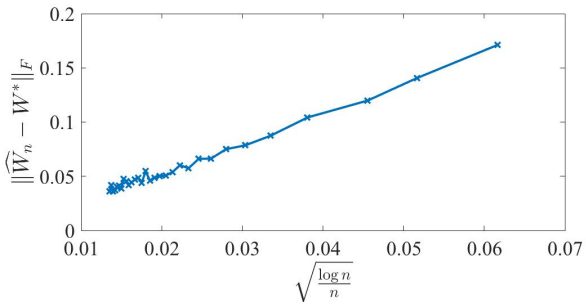


Fig. 10: The relative error of the learned model with the teacher model when n changes

Year 1 Publications

We have two published paper with fully citations as follows. We are about to submit another two manuscripts to IEEE Transactions.

- Ren Wang, Meng Wang, and Jinjun Xiong. Quantized higher-order tensor recovery by exploring low-dimensional structures. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, November 2020.
- Ren Wang, Meng Wang, and Jinjun Xiong. Tensor recovery from noisy and multi-level quantized measurements. *EURASIP Journal on Advances in Signal Processing*, 41, 2020. URL: `https://doi.org/10.1186/s13634-020-00698-z`.

References

[1] Anastasia Aidini, Grigorios Tsagkatakis, and Panagiotis Tsakalides. 1-bit tensor completion. *Electronic Imaging*, 2018(13):261–1, 2018.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[3] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl-Heinz Lüke, and Roland Schwaiger. Incarmusic: Context-aware music recommendations in a car. In *International Conference on Electronic Commerce and Web Technologies*, pages 89–100. Springer, 2011.

[4] Shouyuan Chen, Michael R Lyu, Irwin King, and Zenglin Xu. Exact and stable recovery of pairwise interaction tensors. In *Advances in Neural Information Processing Systems*, pages 1691–1699, 2013.

[5] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3040–3050, 2018.

[6] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

[7] Junil Choi, Jianhua Mo, and Robert W Heath. Near maximum-likelihood detector and channel estimator for uplink multiuser massive mimo systems with one-bit adcs. *IEEE Transactions on Communications*, 64(5):2005–2018, 2016.

[8] Simon S. Du, Jason D. Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint, http://arxiv.org/abs/1709.06129*, 2017.

[9] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL: `https://openreview.net/forum?id=S1eK3i09YQ`.

[10] Pengzhi Gao, Ren Wang, Meng Wang, and Joe H. Chow. Low-rank matrix recovery from noisy, quantized and erroneous measurements. *IEEE Trans. Signal Process.*, 66(11):2918–2932, 2018.

[11] Athinodoros S Georghiades, Belhumeur Peter N, and Kriegman David J. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[12] Navid Ghadermarzy, Yaniv Plan, and Ozgur Yilmaz. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2019.

[13] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *ArXiv preprint arXiv: 2006.13409*, 2020.

[14] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. *arXiv preprint arXiv: 1909.11500*, 2019.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[16] Kuang-Chih Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):684–698, 2005.

[17] Baohua Li, Xiaoning Zhang, Xiaoli Li, and Huchuan Lu. Tensor completion from one-bit observations. *IEEE Transactions on Image Processing*, 28(1):170–180, 2019.

[18] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[19] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2012.

[20] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[21] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Arxiv preprint Arxiv: 2006.06098*, 2020.

[22] Andreas Reinhardt, Frank Englert, and Delphine Christin. Enhancing user privacy by preprocessing distributed smart meter data. In *Proc. Sustainable Internet and ICT for Sustainability (SustainIT)*, pages 1–7, 2013.

[23] Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[24] Ren Wang, Meng Wang, and Jinjun Xiong. Data recovery and subspace clustering from quantized and corrupted measurements. *IEEE J. Sel. Topics Signal Process., Special Issue on Robust Subspace Learning and Tracking: Theory, Algorithms, and Applications*, 12(6):1547–1560, 2018.

[25] Ren Wang, Meng Wang, and Jinjun Xiong. Quantized higher-order tensor recovery by exploring low-dimensional structures. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, November 2020.

[26] Ren Wang, Meng Wang, and Jinjun Xiong. Tensor recovery from noisy and multi-level quantized measurements. *EURASIP Journal on Advances in Signal Processing*, 41, 2020. URL: `https://doi.org/10.1186/s13634-020-00698-z`.

[27] Yangyang Xu, Ruru Hao, Wotao Yin, and Zhixun Su. Parallel matrix factorization for low-rank tensor completion. *arXiv preprint arXiv:1312.1254*, 2013.

[28] Quanming Yao, James Tin-Yau Kwok, and Bo Han. Efficient nonconvex regularized tensor completion with structure-aware proximal iterations. In *International Conference on Machine Learning*, pages 7035–7044, 2019.

[29] Yuki Yoshida and Masato Okada. Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1722–1730. Curran Associates, Inc., 2019.

[30] Xiaoqin Zhang, Di Wang, Zhengyuan Zhou, and Yi Ma. Simultaneous rectification and alignment via robust recovery of low-rank tensors. In *Advances in Neural Information Processing Systems*, pages 1637–1645, 2013.

[31] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

[32] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2055–2064, 2019.

Technical Report

May 30, 2022

1 Accomplishments

1.1 Research Objectives

This project aims to extract useful information from large amounts of networked data obtained by the Air Force. It will develop a framework of computationally efficient and cyber-resilient data acquisition, data recovery, and data classification methods from high-dimensional measurements.

In year 2, the main research objective is to develop efficient and reliable learners with theoretical guarantees in neural network learning, given the limited resources and training samples.

To achieve the goals above, we develop the theoretical foundation of neural network pruning and self-training. Neural network pruning sets some neuron weights to zero to reduce the training time. Self-training uses both a limited number of labeled data (costly to obtain) and a large number of unlabeled data (cheap to obtain) for training. In addition, we design the experiments on synthetic and real data to verify our theoretical findings.

1.2 Accomplishments

1.2.1 Major Activities

This project focuses on two major aspects:

Network pruning. The success of modern deep learning mainly benefits from building overparameterized neural networks, where the number of trainable parameters is significantly higher than the degree of the models. However, larger models require more storage space and take more time to run, which requires more expensive hardware and leads to longer inference time.

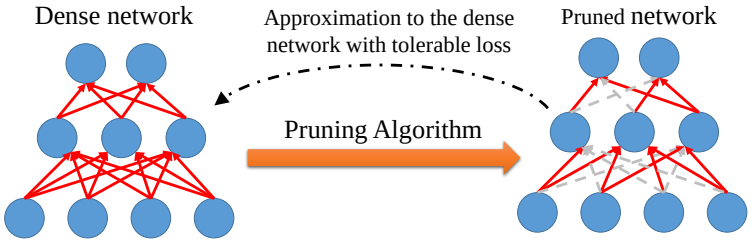


Figure 1: An example of network pruning techniques; dash lines stand for pruned weights, and solid lines stand for unpruned weights.

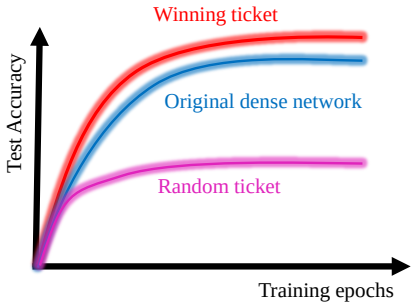


Figure 2: Illustration of "winning tickets" and "random tickets".

The essential idea of network pruning techniques is to eliminate unnecessary weights in the hidden layers, and numerical experiments [5, 9, 10, 13, 22] suggest that over 90% of the parameters can be removed without harming the test accuracy. Therefore, neural network pruning approach can reduce the computational cost of model training and inference significantly [18, 26, 30] and potentially lessen the chance of overfitting [6].

The recent lottery ticket hypothesis (LTH) claims that a randomly initialized dense neural network always contains a so-called "winning ticket," which is a sub-network bundled with the corresponding initialization. This winning ticket can achieve at least the same testing accuracy as that of the original network by running at most the same amount of training time, while a randomly pruned network usually achieves worse performance than the original network (see Figure 2). Inspired by LTH, network pruning algorithms [7, 23] succeed in deeper networks like Residual Networks (Resnet)-50 and Bidirectional Encoder Representations from Transformers (BERT) network [3], where a matched sub-network with sparsity varying from 40% to 90% achieves high test accuracy than the original dense network.

Despite the numerical success of LTH and its derived pruning algorithms, the theoretical foundation of network pruning is limited. The existing theoretical works are mainly from the scope of model compression, i.e., finding a sub-network that achieves a tolerable loss in either expressive power or training accuracy, compared with the original dense network [1, 21, 32]. None of them can provide theoretical support for the improved generalization achieved by winning tickets, i.e., pruned networks with faster convergence and better test accuracy. This project systematically analyzes learning pruned neural networks with a finite number of training samples. Our analytical results also justify the LTH from the perspective of the sample complexity.

Self-training. Self-training [8, 19, 25, 31], one of the most powerful semi-supervised learning (SemiSL) algorithms, augments a limited number of labeled data with unlabeled data so as to achieve improved generalization performance on test data, compared with the model trained by supervised learning using the labeled data only (see Figure 3). In practice, the quality of training data can hardly be guaranteed given a fixed data collection budget or limited time. Also, some labels are not accessible because they may contain sensitive and private information. Although labeled data are often costly to obtain, unlabeled data are usually vastly available. Therefore, self-training has shown empirical success in diversified applications such as few-shot image classification [4, 27–29, 35], objective detection [24], robustness-aware model training against adversarial attacks [2], continual lifelong learning [20], and natural language processing [11, 14].

The terminology "self-training" has been used to describe various SemiSL algorithms in the literature, while this project is centered on the commonly-used iterative self-training method. An initial teacher model (learned from the labeled data) is applied to the unlabeled data to generate pseudo labels. One then trains a student model by minimizing the weighted empirical risk of both the labeled and unlabeled data. The student model is then used as the new teacher to update the pseudo labels of the unlabeled data. This process is repeated multiple times to improve the eventual student model. See Figure 4 for an algorithm flowchart.

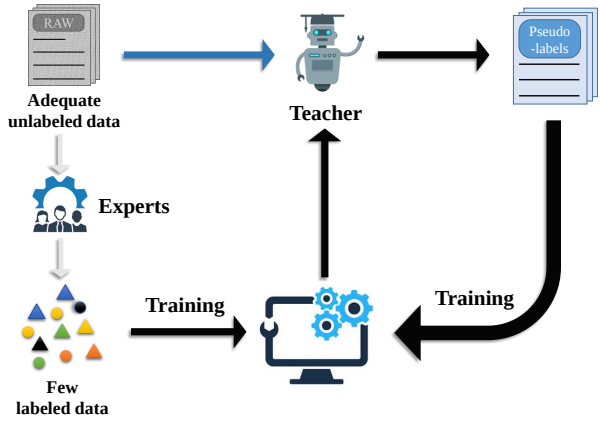


Figure 3: Semi-supervised learning methods by using unlabeled data.

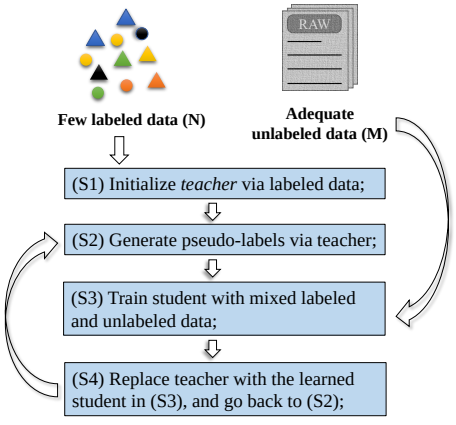


Figure 4: The proposed self-training algorithm in this project.

Despite the empirical achievement of self-training methods with neural networks, the most theoretical justification for such success is limited to linear models. However, the neural networks used in practice are highly nonlinear due to their activation functions and deep layers. To the best of our knowledge, there is no analytical characterization of how the unlabeled data affects the learned model's generalization by iterative

self-training on nonlinear neural networks. Hence, this project provides the theoretical study of iterative self-training on nonlinear neural networks. Specifically, we provide a quantitative analysis of the generalization performance of iterative self-training as a function of the number of labeled and unlabeled samples.

1.2.2 Specific Objectives

The specific objectives of this project include:

1. Exploration of the Iterative Magnitude Pruning algorithm;
2. Exploration of the Iterative Self-training algorithm;
3. Characterization of the landscape of the objective function and identification of the local convex region near the initial point and ground truth;
4. Design of high-order momentum to obtain a good initialization that lies in the characterized local convex region through tensor decomposition;
5. Characterization of the influence of sample amount in affecting the landscape of objective function through concentration theorem;
6. Instructions on the parameters selections in improving network pruning algorithm and self-training algorithm;
7. Design of numerical experiments to verify the theoretical findings.

1.2.3 Significant Results

Network pruning. For network pruning, this project provides the theoretical justification for the improved generalization of a good pruned network. Let $\widetilde{r}$ be the number of non-pruned weights and N be training samples, the informal version of the theorems are summarized below.

Theorem 1 indicates that there is a locally convex region near the ground truth $\boldsymbol{W}^*$, and the radius of the region is affected by $\widetilde{r}$ and N.

Theorem 1 (Informal version of Theorem 1 in [33]) *Let M be the mask matrix of a "winning ticket". Suppose the mapping between the input and output can be approximated by a neural network with some ground truth $\boldsymbol{W}^*$ and bounded noise. Then, the objective function is strictly locally convex at any $\boldsymbol{W}$ such that*

$$\|M \odot \boldsymbol{W} - \boldsymbol{W}^*\|_2 \leq \Theta\Big(\sqrt{\frac{\widetilde{r}}{N}} \cdot \frac{1}{K^2}\Big), \tag{1}$$

where $\odot$ denotes the entry-wise multiplication and K is the number of neurons in the hidden layer.

Theorem 2 shows that the iterations converge linearly to the ground truth $\boldsymbol{W}^*$ with a rate of $1-(1-\sqrt{\widetilde{r}/N}/\sqrt{K})$ up to an statistical error $\sqrt{\widetilde{r}/N}|\xi|$ with noisy labels.

Theorem 2 (Informal version of Theorem 2 in [33]) *Same setting as Theorem 1. When the number of training samples satisfies*

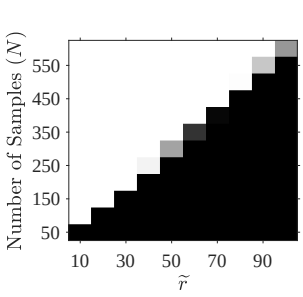
$$N \geq \Theta(K^6 \widetilde{r}), \tag{2}$$

the iterates $\{\boldsymbol{W}^{(t)}\}_{t=1}^T$ converges linearly to $\boldsymbol{W}^*$ as

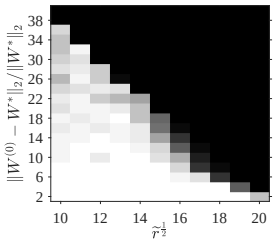
$$\|\boldsymbol{W}^{(T)} - \boldsymbol{W}^*\|_2 \leq \left(1 - \frac{1-\sqrt{\widetilde{r}/N}}{\sqrt{K}}\right)^T \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2 + \sqrt{\frac{\widetilde{r}\log q}{N}} \cdot |\xi|, \tag{3}$$

where ξ is upper bound of the additive noise in the labels.

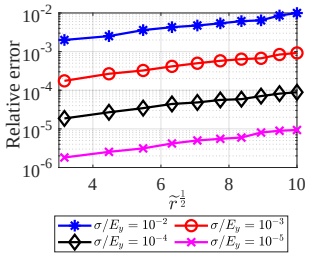
Specifically, our contribution can be summarized in the following aspects.



(a) Sample complexity



(b) The radius of the local convex region; $W^{(0)}$ is the initial point, and W^* is the desired point



(c) Relative error against $\widetilde{r}^{\frac{1}{2}}$ at different noise level

Figure 5: Phase transition of sample complexity and radius of local convex region when the number of the non-pruned weights $\widetilde{r}$ changes by averaging over 100 trials. A black block means the failures of learning in all trails while a white block indicates all success.

1. We provide the first required number of samples for successful convergence, termed the *sample complexity*, for the pruned networks. Our sample complexity bound depends linearly on the number of the non-pruned weights $\widetilde{r}$ (see (2)).

2. We characterize the benign optimization landscape of a pruned network, and the radius of the benign region increases as the non-pruned weights amount increases at the rate of $\sqrt{\widetilde{r}}$ (see (1)). Therefore, we show analytically the objective function has an enlarged convex region for a pruned network, justifying the importance of a good sub-network (i.e., the pruning network obtained through Iterative Magnitude Pruning algorithm).

3. We characterize the improved generalization of a good pruned network. We show that gradient-descent methods converge faster to the oracle model when the neural network is properly pruned, or equivalently, learning on a pruned network returns a model closer to the oracle model with the same number of iterations, indicating the improved generalization of winning tickets (see (3)).

4. We provide numerical experiments to justify the theoretical findings, see Figure 5. Figure 5 shows the numerical results on synthetic data with Gaussian input by averaging over 100 independent trials. The dark color indicates a lower success rate in Figures 5(a)&(b). We can see the curves between white and black regions are straight lines in both Figure 5(a) and (b). The curve in Figure 5(a) verifies our theoretical findings that the sample complexity is a linear function of the amount of the non-pruned weights. In Figure 5(b), the trial is successful if and only if the initial point $W^{(0)}$ lies in the local convex region of W^*, and the curve verifies our theoretical findings that the radius of the benign region is a linear function of $\sqrt{\widetilde{r}}$.

5. We justify the efficiency of network pruning techniques on practical data (Cifar10 [15] and Minst [16]). Figures 6 and 7 show the test performance of learned models by implementing the IMP algorithm on MNIST and CIFAR-10 using Lenet-5 [17] and Resnet-50 [12] architecture, respectively. As we can see, a properly pruned network (i.e., winning ticket) helps reduce the sample complexity required to reach the test accuracy of the original dense model. For example, training on a pruned network returns a model (e.g., P_1 and P_3 in Figures 6 and 7) that has better testing performance than a dense model (e.g., P_2 and P_4 in Figures 6 and 7) trained on a larger data set. Also, the test accuracy drops when the network is overly pruned, which is no longer a "winning ticket".

Self-training. For self-training methods, this project provides a quantitative analysis of the generalization performance of iterative self-training algorithm as the functions of labeled data and unlabeled data amounts. Let $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ be the labeled data and $\{(\widetilde{\boldsymbol{x}}_m, \widetilde{y}_m)\}_{m=1}^M$ be unlabeled data, where $\widetilde{y}_m$ is the pseudo-label.

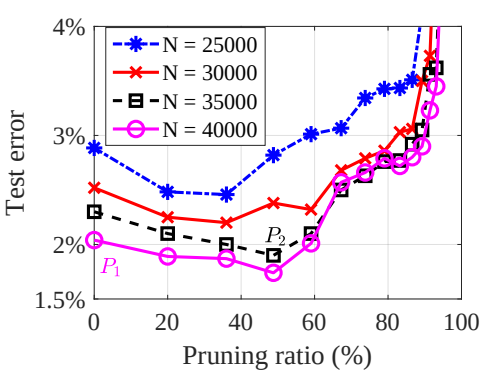


Figure 6: Test accuracy of pruned LeNet-5 on Mnist

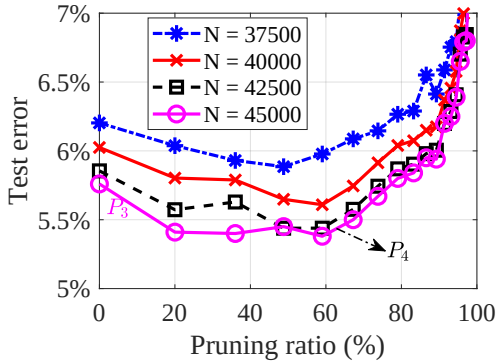


Figure 7: Test accuracy of pruned Resnet-50 on Cifar-10

The objective function of self-training approach is to minimize:

$$\hat{f}(\boldsymbol{W}) = \frac{\lambda}{2N}\sum_{n=1}^{N}\big(y_n - g(\boldsymbol{W};\boldsymbol{x}_n)\big)^2 + \frac{1-\lambda}{2M}\sum_{m=1}^{M}\big(\widetilde{y}_m - g(\boldsymbol{W};\widetilde{x}_m)\big)^2, \tag{4}$$

where λ is the a fixed constant between 0 and 1.

Theorem 1 characterizes the convergence rate of the proposed algorithm and the accuracy of the learned model $\boldsymbol{W}^{(L)}$ in a low labeled-data regime. Specifically, the iterates converge linearly, and the learned model is close to $\boldsymbol{W}^{[\hat{\lambda}]}$ and guaranteed to outperform the initial model $\boldsymbol{W}^{(0)}$.

Theorem 3 (Informal version of Theorem 1 in [34]) *If the following conditions hold:*

$$1/2 \le \lambda \le \sqrt{N/N^*} \quad \textit{and} \quad M \ge \Theta((1-\lambda)^2 K^3 d \log q). \tag{5}$$

Then, the iterations $\{\boldsymbol{W}^{(\ell)}\}_{\ell=0}^{L}$ *converges to* $\boldsymbol{W}^{[\lambda]} = (1-\lambda)\boldsymbol{W}^{(0)} + \lambda \boldsymbol{W}^*$ *as*

$$\begin{aligned}\|\boldsymbol{W}^{(L)} - \boldsymbol{W}^{[\lambda]}\|_F \le &\left(\Big(1+\Theta\big(\frac{1}{\sqrt{M}}\big)\Big)\cdot\frac{1}{K}\right)^{L}\cdot\|\boldsymbol{W}^{(0)} - \boldsymbol{W}^{[\lambda]}\|_2 \\ &+ \Big(1+\Theta\big(\frac{1}{\sqrt{M}}\big)\Big)\cdot\frac{1}{K}\cdot\|\boldsymbol{W}^* - \boldsymbol{W}^{[\lambda]}\|_F.\end{aligned} \tag{6}$$

Theoretical findings can be illustrated by Figure 8. In Figure 8, $\boldsymbol{W}^*$ is the ground truth model, and $\{\boldsymbol{W}^{(\ell)}\}_{\ell=0}^{L}$ stands for the iterations returned by the self-training algorithm at iteration ℓ. Also, $\boldsymbol{W}^{(0)}$ is the initialization, and $\boldsymbol{W}^{(L)}$ stands for the convergent point. The generalization error can be measured by the distance of $\boldsymbol{W}^{(L)}$ and $\boldsymbol{W}^*$. Specifically, we summarize the highlights as follows.

1. We provide the analytical justification of the iterative self-training algorithm using unlabeled data over a supervised learning approach. We proved that the learned model returned by the iterative self-training method converges linearly to a model close to the ground truth.

2. We provide the quantitative justification of improvement by unlabeled data. The distance between the convergent point and the ground truth is reduced as a linear function of $1/\sqrt{M}$(see ε_1 in Figure 8), where M is the unlabeled data amount. In addition, we prove that the improved convergence rate is a linear function of $1/\sqrt{M}$.

3. We provide the insights for the parameter selection in iterative self-training algorithms. Let λ be the weighted sum factor of the loss function for the labeled data. We prove that a large λ is desired because large λ requires a less number of unlabeled data with an improved generalization (a smaller ε_0 in Figure 8). In addition, our theorem shows an upper bound of λ to avoid divergence. A large labeled data amount and good initialization lead to a high upper bound of λ.

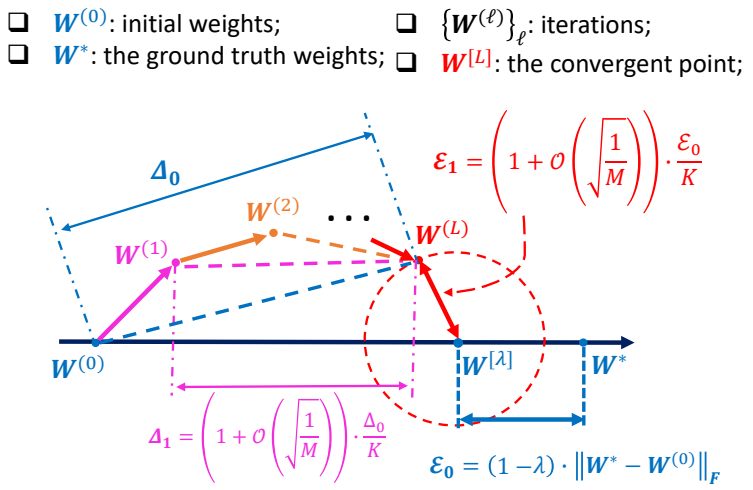


Figure 8: Illustration of the (1) ground truth $\boldsymbol{W}^*$, (2) iterations $\{\boldsymbol{W}^{(\ell)}\}_{\ell=0}^{L}$, (3) convergent point $\boldsymbol{W}^{(L)}$, and (4) $\boldsymbol{W}^{[\hat{\lambda}]} = \hat{\lambda}\boldsymbol{W}^* + (1-\hat{\lambda})\boldsymbol{W}^{(0)}$.

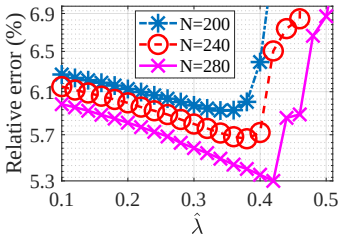


Figure 9: $\frac{\|\boldsymbol{W}^{(L)} - \boldsymbol{W}^*\|_F}{\|\boldsymbol{W}^*\|_F}$ when $\hat{\lambda}$ and N change.

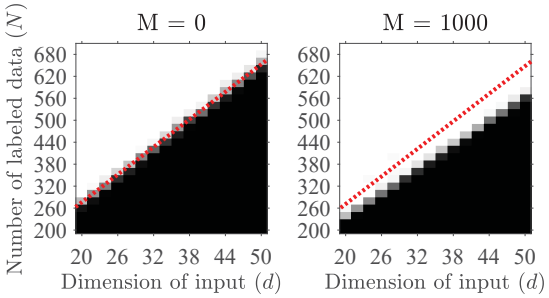
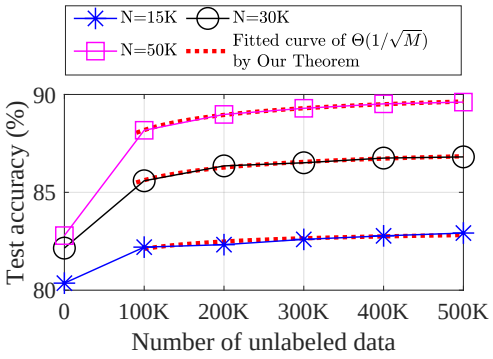


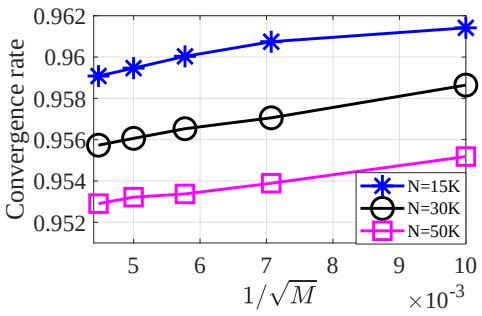
Figure 10: Empirical phase transition of the curves with (a) $M = 0$ and (b) $M = 1000$.

4. We provide the numerical experiments to justify the theoretical findings. Figure 9 shows that the relative errors of $\boldsymbol{W}^{(L)}$ with respect to $\boldsymbol{W}^*$ decrease almost linearly when $\hat{\lambda}$ increases, which is consistent with the theoretical result in (6). Moreover, when λ exceeds a certain threshold positively correlated with N, the relative error increases rather than decreases. That is consistent with the analysis in (5) that $\hat{\lambda}$ has an upper limit, and such a limit increases as N increases. Figure 10 depicts the phase transition of returning $\boldsymbol{W}^{(L)}$. When d increases, the required number of labeled data to learn $\boldsymbol{W}^*$ is linear in d. Thus, the sample complexity bound in (5) is order-wise optimal for d. Moreover, the phase transition line when $M = 1000$ is below the one when $M = 0$. Therefore, with unlabeled data, the required sample complexity of N is reduced.

5. We justify the improved performance by using unlabeled data on practical data. Figure 2 shows the iterative self-training algorithm on the augmented CIFAR-10 dataset, which contains unlabeled data mined from 80 Million Tiny Images. N is the labeled data amount from the original CIFAR-10 dataset, and M is the unlabeled data from Tiny Images. In Figure 11(a), the dash lines stand for the best fitting of the linear functions of $1/\sqrt{M}$ via the least square method. One can see that the test accuracy is improved by up to 7% using unlabeled data, and the empirical evaluations match the theoretical predictions. Figure 11(b) shows the convergence rate calculated based on the first 50 epochs, and the convergence rate is almost a linear function of $1/\sqrt{M}$.



(a) Test accuracy



(b) Convergence rate

Figure 11: The performance of iterative self-training algorithm on CIFAR-10 dataset against the number of unlabeled data; N is the labeled data amount, and M is the unlabeled data amount

1.2.4 Publications

We have the following published paper with fully citations below. Among them, [1-3] are journal publications. [4-6] are conference papers and presented at ICLR22, NeurIPS21, and CISS 22 respectively.

[1] Ming Yi, Meng Wang, Evangelos Farantatos and Tapas Barik. "Bayesian Robust Hankel Matrix Completion with Uncertainty Modeling for Synchrophasor Data Recovery," *ACM SIGENERGY Energy Informatics Review*, 2022.

[2] Ming Yi and Meng Wang. "Bayesian Energy Disaggregation at Substations with Uncertainty Modeling," *IEEE Transactions on Power Systems*, 2022, 37(1): 764-775.

[3] Shuai Zhang, Meng Wang, Jinjun Xiong, Sijia Liu, Pin-Yu Chen, "Improved Linear Convergence of Training CNNs With Generalizability Guarantees: A One-Hidden-Layer Case," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 2(6): 2622-2635, doi: 10.1109/TNNLS.2020.3007399.

[4] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen and Jinjun Xiong, "How Does Unlabeled Data Improve Generalization in Self-training? A one-hidden-layer Theoretical Analysis," in *Proc. the Tenth International Conference on Learning Representations (ICLR)*, April 2022. (acceptance rate: 32.3%)

[5] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen and Jinjun Xiong, "Why Lottery Ticket Wins? A Theoretical Perspective of Sample Complexity on Sparse Neural Networks," in *Proc. of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, December 2021. (acceptance rate: 26%)

[6] Hongkang Li, Shuai Zhang, and Meng Wang, "Learning and generalization of one-hidden-layer neural networks, going beyond standard Gaussian data," in *Proc. 2022 56th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2022.

1.3 Communities of Interest

This project leads to two papers in the top machine learning conferences (NeurIPS'21 and ICLR'22). Records of the presentations on these results are available to the participants in the virtual meetings host by NeurIPS and ICLR, and we participate the poster sessions to illustrate our results.

1.4 Future goals

Our plan to do during the next reporting period includes:

1. Fast learning algorithms on structured data, i.e., graph neural networks and graph structured data;
2. Fairness of the minority groups in machine learning problems.

2 Impacts

2.1 Principal disciplines of the project

The major principal disciplines of this project includes:

1. This project contributes to the theoretical foundation of the generalization guarantee in deep learning. It increases the public trust in incorporating artificial intelligence (AI) technology in critical domains, such as self-driving cars and security systems.

2. This project contributes to the theoretical foundation of network pruning approaches. It encourages learning methods to find a good compressed model, which can significantly save the computation resources. For instance, mobile devices can benefit from faster inference and lower energy costs.

3. This project contributes to the theoretical understanding of the knowledge transformation in deep learning. It reduces the required labeled data amount and improves the efficiency of deep learning development and deployment for multiple tasks. It benefits the AI system in leveraging stored knowledge to solve new challenges. It also guarantees the system to utilize simulated training to prepare for real-world tasks.

2.2 Impact on the development of human resources

This project supports two Ph.D. students and one graduated and became a postdoc at RPI, continuing on this project.

This project provides opportunities for underrepresented groups of students in engineering to engage in AI-related topics. For instance, this project involves one female undergraduate student in the Electrical, Computer, and System Engineering department at Rensselaer Polytechnic Institute. She was able to access the computation resources in the lab to test her developed algorithms.

2.3 Impact on teaching and educational experiences

The codes for numerical evaluations of the methods in this project are publicly available.

2.4 Impact on physical, institutional, and information resources

Not applicable.

2.5 Impact on society beyond science and technology

This project focuses on explaining the behaviors of AI systems in decision-making. If the AI system is fully explainable and transparent, the developed AI system will be controllable and can be restricted by laws. It also helps the government formulate policies and laws in promoting and regulating AI, which will increase the acceptance of AI systems in public. The wide usage of AI can dramatically improve the efficiencies of our workplaces and can augment the work humans can do. AI systems can take over repetitive or dangerous tasks, and it frees up the human workforce to do work they are better equipped for, e.g., tasks that involve creativity and empathy. In addition. It could increase overall happiness and job satisfaction for such employee engagement.

3 Changes

Not applicable.

References

[1] M. Ben, M. Osadchy, V. Braverman, S. Zhou, and D. Feldman, "Data-independent neural pruning via coresets," in *International Conference on Learning Representations (ICLR)*, 2020.

[2] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," *Advances in Neural Information Processing Systems*, vol. 32, pp. 11 192–11 203, 2019.

[3] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, "The lottery ticket hypothesis for pre-trained bert networks," *arXiv preprint arXiv:2007.12223*, 2020.

[4] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 243–22 255, 2020.

[5] X. Dong, S. Chen, and S. Pan, "Learning to prune deep neural networks via layer-wise optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 2017, pp. 4857–4867.

[6] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJl-b3RcF7

[7] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin, "Stabilizing the lottery ticket hypothesis," *arXiv preprint arXiv:1903.01611*, 2019.

[8] J. Han, P. Luo, and X. Wang, "Deep self-learning from noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5138–5147.

[9] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.

[10] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in neural information processing systems*, 1993, pp. 164–171.

[11] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," in *International Conference on Learning Representations*, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," *arXiv preprint arXiv:1607.03250*, 2016.

[14] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7084–7088.

[15] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[18] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, 1990, pp. 598–605.

[19] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[20] K. Lee, K. Lee, J. Shin, and H. Lee, "Overcoming catastrophic forgetting with unlabeled data in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 312–321.

[21] L. Liebenwein, C. Baykal, H. Lang, D. Feldman, and D. Rus, "Provable filter pruning for efficient neural networks," in *International Conference on Learning Representations*, 2019.

[22] D. Molchanov, A. Ashukha, and D. Vetrov, "Variational dropout sparsifies deep neural networks," in *International Conference on Machine Learning*, 2017, pp. 2498–2507.

[23] A. Renda, J. Frankle, and M. Carbin, "Comparing rewinding and fine-tuning in neural network pruning," in *International Conference on Learning Representations*, 2019.

[24] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05)-Volume 1-Volume 01*, 2005, pp. 29–36.

[25] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[26] S. Srinivas and R. V. Babu, "Data-free parameter pruning for deep neural networks," *arXiv preprint arXiv:1507.06149*, 2015.

[27] J.-C. Su, S. Maji, and B. Hariharan, "When does self-supervision improve few-shot learning?" in *European Conference on Computer Vision*. Springer, 2020, pp. 645–666.

[28] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[29] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.

[30] T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5687–5695.

[31] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196.

[32] M. Ye, C. Gong, L. Nie, D. Zhou, A. Klivans, and Q. Liu, "Good subnetworks provably exist: Pruning via greedy forward selection," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 820–10 830.

[33] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Why lottery ticket wins? a theoretical perspective of sample complexity on pruned neural networks," in *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[34] ——, "How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis," in *International Conference on Learning Representations*, 2022.

[35] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in Neural Information Processing Systems*, vol. 33, 2020.