Research Statement – Resource-Efficient Machine Learning

Xiaotian Han · Texas A&M University https://ahxt.github.io

Overview My research is centered on the fields of artificial intelligence, machine learning, and data science. Specifically, I specialize in designing deep learning algorithms designed for scenarios with limited labeled data or restricted computational capabilities. Instead of merely aiming to reduce costs or expedite existing models training, I envision democratizing cutting-edge machine learning for high-impact societal applications with limited resources, such as predicting rare diseases and modeling low-resourced languages. In doing so, I aim to democratize cutting-edge machine learning techniques, unlocking their potential for wider applications and fostering substantial societal impact.

Research Achievements In the domain of resource-efficient machine learning, I have contributed through peer-reviewed papers at the top machine learning conferences and journals, including ICML, ICLR, TMLR, WWW, KDD, IJCAI, AAAI, TKDE, etc. I was a recipient of ICML2022 Outstanding Paper Award (first author) and Excellent Ph.D. Student Award (one per year) from our department. I have published 16 papers and have gained ~1000 citations, with an h-index of 13 and an i10-index of 14.

Previous Work Over the past few years, my aim has been to democratize machine learning, making it more accessible and efficient, especially in scenarios characterized by limited data and computational resources. To alleviate data scarcity, I have developed methods to enhance machine learning applications, particularly focusing on 1) addressing graph data scarcity: I developed $\mathcal{G}$-Mixup [1] technique innovatively augments graph data, benefiting applications such as drug discovery, 2) addressing clinical text data scarcity for healthcare [2]: Synthetic data generation using ChatGPT addresses the scarcity of clinical text while ensuring privacy, significantly boosting the performance of local models in healthcare. and 3) semi-supervised learning with missing annotations: I proposed G^2R [3] to leverage unsupervised learning for multi-task applications on social networks, efficiently utilizing structural information. To reduce computational resources, I have paved the way for more accessible and efficient machine learning practices such as 1) accelerating LLMs pre-training with less training resources [4]: I proposed the Growlength method to accelerate the pre-training of Large Language Models, reducing computational costs and time, thereby democratizing access to these powerful tools. 2) accelerating GNN training on large graphs [5, 6]: Innovations in training Graph Neural Networks, such as MLP Initialization and the application of Mixup techniques, optimize the training process and reduce resource consumption. and 3) optimizing model deployment costs [7]: I proposed method to provides a flexible and cost-effective solution for model deployment, allowing for in-situ adjustments of accuracy-fairness trade-offs during inference. Besides, I also work on fairness in machine learning [8, 9, 10, 11, 12].

Future Plan I aim to extend, enhance, and popularize resource-efficient machine learning techniques, particularly in large language models, healthcare and AI for scientific discovery. For LLMs, my primary future work will focus on efficient LLMs training and inference, making them more accessible and cost-effective, particularly under computational constraints. This includes creating adaptive computing strategies, dataset distillation techniques, and tailored hardware specific methods. In healthcare, I plan to utilize generative AI for data synthesis in data-constrained settings, addressing data scarcity while maintaining data privacy. Furthermore, I will explore data- and model-efficient methods to advance AI in scientific discovery, such as drug discovery, materials science, and climate science, tackling challenges posed by sparse labeled data, and the need for extensive computational resources.

Previous Achievements & Ongoing Work

1. Democratizing Machine Learning in Scenarios with Limited Data Resources

1.1 Overcoming Graph Data Scarcity with $\mathcal{G}$-Mixup

Despite its growing relevance in various fields, including the critical area of affordable and life-saving drug discovery, graph data often suffers from significant limitations in both quantity and diversity. This scarcity of data can lead to overfitting and reduced generalizability of graph neural networks (GNNs). In response to this, we introduce a novel augmentation technique specifically designed for graph data, termed $\mathcal{G}$-Mixup [1]. While traditional Mixup methods are effective for regular, grid-like, and Euclidean data types like images or tabular data, they are not directly applicable to graph data due to several inherent challenges such as graphs can vary in terms of the number of nodes they contain, they may not align easily. To overcome these issues, we propose $\mathcal{G}$-Mixup to augment graphs for graph classification by interpolating the generator (i.e., graphon) of different classes of graphs. Specifically, we first use graphs within the same class to estimate a graphon. Then, instead of directly manipulating graphs, we interpolate graphons of different classes in the Euclidean space to get mixed graphons, where the synthetic graphs are generated through sampling based on the mixed graphons.

1.2 Tackling Clinical Text Scarcity with Synthetic Data

Utilizing online Large Language Models (LLMs) such as ChatGPT for clinical text mining raises privacy concerns, as it is imperative to prevent any leakage of sensitive patient information. This results in an inadequate supply of in-domain training data for the online model. Despite this, recent advancements in LLMs have demonstrated remarkable capabilities. Their application to the healthcare sector holds great promise and has the potential to yield significant societal benefits. To navigate this dilemma and democratize LLMs for healthcare sectors, we propose an innovative solution that leverages the capabilities of online LLMs while ensuring the protection of sensitive patient information [2]. Our approach involves training high-performing local models using synthetic labeled data generated by ChatGPT for clinical text mining tasks. We prompt ChatGPT to produce high-quality annotated sentences and then train local models, which exhibit substantial improvements, nearing state-of-the-art performance. This strategy proves to be an effective method for acquiring premium training data from LLMs, circumventing potential privacy issues, and thereby democratizing the responsible and ethical use of artificial intelligence in the healthcare sector.

1.3 Addressing Missing Label Annotations for Multi tasks

Online social networks often serve multiple functions, necessitating a variety of task-specific labels for effective data analysis and application. Addressing this need, our work adopts a semi-supervised learning approach, leveraging solely the graph's structural information to learn a comprehensive representation adaptable to various tasks. Graph Neural Networks (GNNs), while powerful, are complex and computationally intensive, which can limit their online deployment. This challenge highlights the need for unsupervised representation learning within GNN frameworks, a need particularly acute in industrial contexts where computational resources may be scarce. Thus, we propose Geometric Graph Representation Learning (G^2R) [3] to learn node representations in an unsupervised manner via maximizing rate reduction, which maps nodes into distinct groups (implicitly stored in the adjacency matrix) into different representation subspaces. The geometric difference between subspaces makes the node representation enjoy rich semantic information.

2. Democratizing Machine Learning in Scenarios with Limited Computational Resources

2.1 Democratizing LLMs through Accelerated Pre-Training

The democratization of Large Language Models (LLMs) aims to make these powerful tools accessible and beneficial for a broader spectrum of users. However, it remains a significant challenge due to the substantial computational resources and time required for pre-training, which are often beyond the reach of smaller organizations and researchers. Addressing this disparity is crucial for fostering innovation and ensuring equitable access to advanced AI technologies. In response to these issues, we introduce a novel, simple, and effective method named Growlength [4] to accelerate the pre-training process of LLMs. Our method progressively increases the training sequence length throughout the pre-training phase, thereby mitigating computational costs and enhancing efficiency. Our method not only converges more swiftly but also exhibits superior performance compared to existing methods, and does not require any additional engineering efforts, making it a practical solution. Making LLMs more accessible will foster a diverse and inclusive community, opening up opportunities for smaller organizations and independent researchers to contribute to and benefit from advancements in AI.

2.2 Accelerating GNN Training on Large Graphs

Training Graph Neural Networks (GNNs) on large graphs is a resource-intensive and time-consuming process, hindering rapid updates for daily applications and contributing significantly to carbon dioxide emissions. Optimizing and accelerating GNN training is crucial for GNN applications that require constant updates, such as social analysis [13, 14, 15] and road network optimization. To optimize GNN training, we present two innovative strategies: MLP Initialization (MLPInit) and a novel application of Mixup techniques 1) The MLPInit method [5], primarily focused on accelerating GNN training, addresses the inherent complexities and time-intensive nature of training GNNs on large graphs. This is achieved by initializing GNNs with weights from a fully trained analog Multi-Layer Perceptron (MLP), referred to as PeerMLP. This initialization not only results in training speedup but also often enhances prediction performance. 2) We established a connection between graph convolution and Mixup techniques [6], revealing a unified approach towards feature representation. We treated graph convolution as a specialized form of Mixup applied during both training and testing phases and achieved comparable performance of GNNs by training equivalent MLPs with Mixup. This revelation opens avenues for efficient GNN design. With the accelerated GNN training, the GNNs will ultimately contribute to the broader goal of refining GNN performance for various applications.

2.3 Optimizing Model Deployment Costs

During model deployment, practitioners have to balance multiple, sometimes conflicting, objectives. This complexity often necessitates the training of multiple models, each tailored to a specific requirement, which can be resource-intensive. Moreover, it is crucial to have fine-grained control over the trade-offs between these multiple objectives during inference. In light of these challenges, we propose to provide a more adaptable and cost-effective solution in the domain of fairness, where flexible accuracy-fairness trade-offs are practically desired due to diverse regularization on fairness. Existing fairness methods typically offer a fixed accuracy-fairness trade-off. To reduce the heavy computational resources, we propose *You Only Debias Once* [7] to achieve in-situ flexible accuracy-fairness trade-offs at inference time, using *a single model* that trained only once. Instead of pursuing one fairness-optimum model, we aim to find a "line" that connects the accuracy-optimum and fairness-optimum models. Points (models) on this line implement varying levels of accuracy-fairness trade-offs. At inference time, by manually selecting the position on the "line", our proposed method can achieve arbitrary accuracy-fairness trade-offs.

Future Research Plans

1. Democratizing Large Language Models through Resource Efficiency As Large Language Models (LLMs) continue to grow in size, ensuring their efficient application becomes an increasing challenge during both the training and inference phases. Over the next few years, I aim to extend my prior research on enhancing LLMs with constrained data and hardware resources. My goal is to make LLMs more accessible and cost-effective, ultimately contributing to the democratization of these powerful machine learning tools.

1.1 LLMs with limited Computational Resources As LLMs grow in size and complexity, computational constraints have become a bottleneck for both training and inference. In the near future, I will explore the following key areas to democratize LLMs: i) more efficient training algorithms that can accelerate the convergence of LLMs. ii) investigating adaptive computing, where different tasks or data segments utilize varying model sizes or computational resources based on their complexity. iii) Developing dataset distillation techniques to compress the large dataset to a coreset, smaller while preserving most of their capabilities. iv) Investigate the hardware-awared LLMs inference techniques.

1.2 LLMs For Limited Training Data The high quality data is limited. Improving the performance of LLMs further needs more text data or a higher-quality subset of dataset. With the constraints on high-quality data availability, it becomes essential to derive more from less. My future research will focus on: 1) Leverage semi-supervised and unsupervised learning, effectively exploiting limited labeled data alongside abundant unlabeled data. 2) Exploring data augmentation methods tailored specifically for text, with the aim of diversifying available qulity training data for LLMs.

2. Utilizing Generative AI in Resource-Constrained Healthcare The adoption of machine learning (ML) in healthcare necessitates strict compliance with data privacy and security requirements, as the highly sensitive healthcare data, such as medical images and clinical notes, often result in sparse training data. Recognizing the sensitive nature of healthcare data, my previous work made a first step to utilize LLMs for data synthesis and training localized models. But the challenges of data scarcity highlight the further need for more advanced techniques in the healthcare domain. Building on my prior work, I aim to delve into state-of-the-art generative AI applications for data synthesis in healthcare settings characterized by data scarcity. This advanced approach has the potential to significantly alleviate the challenges posed by data scarcity, enabling the generation of synthetic yet realistic data samples that can augment our training datasets. Furthermore, the use of generative AI for data synthesis also aligns with the need to uphold data privacy, which can ensure that we are not compromising patient privacy.

3. Advancing Scientific Discovery via Resource-Efficient Machine Learning Artificial Intelligence shows great promise in driving scientific discovery forward, particularly through the adoption of resource-efficient machine learning techniques, which are essential given the complex and voluminous nature of scientific data. I plan to delve into advanced data- or model-efficient methods to advance AI for scientific discovery. In drug discovery, the scarcity of labeled data poses a significant challenge, as extensive wet lab experiments are required to create and analyze drug candidates for model pre-training. This necessitates the use of data augmentation or data generation techniques. Similarly, in materials science, researchers employ machine learning models to predict the properties of novel materials and also face data scarcity. This challenge requires access to laboratory resources and equipment. Additionally, massive computational resources are needed for simulations and data analysis in materials science. In climate science, substantial high-performance computing resources are required for climate modeling.

References

[1] **Xiaotian Han**, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-mixup: Graph data augmentation for graph classification. In *ICML*, 2022. URL https://proceedings.mlr.press/v162/han22c/han22c.pdf.

[2] **Xiaotian Han***, Ruixiang Tang*, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? In *AMIA*, 2023. URL https://arxiv.org/pdf/2303.04360.pdf.

[3] **Xiaotian Han**, Zhimeng Jiang, Ninghao Liu, Qingquan Song, Jundong Li, and Xia Hu. Geometric graph representation learning via maximizing rate reduction. In *WWW*, 2022. URL https://doi.org/10.1145/3485447.3512170.

[4] **Xiaotian Han***, Hongye Jin*, Jingfeng Yang, Zhimeng Jiang, Chia-Yuan Chang, and Xia Hu. Growlength: Accelerating llms pretraining by progressively growing training length. *arXiv preprint arXiv:2310.00576*, 2023. URL https://arxiv.org/pdf/2310.00576.pdf.

[5] **Xiaotian Han**, Tong Zhao, Yozen Liu, Xia Hu, and Neil Shah. MLPInit: Embarrassingly simple GNN training acceleration with MLP initialization. In *ICLR'23*, 2023. URL https://openreview.net/forum?id=P8YIphWNEGO.

[6] **Xiaotian Han**, Hanqing Zeng, Yu Chen, Shaoliang Nie, Jingzhou Liu, Kanika Narang, Zahra Shakeri, Karthik Abinav Sankararaman, Song Jiang, Madian Khabsa, et al. On the equivalence of graph convolution and mixup. *arXiv preprint arXiv:2310.00183*, 2023. URL https://arxiv.org/pdf/2310.00183.pdf.

[7] **Xiaotian Han**, Tianlong Chen, Kaixiong Zhou, Zhimeng Jiang, Zhangyang Wang, and Xia Hu. You only debias once: Towards flexible accuracy-fairness trade-offs at inference time. *arXiv preprint*, 2023.

[8] **Xiaotian Han**, Zhimeng Jiang, Hongye Jin, Zirui Liu, Na Zou, Qifan Wang, and Xia Hu. Retiring δDP: New distribution-level metrics for demographic parity. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=LjDFIWWVVa.

[9] **Xiaotian Han***, Zhimeng* Jiang, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. Chasing fairness under distribution shift: a model weight perturbation approach. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=DVjyq5eCAD.

[10] **Xiaotian Han**, Kaixiong Zhou, Ting-Hsiang Wang, Jundong Li, Fei Wang, and Na Zou. Marginal nodes matter: Towards structure fairness in graphs. *KDDExploration*, 2023. URL https://arxiv.org/pdf/2310.14527.pdf.

[11] Zhimeng Jiang, **Xiaotian Han**, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YigKlMJwjye.

[12] **Xiaotian Han**, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. *arXiv preprint arXiv:2306.09468*, 2023. URL https://arxiv.org/pdf/2306.09468.pdf.

[13] Shaohua Fan, Junxiong Zhu, **Xiaotian Han**, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. Metapath-guided heterogeneous graph neural network for intent recommendation. In *KDD*, pages 2478–2486, 2019. URL https://dl.acm.org/doi/10.1145/3292500.3330673.

[14] Chuan Shi, **Xiaotian Han**, Li Song, Xiao Wang, Senzhang Wang, Junping Du, and S Yu Philip. Deep collaborative filtering with multi-aspect information in heterogeneous networks. *IEEE transactions on knowledge and data engineering*, 33(4):1413–1425, 2019. URL https://arxiv.org/pdf/1909.06627.pdf.

[15] **Xiaotian Han**, Chuan Shi, Senzhang Wang, S Yu Philip, and Li Song. Aspect-level deep collaborative filtering via heterogeneous information networks. In *IJCAI*, volume 18, pages 3393–3399, 2018. URL https://www.ijcai.org/proceedings/2018/0471.pdf.