Field of Interest (Limit: 2250 char)

1. *In terms a general audience would understand, describe an important, outstanding challenge in mathematics, statistics or computer science that you would like to pursue in your research. (1/3)*
2. *Describe the particular mathematics, statistics or computer science problem that you would like to pursue in your research. What would be the impact on high-performance computing and on science, engineering and/or society in general if this challenge could be successfully addressed? (2/3)*

I propose to develop new, computationally efficient Adaptive Privacy-Preserving LEarning (APPLE) methods for building large privacy-preserving machine learning (ML) models. In so doing, I will enable a wide range of DOE science applications to use extremely large datasets and powerful HPC systems to build exquisitely detailed ML models–without risk of leaking sensitive (e.g., protected or private research) data and by reducing unnecessary computation.

As an example of where APPLE methods will apply, consider the problem (at the Advanced Photon Source: APS) of analyzing ptychography data to identify counterfeit and malicious integrated circuits (ICs). ML models trained on many real circuits can accelerate analysis enormously, but risk "memorizing" sensitive information in their weights [1]. Similarly, models trained to predict environmental justice implications of climate policies may leak sensitive household information, and predictive models of power grid stability may leak sensitive utility provider information. Additionally, given the size of these datasets (e.g., IC Dataset is >10PB per chip), training these models requires enormous computing resources[2].

I will study ways to efficiently organize and process large datasets and to train models in a manner that allows trained models to forget specific (e.g., inaccurate or sensitive) portions of data, without retraining the model from scratch. Such trained models will have malleable weights that can be version controlled to track which data were used for which model version. I will investigate, for example, teacher models to selectively train obfuscated sets of student models [3]; federated learning to adaptively combine isolated submodels based on security needs [4]; and segmented learning to dynamically reshape networks based on data impact.

APPLE models will enable more efficient use of HPC resources because a single model can be maintained for each domain, and then refined for specific tasks by redacting specific data sources–in effect, dynamically adapting (without retraining) the model for different security levels. APPLE will thus enable more secure and comprehensive scientific discoveries from ML while reducing resource usage and researcher time.

Works Cited:

[1] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," http://arxiv.org/abs/1802.08232

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” doi: 10.1145/3442188.3445922.

[3] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” http://arxiv.org/abs/1503.02531

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” http://arxiv.org/abs/1602.05629

Given the size of the IC datasets (>10PB each), training requires enormous computing resources—a problem that will only become more significant as ML models continue to grow in size and complexity[2].

---------- Needed to cut characters in citations ----------

[1] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks,” *arXiv:1802.08232 [cs]*, Jul. 2019, Accessed: Jan. 05, 2022. [Online]. Available: http://arxiv.org/abs/1802.08232

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[3] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” *arXiv:1503.02531 [cs, stat]*, Mar. 2015, Accessed: Jan. 05, 2022. [Online]. Available: http://arxiv.org/abs/1503.02531

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” *arXiv:1602.05629 [cs]*, Feb. 2017, Accessed: Jan. 05, 2022. [Online]. Available: http://arxiv.org/abs/1602.05629

High-Performance Computing (Limit: 2250 char)

1. *What is the most complex calculation you have run on a high-performance machine as part of your research experience? Or if you haven't run a high-performance computing system, tell us about the most complex computational problem you have tackled. (1/2)*
2. *Imagine if you were given access to resources 100 times more powerful than what you have access to. What would that enable you to do, and what do you perceive the mathematical and computer science challenges to be? (1/2)*

I trained a vision transformer (ViT) neural network with 22 million trainable parameters for image classification on ThetaGPU at Argonne National Laboratory to benchmark system performance. This training was performed using TensorFlow and distributed using the Horovod framework across 1, 2, 3, 4, and 8 compute nodes to understand how training/testing scales. Each node includes 8 NVIDIA A100 graphics processing units (GPUs). When training the ViT on 1 node, my code processed about 23,000 images/second during training and about 96,000 images/second during testing. When the training was scaled to 4 nodes, there was about a 3.8x and 4x performance increase for training and testing, respectively (near perfect scaling). However, when scaled to 8 nodes (64 GPUs), there was only about a 6.7x and 6.5x performance increase for training and testing, respectively. These results may demonstrate a system-level threshold for scalability; for more than 4 nodes, the communication-induced delay may be too great and prevent perfect scaling.

As I have utilized 8 nodes on ThetaGPU (~1.33 PFLOP/s), a 100x increase in resources would resemble ~100 PFLOP/s, or roughly 1/10th the computing power of the upcoming Aurora supercomputer at Argonne. At this scale, I would be able to build an APPLE model to efficiently tackle the APS IC dataset (>10 PB) as this dataset is many magnitudes larger than the largest language dataset presently available to science[1]. However, a linear 100x performance increase is not likely as stipulated by Ahmdal's law[2]. Challenges to efficiently scaling the problem to exascale include balancing and minimizing inter-node communication cost with memory-access cost[3], considering energy consumption[4], and conducting time-consuming fault detection and debugging[5]. Software bottlenecks may also hinder scaling; developers must use concurrent and latency tolerant algorithms and ensure that software maximizes hardware throughput. I expect my proposed APPLE approach to enhance scalability by introducing algorithms with greater degrees of freedom (e.g., decoupled teacher/student training, parallelized federated learning, and sparse model training) and significantly reducing use of HPC resources for repetitive tasks.

Works Cited:

[1] L. Gao *et al.*, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," http://arxiv.org/abs/2101.00027

[2] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," Doi: 10.1145/1465482.1465560.

[3] J. G. Pauloski *et al.*, “KAISA: an adaptive second-order optimizer framework for deep neural networks,” doi: 10.1145/3458817.3476152.

[4] J. Mair, Z. Huang, D. Eyers and Y. Chen, "Quantifying the Energy Efficiency Challenges of Achieving Exascale Computing," doi: 10.1109/CCGrid.2015.130.

[5] A. Gainaru, F. Cappello, M. Snir, and W. Kramer, “Fault prediction under the microscope: A closer look into HPC systems,” doi: 10.1109/SC.2012.57.

Had to cut down on characters for references:

[1] L. Gao *et al.*, “The Pile: An 800GB Dataset of Diverse Text for Language Modeling,” *arXiv:2101.00027 [cs]*, Dec. 2020, Accessed: Jan. 10, 2022. [Online]. Available: http://arxiv.org/abs/2101.00027

[2] G. M. Amdahl, “Validity of the single processor approach to achieving large scale computing capabilities,” in *Proceedings of the April 18-20, 1967, spring joint computer conference on - AFIPS ’67 (Spring)*, Atlantic City, New Jersey, 1967, p. 483. Doi: 10.1145/1465482.1465560.

[3] J. G. Pauloski *et al.*, “KAISA: an adaptive second-order optimizer framework for deep neural networks,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, St. Louis Missouri, Nov. 2021, pp. 1–14. doi: 10.1145/3458817.3476152.

[4] J. Mair, Z. Huang, D. Eyers and Y. Chen, "Quantifying the Energy Efficiency Challenges of Achieving Exascale Computing," *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2015, pp. 943-950, doi: 10.1109/CCGrid.2015.130.

[5] A. Gainaru, F. Cappello, M. Snir, and W. Kramer, “Fault prediction under the microscope: A closer look into HPC systems,” in *2012 International Conference for High Performance Computing, Networking, Storage and Analysis*, Salt Lake City, UT, Nov. 2012, pp. 1–11. doi: 10.1109/SC.2012.57.

Program of Study (Limit: 2250 char)

Describe how the courses listed in your planned program of study would help prepare you to address the challenges you have described in questions 1 and 2. Discuss your rationale for choosing these courses. How will the science or engineering application courses you have selected impact your research?

My research focus is on creating efficient ML workflows for HPC systems, so I choose to concentrate my computer science and HPC courses on foundational systems topics like parallel computing and distributed computing. I also include a course on large systems for Deep Learning, as it explores modern HPC approaches to tackling large ML problems. I also want to strengthen my understanding of Operating Systems, as this is an area I have not formally studied, but know to be integral to system performance.

As I have taken foundational ML classes during undergrad, I look forward to extending my knowledge of core mathematical concepts behind modern ML through classes such as Linear Algebra and Fourier Analysis. I am also excited to take a course on computational Numerical Methods; I chose a course with a focus on partial differential equations (PDEs) because the skills from this class will be transferable to many methods/applications in computational science.

I focus my science application courses in environmental and climate science, areas for which I developed a passion during my undergraduate career. (I pursued a minor in Environmental Science and spent a semester at the UNC Institute of Marine Science conducting two research projects. The first used deep learning (DL) to quantify blue whale behavior from drone imagery and the second attempted to model the geomorphological and hydrological changes of a barrier island post-Hurricane Dorian.) Through this work, I was exposed to the challenges of developing and revising accurate models with large and dynamic (and sometimes sensitive) datasets and insufficient computational capabilities. For example, my DL/Whale project constantly exhausted the computational limits of my lab as I was repeatedly training (and retraining) multiple ML models with variations of the same datasets; this was an application that would have benefited from an APPLE model framework. I look forward to using my science classes to identify additional applications that can benefit from my APPLE research, uncover new environmental science research areas, and understand the computational capabilities needed by environmental science researchers.

Edit	Course number	Course Title	Credit hours	Term and Year	Grade	Academic Level
Science/Engineering						
Modify, *Delete*	GEOS 34220	Climate Foundations	3Q	Fall 2022		G
Modify, *Delete*	GEOS 34250	Geophysical Fluid Dynamics: Understanding the Motions of the Atmosphere and Oceans	3Q	Spring 2022		G
Modify, *Delete*	GEOS 34600	Introduction to Atmosphere, Ocean, and Climate Modeling	3Q	Winter 2022		G
Mathematics and Statistics						
Modify, *Delete*	CAAM 31100	Mathematical Computation III: Numerical Methods for PDEs	3Q	Winter 2023		G
Modify, *Delete*	CAAM 31430	Applied Linear Algebra	3Q	Spring 2022		G
Modify, *Delete*	CAAM 31460	Applied Fourier Analysis	3Q	Fall 2023		G
High-Performance Computing						
Modify, *Delete*	CMSC 23010	Parallel Computing	3Q	Fall 2022		G
Insert a new course in High-Performance Computing						
Computer Science						
Modify, *Delete*	CMSC 33310	Advanced Distributed Systems	3Q	Winter 2022		G
Modify, *Delete*	CMSC 33000	Operating Systems	3Q	Winter 2023		G

Modify, *Delete*	CMSC 35200	Deep Learning Systems	3Q	Fall 2023		G