Deep Learning on Supercomputers

Zhao Zhang[1], Valeriu Codreanu[2], Ian Foster[3]
[1]Texas Advanced Computing Center, [2]SURFsara, [3]University of Chicago & Argonne National Laboratory

1 Abstract

The Deep Learning (DL) on Supercomputers workshop provides a forum for practitioners working on any and all aspects of DL for science and engineering in the High Performance Computing (HPC) context to present their latest research results and development, deployment, and application experiences. The general theme of this workshop series is the intersection of DL and HPC; the theme of this particular workshop is the applications of DL methods in science and engineering: novel uses of DL methods, e.g., convolutional neural networks (CNN), recurrent neural networks (RNN), generative adversarial networks (GAN), and reinforcement learning (RL), in the natural sciences, social sciences, and engineering, to enhance innovative applications of DL in traditional numerical computation. Its scope encompasses application development in scientific scenarios using HPC platforms; DL methods applied to numerical simulation; fundamental algorithms, enhanced procedures, and software development methods to enable scalable training and inference; hardware changes with impact on future supercomputer design; and machine deployment, performance evaluation, and reproducibility practices for DL applications with an emphasis on scientific usage. This workshop will be centered around published papers. Submissions will be peer-reviewed, and accepted papers will be published as part of the Joint Workshop Proceedings by Springer.

Topics include but are not limited to:

- DL as a novel approach of scientific computing
  - Emerging scientific applications driven by DL methods
  - Novel interactions between DL and traditional numerical simulation
  - Effectiveness and limitations of DL methods in scientific research
  - Algorithms and procedures to enhance reproducibility of scientific DL applications
- DL for Science workflows
  - Data management through the life cycle of scientific DL applications
  - DL performance evaluation and analysis on deployed systems
- Scalable DL methods to address the challenges of demanding scientific applications
  - General algorithms and procedures for efficient and scalable DL training
  - General algorithms and systems for large scale model serving for scientific use cases
  - New software, and enhancements to existing software, for scalable DL
  - DL communication optimization at scale
  - I/O optimization for DL at scale
  - DL performance modeling and tuning of DL on supercomputers
  - DL benchmarks on supercomputers
- Novel hardware designs for more efficient DL
  - Processors, accelerators, memory hierarchy, interconnect changes with impact on DL in the HPC context

As part of the reproducibility initiative, the workshop will require authors to provide information such as the algorithms, software releases, datasets, and hardware configurations used. For performance evaluation studies, we will encourage authors to use well-known benchmarks or applications with open accessible datasets: for example, MLPerf (https://github.com/mlperf/training) and ResNet-50 with the ImageNet-1K dataset (http://www.image-net.org/archive/stanford/fall11_whole.tar).

2 Relevance and Impact of the Workshop for ISC

Deep learning (DL) is a class of machine learning algorithms in which multiple layers of nonlinear processing units are used for feature extraction and transformation, with each successive layer taking the output from the previous layer as input [20]. Advances in both implementation approaches and hardware capabilities have contributed to a recent renaissance in such methods, with impressive performance in applications ranging from autonomous driving [7] and board games [24] to the checkout-free grocery store [3]. Fascinating results have been obtained on scientific problems, such as supernova classification in astronomy [9], target extension in drug discovery [4], melanoma recognition [13] and

early heart failure detection [12] in disease diagnosis, exotic particle (Higgs boson) search in high energy physics [5], and neuroimaging analysis [23], gene annotation [11] and molecular dynamics [21] in bioinformatics. More recently, we have witnessed outstanding results from research groups using the world's largest supercomputers to tackle relevant scientific problems with deep learning. We can see successful examples in climate modeling for detecting extreme weather events [18], using physics-informed generative models for approximating solutions of stochastic PDEs [26], or approximating the solutions for difficult inverse problems in material imaging [19], all reaching peak performance in excess of 1 ExaFLOP/s. Recently, we also see interesting developments combining probabilistic programming with deep learning for tackling key scientific problems in high-energy physics [6] on large-scale supercomputers.

Most current DL research uses open source DL frameworks such as TensorFlow [1], PyTorch [22], Caffe [16], Keras, CNTK [28], MXNet [10], GeePS [15], Chainer [25], and Poseidon [29]. These frameworks are usually optimized for various heterogeneous architectures based on CPUs, GPUs, TPUs, and/or FPGAs. Many use MPI as their communication layer, indicating their readiness for supercomputers from a technical perspective.

We have seen growing interest in the fusion of deep learning and HPC in past years. The organizers of this proposed workshop have organized 1st Deep Learning on Supercomputers workshop[1] in SC18 in Dallas, TX, USA. The workshop had one keynote speech and eight invited talks from academia, industry, and national labs. The workshop attracted over 200 attendees on the very last day of the conference. Afterwards, we have organized the 1st Deep Learning for Science workshop[2] in ISC19 in Frankfurt, Germany. The workshop, part of the same *Deep Learning on Supercomputers* series, contained a keynote speech, eight invited talks, and a panel discussion at the end that allowed for in-depth interaction between the audience and the speakers. Finally, during SC19 in Denver, CO, USA, we have organized the 3rd Deep Learning on Supercomputers workshop[3], where we have received 23 paper submissions and have selected 11 excellent contributions. The workshop was very popular and attended by more than 400 people.

DL is an opportunity for the HPC community in two ways. First, the computation and communication patterns of DL are a natural fit for modern supercomputers, due to the unparalleled computing power, exceptional memory architectures, and low-latency high-bandwidth interconnects found in such systems. Several research groups have reported using supercomputers to reduce DL training time from hours to minutes [2, 14, 17, 27]. Second, it is becoming clear that DL has as important role to play in scientific computing, both as a standalone data science method [4, 5, 13, 21] and as a method for driving traditional simulations [8].

However, DL poses new challenges to the scientific research and HPC communities, such as efficiently handling complex large-scale data. Unlike the data used for tasks such as image classification on 2D, 3-channel images, scientific data tends to have more channels or be multidimensional in nature (3D, 4D). Most of these real-life scientific problems have large datasets, sometimes in the multi-TB range. Efficiently handling this large-scale complex data volume usually requires both scaling neural network training across a large number of compute nodes, as well as making sure I/O bottlenecks are alleviated and do not become a showstopper on today's supercomputers. Another important challenge is maintaining the final solution accuracy of these networks when trained at scale. Naively scaling DL algorithms by just increasing the batch size is likely to result in either low machine utilization or accuracy degradation. It is very important to design a good network architecture and to perform hyperparameter tuning. This comes however with its own challenge due to the large computational requirements of individual training runs. Yet another challenge is encountered when trying to do performance evaluation of various combinations of frameworks, DL algorithms, and hardware platforms. This is due to the rapid development cycles used by the community, and may lead to either unreproducible or very difficult to reproduce results. HPC vendors and operators will encounter these challenges as more scientists seek to run DL applications on supercomputers.

The objective of this proposed workshop is to provide a venue in which engineers and researchers can exchange the latest developments and results in this field. We expect the workshop to enable insightful discussions of challenges such as those listed above and to incubate collaborations across institutions. The ultimate goal is to unite people in the HPC and scientific computing communities in the pursuit of new methods that will expedite the use of supercomputers for scientific discoveries.

The impact of this workshop will be both deep from a technical perspective and broad in terms of the areas that it influences. It will:

- help system designers learn requirements of scientific users;
- enable scientists to benefit from the rapid pace of innovation in the DL community;

[1] https://www.tacc.utexas.edu/workshop/2018/deep-learning
[2] https://dlonsc.github.io/
[3] https://dlonsc19.github.io/

- increase the capacity and capability of supercomputers;
- increase coherence between modeling & simulation and data analytic computing; and
- disseminate DL innovations to a wide range of industrial and scientific domains

3 Tentative Program Committee

We have recruited for the workshop program committee leading researchers and engineers working on deep learning applications, infrastructure, and hardware. Members are from USA, Japan, Spain, Netherlands, and Russia, and from academia and industry. Drs. Zhao Zhang, Valeriu Codreanu, and Ian Foster will co-chair the workshop. The following people have all agreed to participate in the committee.

- Valeriu Codreanu (co-chair), SURFsara, Netherlands
- Ian Foster (co-chair), UChicago & ANL, USA
- Zhao Zhang (co-chair), TACC, USA
- Weijia Xu (proceeding chair), TACC, USA
- Ahmed Al-Jarro, Fujitsu Laboratories of Europe, UK
- Takuya Akiba, Preferred Networks, Japan
- Thomas S. Brettin, ANL, USA
- Maxwell Cai, SURFsara, Netherlands
- Erich Elsen, Google Brain, USA
- Steve Farrell, LBNL, USA
- Song Feng, IBM Research, USA
- Boris Ginsburg, NVIDIA, USA
- Torsten Hoefler, ETH, Switzerland
- Jessy Li, UT Austin, USA
- Zhengchun Liu, ANL, USA
- Peter Messmer, Nvidia, USA
- Damian Podareanu, SURFsara, Netherlands
- Simon Portegies Zwart, Leiden Observatory, Netherlands
- Judy Qiu, Indiana University, USA
- Arvind Ramanathan, ORNL, USA
- Vikram Saletore, Intel, USA
- Mikhail E. Smorkalov, Intel, Russia
- Rob Schreiber, Cerebras, USA
- Dan Stanzione, TACC, USA
- Rick Stevens, UChicago & ANL, USA
- Wei Tan, Citadel, USA
- Jordi Torres, Barcelona Supercomputing Center, Spain
- Daniela Ushizima, LBNL, USA
- Sofia Vallecorsa , CERN, Switzerland
- David Walling, TACC, USA
- Markus Weimer, Microsoft, USA
- Kathy Yelick, UC Berkeley & LBNL, USA

4 Workshop Format

We believe that the tremendous interest in this topic, and the breadth of expected contributions, requires a full day workshop. We will center the workshop around published technical papers. Authors will be invited to submit papers with unpublished, original work with a minimum of 6 pages and a maximum of 12 pages in single column text with LNCS style.

We will organize the day as a keynote speech (60 minutes including Q&A), four 75-minute technical sessions, and one panel discussion (60 minutes). The workshop will start at 9:00AM and end at 6:00PM. We plan to accept 8-12 technical papers, although this number is subject to change based on the quality of the submissions.

We will invite a keynote speaker with a reputation in using deep learning methods for scientific research. The program committee will discuss the candidate list once the workshop is approved.

5 Expected Outcome

We expect between 100 to 150 attendees to attend the entirety of the workshop, as our previous workshop in SC19 attracted over 400 attendees. Participants are likely to include those working on all aspects of deep learning methods on supercomputers: scientists from all research domains, deep learning infrastructure designers, hardware designers, framework developers, deployment engineers, and application developers; and also representatives of funding agencies, supercomputing centers, cloud service providers, and supercomputer vendors. The accepted papers will be published in the Joint Workshop Proceeding by Springer.

Efficient execution of deep learning algorithms on supercomputer is a rapidly developing topic. It is challenging for engineers and researchers in different countries and disciplines to track progress in this area. The proposed workshop thus meets an important need. We expect that by enabling the timely dissemination of the latest research and engineering progress to users, we will allow the HPC community to better facilitate industrial and scientific discoveries with state-of-the-art DL technology.

6 Advertising and Attracting Attendees

We will advertise the workshop via a combination of in-person communication, community email lists, computer center user databases, and social networks.

Our program committee includes leading deep learning researchers from both academia and industry, and spans domain scientists, computation providers, hardware vendors, software providers, service providers, and application developers. We will ask committee members to advertise the workshop in their home institutions, within their social networks, and via social media, e.g., Twitter, Facebook, and Linkedin.

We will also use email lists in the HPC, deep learning, scientific computing, cloud computing, and domain application communities to spread the call for papers. We will use TACC's, Argonne's, and SURFsara's computer center user database to encourage domain researchers who are not familiar or new to DL to attend this workshop. We will leverage the workshop organizers and sponsoring institutions' social media, to advertise the call for papers and attract attendees.

7 Expected Timeline

We propose the following timeline for the workshop:

- Feb 21st, Organizational meeting,
- Feb 22nd – Aug 31st, Inviting keynote speaker, website setup, advertising, contacting publishers for proceedings
- Mar 31st, Technical paper due
- Apr 15th, Technical paper review due
- Apr 16th, Technical program released
- May 15th, Camera-ready due
- June 25th, Workshop day

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*, 2016.

[2] T. Akiba, S. Suzuki, and K. Fukuda. Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*, 2017.

[3] Alba, Davey. Only Amazon Could Make a Checkout-Free Grocery Store a Reality, 2016. https://www.wired.com/2016/12/amazon-go-grocery-store/.

[4] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, 2016.

[5] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

[6] A. G. Baydin, L. Shao, W. Bhimji, L. Heinrich, L. Meadows, J. Liu, A. Munk, S. Naderiparizi, B. Gram-Hansen, G. Louppe, et al. Etalumis: Bringing probabilistic programming to scientific simulators at scale. *arXiv preprint arXiv:1907.03382*, 2019.

[7] Carol Reiley. Deep Driving, 2016. https://www.technologyreview.com/s/602600/deep-driving/.

[8] J. Carrasquilla and R. G. Melko. Machine learning phases of matter. *Nature Physics*, 2017.

[9] T. Charnock and A. Moss. Deep recurrent neural networks for supernovae classification. *arXiv preprint arXiv:1606.07442*, 2016.

[10] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[11] D. Chicco, P. Sadowski, and P. Baldi. Deep autoencoder neural networks for gene ontology annotation predictions. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 533–540. ACM, 2014.

[12] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, page ocw112, 2016.

[13] N. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. R. Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *arXiv preprint arXiv:1610.04662*, 2016.

[14] V. Codreanu, D. Podareanu, and V. Saletore. Scale out for large minibatch sgd: Residual network training on imagenet-1k with improved accuracy and reduced time to train. *arXiv preprint arXiv:1711.04291*, 2017.

[15] H. Cui, H. Zhang, G. R. Ganger, P. B. Gibbons, and E. P. Xing. GeePS: Scalable deep learning on distributed gpus with a gpu-specialized parameter server. In *Proceedings of the Eleventh European Conference on Computer Systems*, page 4. ACM, 2016.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[17] T. Kurth, J. Zhang, N. Satish, E. Racah, I. Mitliagkas, M. M. A. Patwary, T. Malas, N. Sundaram, W. Bhimji, M. Smorkalov, J. Deslippe, M. Shiryaev, S. Sridharan, Prabhat, and P. Dubey. Deep learning at 15pf: Supervised and semi-supervised classification for scientific data. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '17, pages 7:1–7:11, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5114-0. doi:10.1145/3126908.3126916. URL http://doi.acm.org/10.1145/3126908.3126916.

[18] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, et al. Exascale deep learning for climate analytics. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, page 51. IEEE Press, 2018.

[19] N. Laanait, J. Romero, J. Yin, M. T. Young, S. Treichler, V. Starchenko, A. Borisevich, A. Sergeev, and M. Matheson. Exascale deep learning for scientific inverse problems. *arXiv preprint arXiv:1909.11150*, 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[21] P. D. Lena, K. Nagata, and P. F. Baldi. Deep spatio-temporal architectures and learning for protein structure prediction. In *Advances in neural information processing systems*, pages 512–520, 2012.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[23] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, and V. D. Calhoun. Deep learning for neuroimaging: a validation study. *arXiv preprint arXiv:1312.5847*, 2013.

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[25] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, pages 1–6, 2015.

[26] L. Yang, S. Treichler, T. Kurth, K. Fischer, D. Barajas-Solano, J. Romero, V. Churavy, A. Tartakovsky, M. Houston, G. Karniadakis, et al. Highly-scalable, physics-informed gans for learning solutions of stochastic pdes. *arXiv preprint arXiv:1910.13444*, 2019.

[27] Y. You, Z. Zhang, C. Hsieh, J. Demmel, and K. Keutzer. Imagenet training in minutes. *in submission to CVPR 2018*, 2018.

[28] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, J. Droppo, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, B. Peng, A. Stolcke, and M. Slaney. An introduction to computational networks and the computational network toolkit. *Microsoft Technical Report MSR-TR-2014–112*, 2014.

[29] H. Zhang, Z. Hu, J. Wei, P. Xie, G. Kim, Q. Ho, and E. Xing. Poseidon: A system architecture for efficient gpu-based deep learning on multiple machines. *arXiv preprint arXiv:1512.06216*, 2015.

8 Call for Papers

The Deep Learning (DL) on Supercomputers workshop provides a forum for practitioners working on any and all aspects of DL for scientific research in the High Performance Computing (HPC) context to present their latest research results and development, deployment, and application experiences. The general theme of this workshop series is the intersection of DL and HPC, while the theme of this particular workshop is centered around the applications of deep learning methods in scientific research: novel uses of deep learning methods, e.g., convolutional neural networks (CNN), recurrent neural networks (RNN), generative adversarial network (GAN), and reinforcement learning (RL), for both natural and social science research, and innovative applications of deep learning in traditional numerical simulation. Its scope encompasses application development in scientific scenarios using HPC platforms; DL methods applied to numerical simulation; fundamental algorithms, enhanced procedures, and software development methods to enable scalable training and inference; hardware changes with impact on future supercomputer design; and machine deployment, performance evaluation, and reproducibility practices for DL applications with an emphasis on scientific usage. This workshop will be centered around published papers. Submissions will be peer-reviewed, and accepted papers will be published as part of the Joint Workshop Proceeding by Springer.

Topics include but are not limited to:

- DL as a novel approach of scientific computing
  - Emerging scientific applications driven by DL methods
  - Novel interactions between DL and traditional numerical simulation
  - Effectiveness and limitations of DL methods in scientific research
  - Algorithms and procedures to enhance reproducibility of scientific DL applications
- DL for Science workflows
  - Data management through the life cycle of scientific DL applications
  - DL performance evaluation and analysis on deployed systems
- Scalable DL methods to address the challenges of demanding scientific applications
  - General algorithms and procedures for efficient and scalable DL training
  - General algorithms and systems for large scale model serving for scientific use cases
  - New software, and enhancements to existing software, for scalable DL
  - DL communication optimization at scale
  - I/O optimization for DL at scale
  - DL performance modeling and tuning of DL on supercomputers
  - DL benchmarks on supercomputers
- Novel hardware designs for more efficient DL
  - Processors, accelerators, memory hierarchy, interconnect changes with impact on DL in the HPC context

As part of the reproducibility initiative, the workshop requires authors to provide information such as the algorithms, software releases, datasets, and hardware configurations used. For performance evaluation studies, we will encourage authors to use well-known benchmarks or applications with open accessible datasets: for example, MLPerf (https://github.com/mlperf/training) and ResNet-50 with the ImageNet-1K dataset (http://www.image-net.org/archive/stanford/fall11_whole.tar).

Important Dates.

- Papers due: Mar 31st, 2020
- Acceptance notification: Apr 16th, 2020
- Camera ready due: May 15th, 2020
- Workshop date: June 25th, 2020

Paper Submission. Authors are invited to submit unpublished, original work with a minimum of 6 pages and a maximum of 12 pages in single column text with LNCS style. All submissions should be in LNCS format (templates are available at http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines) and submitted using Linklings at https://ssl.linklings.net/conferences/isc_hpc/ tentatively.