Improved Neural Network Training: Layer-Parallelism, Least-squares and Initialization

Eric C. Cyr[a], Mamikon Gulian, Ravi Patel, Mauro Perego, Nat Trask (SNL), Stefanie Guenther (LLNL), Lars Ruthotto (Emory University), Jacob B. Schroder (UNM), Nico R. Gauger (TU Kaiserslautern)

[a] *Computational Mathematics Department, Sandia National Laboratories*

Deep neural networks are a powerful machine learning tool with the capacity to "learn" complex nonlinear relationships described by large data sets. Despite their success, training these models remains sensitive to parameters and is a computationally intensive undertaking. In this talk we will present two separate advances designed to tackle these challenges. In both cases, the approaches developed are based on numerical techniques that have been developed in the scientific computing community. Adapting these approaches to neural networks has yielded improvements both in the training methodology and in the technologies used to produce these improvements.

The first advance is a new layer-parallel training algorithm that exploits a multigrid scheme to accelerate both forward and backward propagation. Introducing a parallel decomposition between layers requires inexact propagation of the neural network. The multigrid method used in this approach stitches these subdomains together with sufficient accuracy to ensure rapid convergence to the same solution as the serial algorithm. We demonstrate an order of magnitude wall-clock speedup over the serial approach, opening a new avenue for scalable parallelism that is complementary to existing approaches. Results for this talk can be found in [1].

The second advance is motivated by taking an adaptive basis viewpoint of deep neural networks [2, 3, 4, 5]. This perspective leads to novel initializations and a hybrid least squares/gradient descent optimizer. We provide analysis of these techniques, and illustrate via numerical examples dramatic increases in accuracy and convergence rate for benchmarks characterizing applications of DNNs, including regression problems and physics-informed neural networks for the solution of partial differential equations.

Acknowledgement Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] S. Guenther, L. Ruthotto, J. B. Schroder, E. C. Cyr, N. R. Gauger, Layer-Parallel Training of Deep Residual Neural Networks, Accepted to SIMODs, 2019.

[2] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[3] Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. ReLU deep neural networks and linear finite elements. *arXiv preprint arXiv:1807.03973*, 2018.

[4] Daria Fokina and Ivan Oseledets. Growing axons: greedy learning of neural networks with application to function approximation. *arXiv preprint arXiv:1910.12686*, 2019.

[5] Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, and Qiang Qiu. Stochastic conditional generative networks with basis decomposition. *arXiv preprint arXiv:1909.11286*, 2019.