The 8th European Congress on Computational Methods in Applied Sciences and Engineering
ECCOMAS Congress 2022
5-9 June 2022, Oslo, Norway

Scalability analysis and performance modelling of layer-parallel training of deep residual networks using a non-linear multigrid-in-time algorithm

Chinmay Datar[1], **Harald Köstler**[12*]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, chinmay.datar@fau.de

[2] Zentrum für Nationales Hochleistungsrechnen Erlangen (NHR@FAU) Martensstraße 1, 91058 Erlangen, harald.koestler@fau.de

Keywords: *residual networks, multigrid-in-time, performance modelling*

The sequential propagation of data across the layers of the neural network causes the runtimes to scale linearly with the number of layers, and this creates a scalability barrier. In particular, the Residual Network (ResNet) architecture can be interpreted as the forward Euler discretization in time of an optimal control problem, where the trainable network parameters represent the dynamic control variables[1]. Formulating the serial propagation of data across the network as a system of equations, it can then be solved iteratively leveraging the highly parallel multigrid in time methods from optimal control. This is successfully demonstrated for classification tasks in [2]. In this talk, we show how to extend it to include regression tasks. One observes very small to negligible increments in runtimes, when the number of layers and computational resources are scaled by the same constant. In addition, fast convergence of this algorithm for networks with more than 1900 layers is demonstrated, and substantial speed-ups over the layer-serial version in experiments with classification and regression tasks. This is especially attractive for applications requiring increasingly deeper residual networks, such as image classification or networks approximating numerical simulations. An analytical communication model for the performance on distributed-memory architectures is also presented.

REFERENCES

[1] E. Haber and L. Ruthotto, *Stable architectures for deep neural networks*. Inverse Probl., 34 (2017), 014004, https://doi.org/10.1088/1361-6420/aa9a90

[2] S. Günther, L. Ruthotto, J. B. Schroder, E. C. Cyr, and N. R. Gauger, *Layer-parallel training of deep residual neural networks*, SIAM Journal on Mathematics of Data Science 2, no. 1 (2020): 1-23.