Contribution ID: **18** Type: **Regular talk**

Reduced Precision Strategies for Deep Learning: 3DGAN Use Case

Wednesday 21 October 2020 10:40 (20 minutes)

Deep learning simulations are known as computational heavy with the need of a lot of memory and bandwidth. A promising approach to make deep learning more efficient and to reduce its hardware workload is to quantize the parameters of the model to lower precision. This approach results in lower execution inference time, lower memory footprint and lower memory bandwidth.
We will research the effects of low precision inference of a deep generative adversarial network [1] model which consists of a convolutional neural network. The use case is for calorimeter detector simulations of subatomic particle interactions in accelerator based high energy physics. We are comparing the inference results of the generated electron showers with the training data for different numerical bit formats and benchmark these in terms of computation and physics accuracy. The model we are quantizing, is a modified 3DGAN [2] prototype based on 2D convolutional layers. With this prototype we gained a factor 3 runtime speed up.
+

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks,"2014.
[2] G. r. Khattak, S. Vallecorsa, and F. Carminati, "Three dimensional energy parametrized generative adversarial networks for electromagnetic shower simulation,"in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 3913–3917.

Authors: Mr REHM, Florian (Hochschule Coburg (DE)); Mrs VALLECORSA, Sofia

Co-authors: Mr SALETORE, Vikram; PABST, Hans; CHAIBI, Adel

Presenter: Mr REHM, Florian (Hochschule Coburg (DE))

Session Classification: Workshop

Track Classification: 3 ML for simulation and surrogate model : Application of Machine Learning to simulation or other cases where it is deemed to replace an existing complex model