Contribution ID: **264** Type: **talk**

Fast RNN Inference on an FPGA

Tuesday, 13 July 2021 15:00 (15 minutes)

The hls4ml library [1] is a powerful tool that provides automated deployment of ultra low-latency, low-power deep neural networks. We extend the hls4ml library to recurrent architectures and demonstrate low latency by considering multiple benchmark applications. We consider Gated Recurrent Units (GRU) and Long Short Term Memory(LSTM) Models trained using the CERN Large Hadron Collider Top tagging data [2], jet flavor tagging data [3], and the quickdraw dataset as our benchmark applications. By using a large parameter range in between these benchmark models, we demonstrate that low-latency inference across a wide variety of model weights, and show that resource utilization of recurrent neural networks can be significantly reduced with little loss to model accuracy.

Reference:
[1] J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics",JINST13(2018), no. 07,P07027, doi:10.1088/1748-0221/13/07/P07027,arXiv:1804.06913
[2] CERNbox, .https://cernbox.cern.ch/index.php/s/AgzB93y3ac0yuId?path=%2F, 2016
[3] Guest, Daniel, et al. "Jet flavor classification in high-energy physics with deep neural networks." Physical Review D 94.11 (2016): 112002.
[4] Google, "Quick, Draw!", https://quickdraw.withgoogle.com/

Are you are a member of the APS Division of Particles and Fields?

No

Primary authors: WANG, Aaron; Mr PAIKARA, Chaitanya (University of Washington); KHODA, Elham E (University of Washington (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); SUMMERS, Sioni Paris (CERN)

Presenter: WANG, Aaron

Session Classification: Computation, Machine Learning, and AI

Track Classification: Computation, Machine Learning, and AI