Abstract Submitted
for the APR21 Meeting of
The American Physical Society

Fast RNN Inference on an FPGA CHAITANYA PAIKARA, University of Washington, PHILIP HARRIS, Massachusetts Institute of Technology, SCOTT HAUCK, SHIH-CHIEH HSU, RICHA RAO, University of Washington, SIONI SUMMERS, European Organization for Nuclear Research (CERN), UNIVERSITY OF WASHINGTON COLLABORATION, MASSACHUSETTS INSTITUTE OF TECHNOLOGY COLLABORATION, EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN) COLLABORATION — In this work, we will present the implementation templates for two types of recurrent neural network layers within the HLS4ML library Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). These templates provide the lower-level hardware implementation for a neural network model based on these layers, allowing them to be mapped to an FPGA. Using the HLS4ML library, the latency per inference and resource utilization can be adjusted for the targeted FPGA, and the application requirements. Several particle physics problems are used to characterize the template and test its efficiency after the High-Level Synthesis. Design space exploration was performed across different features - resource utilization, latency, model performance, and fixed-point precision. As an example, LSTM and GRU based models for the task of jet identification on simulated proton-proton collision at the Large Hadron Collider were considered. Also, the implementation templates were evaluated against varying numbers of model parameters, and synthesized for larger neural network models based on LSTM and external recursion for jet flavor classification in high energy collision for an FPGA.

Chaitanya Paikara
University of Washington

Date submitted: 02 Jan 2021

Electronic form version 1.4