


From: Riccardo Travaglini Riccardo.Travaglini@bo.infn.it 
Subject: alcuni aggiornamenti sulle attivita' di (Deep) Machine Learning su FPGA
Date: 18 July 2019 at 15:39



To: G.Marco Dallavalle Marco.Dallavalle@bo.infn.it, Daniele Bonacorsi Daniele.Bonacorsi@bo.infn.it, Luigi Guiducci Luigi.Guiducci@bo.infn.it, Andrea Perrotta Andrea.Perrotta@bo.infn.it, Casimiro Baldanza casimiro.baldanza@bo.infn.it, Fabrizio Odorici Fabrizio.Odorici@bo.infn.it, Alessandro Montanari montanari@bo.infn.it, Graziano Bruni graziano.bruni@bo.infn.it, Vincenzo Vagnoni Vincenzo.Vagnoni@bo.infn.it, Gaetano Maron Gaetano.Maron@Inl.infn.it, Davide Falchieri@bo.infn.it, Gabriele Balbi@bo.infn.it, Renato Campanini renato.campanini@bo.infn.it, Francesca Del Corso francesca.delcorso@bo.infn.it, Alberto Cervelli alberto.cervelli@bo.infn.it, Alessia D'Orazio alessia.dorazio@bo.infn.it, Nico Giangiacomini@bo.infn.it, Roberto Giacomelli Roberto.Giacomelli@bo.infn.it, Fabrizio Alfonsi@bo.infn.it
Cc: G.Marco Dallavalle marcovd@bo.infn.it, Tommaso Tommaso.Diotalevi@bo.infn.it

Cari colleghi,

vi informiamo sullo stato di un primo esercizio di implementazione di algoritmo di Machine learning su FPGA che noi stiamo portando avanti e sugli strumenti che stiamo utilizzando.

L'esercizio consiste in una rete neurale di 3 layers (1 input, 1 hidden, 1 output) per la classificazione (in 3 classi) di un dataset noto in ML (<https://archive.ics.uci.edu/ml/datasets/Iris>) contenente caratteristiche di 150 fiori appartenenti a 3 specie diverse di Iris. E' un esercizio relativamente semplice ma il dataset e' un benchmark molto noto nella comunita' dei data scientists e abbiamo pensato fosse ideale per provare metodi e strumenti prima di affrontare problemi specifici.

Per la realizzazione e il training della rete abbiamo usato il framework Keras (<https://keras.io/>) in Python.

Come ulteriore strumento stiamo usando due prodotti simili tra loro: Jupyter Notebook (<https://jupyter.org/>) e Google Colaboratory (<https://colab.research.google.com/notebooks/welcome.ipynb#recent=true>) che permettono di sviluppare documenti web-based interattivi composti, non solo, da codice python eseguibile e commenti testuali.

Il primo passo che stiamo concludendo e' di implementare la rete, fare il training ed esportare i files necessari per eseguire il tool HLS4ML (<https://hls-fpga-machine-learning.github.io/hls4ml/>). Questo strumento e' sviluppato da un gruppo di CMS basato al CERN e permette, tra l'altro, di tradurre reti neurali sviluppate in Keras in codice di alto livello (C++) che puo' essere sintetizzato in FPGA della Xilinx utilizzando il programma Vivado HLS.

Al momento abbiamo un "notebook" in grado di fare il training della rete, misurare alcune metriche di performance ed esportare i file corrispondenti (architettura, pesi, file di test). Stiamo verificando la conversione con HLS4ML.

Alcune informazioni aggiuntive:

- Il codice su notebook della rete neurale e' parte del materiale delle esercitazioni del corso che Daniele Bonacorsi tiene al dottorato in Data Science and Computation, e alla magistrale in Bioinformatics;
- io, Daniele Bonacorsi e Tommaso Diotallevi abbiamo seguito un tutorial su HLS4ML a Zurigo a Febbraio 2019 in cui veniva mostrato l'utilizzo su alcune semplici reti per la categorizzazione di eventi di fisica;
- non abbiamo ancora stabilito su quale FPGA provare l'esercizio; il sistema ATCA e' appena arrivato a Bologna ma non ancora operativo; per questo stiamo valutando se provarlo su una FPGA di classe medio-bassa o su Cloud FPGA.

Da ultimo, non e' semplice per noi organizzare in questo periodo estivo una presentazione o seminario in proposito, a vantaggio di tutti i membri di questo gruppo di lavoro, ma vi terremo aggiornati in lista e nei meeting sui progressi, e siamo sempre disponibili per commenti o chiarimenti.

Riccardo, Daniele e Tommaso

--

Dr. Riccardo Travaglini
c/o I.N.F.N
(Italian Institute of Nuclear Physics)
v.le B. Pichat 6/2
40127 Bologna
Italy

phone: +39 051 20 95275
fax : +39 051 20 95286
