A Framework for Semi-Automatic Precision and Accuracy Analysis for Fast and Rigorous Deep Learning

Anastasia Volkova*[1] and Christoph Lauter[3,2]

[1]Laboratoire des Sciences du Numérique de Nantes – Université de Nantes, Université de Nantes – France

[3]Sorbonne Université – Sorbonne Université, CNRS, LIP6 – France

[2]University of Anchorage in Alaska – États-Unis

Résumé

Deep Neural Networks (DNN) represent a performance-hungry application. Floating-Point (FP) and custom floating-point-like arithmetic satisfies this hunger. While there is need for speed, inference in DNNs does not seem to have any need for precision. Many papers experimentally observe that DNNs can successfully run at almost ridiculously low precision. The aim of this paper is twofold: first, to shed some theoretical light upon why a DNN's FP accuracy stays high for low FP precision. We observe that the loss of relative accuracy in the convolutional steps is recovered by the activation layers, which are extremely well-conditioned. We give an interpretation for the link between precision and accuracy in DNNs. Second, the paper presents a software framework for semi-automatic FP error analysis for the inference phase of deep-learning. Compatible with common Tensorflow/Keras models, it leverages the frugally-deep Python/C++ library to transform a neural network into C++ code in order to analyze the network's need for precision. This rigorous analysis is based an Interval and Affine arithmetics to compute absolute and relative error bounds for a DNN. We demonstrate our tool with several examples.

*Intervenant