ACAT 2021

Contribution ID: **599** Contribution code: **contribution ID 599** Type: **Oral**

GPU Acceleration of Automatic Differentiation in C++ with Clad

Tuesday, 30 November 2021 18:40 (20 minutes)

Automatic Differentiation (AD) is instrumental for science and industry. It is a tool to evaluate the derivative of a function specified through a computer program. The range of AD application domain spans from Machine Learning to Robotics to High Energy Physics. Computing gradients with the help of AD is guaranteed to be more precise than the numerical alternative and have at most a constant factor (4) more arithmetical operations compared to the original function. Moreover, AD applications to domain problems typically are computationally bound. They are often limited by the computational requirements of high-dimensional transformation parameters and thus can greatly benefit from parallel implementations on graphics processing units (GPUs).

Clad aims to enable differentiable analysis for C/C++ and CUDA and is a compiler-assisted AD tool available both as a compiler extension and in ROOT. Moreover, Clad works as a compiler plugin extending the Clang compiler; as a plugin extending the interactive interpreter Cling; and as a Jupyter kernel extension based on xeus-cling.

In this talk, we demonstrate the advantages of parallel gradient computations on graphics processing units (GPUs) with Clad. We explain how to bring forth a new layer of optimisation and a proportional speed up by extending the usage of Clad for CUDA. The gradients of well-behaved C++ functions can be automatically executed on a GPU. Thus, across the spectrum of fields, researchers can reuse their existing models and have workloads scheduled on parallel processors without the need to optimize their computational kernels. The library can be easily integrated into existing frameworks or used interactively, and provides optimal performance. Furthermore, we demonstrate the achieved application performance improvements, including (˜10x) in ROOT histogram fitting and corresponding performance gains from offloading to GPUs.

Significance

References

Speaker time zone

No preference

Primary authors: IFRIM, Ioana (Princeton University (US)); VASILEV, Vasil Georgiev (Princeton University (US))

Presenter: IFRIM, Ioana (Princeton University (US))

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research