Contribution ID: **126** Type: **Short Talk**

C++ Code Generation for Fast Inference of Deep Learning Models in ROOT/TMVA

Tuesday, May 18, 2021 10:50 AM (13 minutes)

We report the latest development in ROOT/TMVA, a new system that takes trained ONNX deep learning models and emits C++ code that can be easily included and invoked for fast inference of the model, with minimal dependency. We present an overview of the current solutions for conducting inference in C++ production environment, discuss the technical details and examples of the generated code, and demonstrates its development status with a preliminary benchmark against popular tools.

Primary authors: AN, Sitong (CERN, Carnegie Mellon University (US)); MONETA, Lorenzo (CERN)

Presenter: AN, Sitong (CERN, Carnegie Mellon University (US))

Session Classification: Artificial Intelligence

Track Classification: Offline Computing