Contribution ID: **166** Type: **Oral presentation**

Keras2c: A simple library for converting Keras neural networks to real-time friendly C

Wednesday 14 October 2020 13:35 (20 minutes)

With the growth of machine learning models and neural networks in measurement and control systems comes the need to deploy these models in a way that is compatible with existing systems. Existing options for deploying neural networks either introduce very high latency, requires expensive and time consuming work to integrate into existing code bases, or only support a very limited subset of model types. We have therefore developed a new method, called Keras2c, which is a simple library for converting Keras/TensorFlow neural network models into real time compatible C code. It supports a wide range of Keras layer and model types, including multidimensional convolutions, recurrent layers, well as multi-input/output models, and shared layers. Keras2c re-implements the core components of Keras/TensorFlow required for predictive forward passes through neural networks in pure C, relying only on standard library functions. The core functionality consists of only ~1200 lines of code, making it extremely lightweight and easy to integrate into existing codebases. Keras2c has been sucessfully tested in experiments and is currently in use on the plasma control system at the DIII-D National Fusion Facility at General Atomics in San Diego.

Minioral

Yes

IEEE Member

Yes

Are you a student?

Yes

Author: CONLIN, Rory (Princeton University)

Co-authors: ERICKSON, Keith (Princeton University); ABBATE, Joe (Princeton University); KOLEMEN, Egemen

Presenter: CONLIN, Rory (Princeton University)

Session Classification: Oral presentations RTA01

Track Classification: Deep Learning and Machine Learning