ELL: the Microsoft Embedded Learning Library

Byron Changuion, Principal Engineering Manager, Embedded Learning Library, Machine Learning and Optimization Group, Microsoft Research AI

The Microsoft Embedded Learning Library (ELL) is a cross-platform open source framework, designed and developed by Microsoft Research over the past three years. ELL is part of Microsoft's broader efforts around intelligent edge computing.

At its core, ELL is an optimizing cross-compiler toolchain for AI pipelines, geared towards small resource-constrained devices and microcontrollers. ELL takes as input an end-to-end AI pipeline, such as an audio keyword detection pipeline or a vision-based people counting pipeline. It compresses the model and generates optimized machine executable code for an embedded target platform. AI pipelines compiled with ELL run locally on the target platform, without requiring a connection to the cloud and without relying on other runtime frameworks. While ELL can generate code for any platform, it is primarily optimized for standard off-the-shelf processors, such as the ARM Cortex A-class and M-class architectures that are prevalent in single-board computers. In addition to its functionality as a compiler, the ELL project provides an online gallery of pre-trained AI models and a handful of tutorials written for makers, technology enthusiasts, and developers who aspire to build intelligent devices and AI-powered gadgets.

In this talk, the vision behind ELL will be presented, and design, scope, and roadmap for the future will be presented. Design considerations that led Microsoft to build an AI compiler, rather than the more conventional choice of building an AI runtime will be addressed. In addition this talk will explain how Microsoft was able to move past the standard academic metrics, instead using real-world criteria to guide our work and priorities.