Bringing TinyML to RISC-V With Specialized Kernels and a Static Code Generator Approach

2022/06/21

@embedded world 2022, 21.6.2022 11:00: https://www.embedded-world.de/en/conferences-programme/programme-overvie... [1]

Rafael Stahl (Technical University of Munich)

Ultra-low-power deep learning, also known as TinyML, has been successfully implemented for applications such as keyword spotting, anomaly detection or gesture recognition. Low-power microcontrollers are a widely-used TinyML target, where especially RISC-V gained interest due to its extensible and scalable ISA. We present an efficient TinyML RISC-V backend for TensorFlow Lite for Microcontrollers (TFLM). Firstly, it uses a specialized kernel library muriscv_nn to exploit the V- and P-Extensions of RISC-V. Secondly, it avoids the overheads of the so-called TFLM interpreter by introducing an additional code generator step that generates a hard-coded inference code for the input model. The beneficial effect of code generation and specialized RISC-V kernels is demonstrated through simulation. For TinyMLPerf models, the P-Extension optimized kernels have 6.5x reduced number of executed instructions compared to the TFLM reference kernels. The static code generation approach reduces their RAM usage by 1.5x and ROM usage by 1.4x.

Das Projekt Scale4Edge wird unter den Förderkennzeichen 16ME0122K-140, 16ME0465, 16ME0900, 16ME0901 im Förderprogramm ZuSE durch das deutsche Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Links:
[1] https://www.embedded-world.de/en/conferences-programme/programme-overview?lectureId=Ges1PhMIDTqRxBWsWZGz