“Low-cost neural network inferencing on the edge with xcore.ai”

Laszlo Kindrat – XMOS
May 28, 2020
tinyML Talks Sponsors

Edge Impulse

www.edgeimpulse.com

Additional Sponsorships available – contact Bette@tinyML.org for info
# Next tinyML Talk

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday</td>
<td><strong>Igor Fedorov</strong></td>
<td>SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers</td>
</tr>
<tr>
<td>June 9</td>
<td>ARM Machine Learning Lab,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arm ML Research</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Dominic Binks</strong></td>
<td>tinyML doesn’t need Big Data, it needs Great Data</td>
</tr>
<tr>
<td></td>
<td>VP Technology,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Audio Analytic Ltd.</td>
<td></td>
</tr>
</tbody>
</table>

Webcast start time is 8 am Pacific time
Each presentation is approximately 30 minutes in length

Please contact talks@tinyml.org if you are interested in presenting
Laszlo Kindrat

Laszlo Kindrat is currently in the role of Machine Learning Engineer at XMOS in Hampton, NH. He has aided the design of the XS3 architecture's vector unit and plays a key role in developing inference tools for the xCORE-AI platform. Laszlo also has experience with traditional and machine learning algorithm development. He holds a PhD in applied mathematics from the University of New Hampshire.
XCORE.AI CHIP ARCHITECTURE

![Diagram of XCORE.AI Chip Architecture]

- **IO pins**
  - Hardware ports
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - ALU (FP + int) vector unit
  - SRAM
  - High-Speed USB PHY

- **xtime scheduler**
  - Core PLL
  - App PLL
  - JTAG

- **switch**

- **external LPDDR interface**

- **IO pins**
  - Hardware ports
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - xcore logical core
  - ALU (FP + int) vector unit
  - SRAM
  - MIPI D-PHY

- **OTP**
  - High-Speed USB PHY
  - MIPI D-PHY

- **High-Speed USB PHY**

- **MIPI D-PHY**
XCORE.AI VECTOR UNIT

- 256 bit vector registers, single cycle memory access
- Native support for int8/16/32 and binarized modes
- Most instructions perform load + arithmetic operation
- Scalar instructions (e.g. pointer math) can be dual issued
- Long uninterrupted sequence of VLMACC(R) ops yield high arithmetic density

VLMACCR: Vector Load and MACC reduce

MEMORY

REGISTER

ACCUMULATOR

RESULT

BUFFER

MEMORY
MODEL DEPLOYMENT WORKFLOW

TensorFlow
Trained floating point network

Lite converter (python API)

Run TFL to xcore.ai converter

my_model.tflite
lib_xs3_ai

xcore.ai

PyTorch
Caffe2
mxnet

ONNX

my_model.onnx

Alternative framework flow

ONNX component
User component
XMOS component
TensorFlow component

Key
DEMO TIME!
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org