Accelerated Neural Networks on OpenCL Devices Using SYCL-DNN

John Lawson
Project lead SYCL-DNN

IWOCL — May 2019
ImageNet classification error

ImageNet Large Scale Visual Recognition Challenge image-net.org/challenges/LSVRC.
2012: Alexnet 60 million parameters
2014: VGG  138 million parameters
2014: GoogLeNet  5 million parameters
Inception block

Diagram of an Inception block with various convolutional layers and a concatenation layer.
2014: GoogLeNet  5 million parameters
2015: InceptionV2  23 million parameters
2015: ResNet-50  25 million parameters
Resnet block

Add

1x1/s1

3x3/s1

1x1/s1

\[ \text{Codeplay Software Ltd.} \]
2015: ResNet-50  25 million parameters
2015: ResNet-152  60 million parameters
Accelerator adoption  TOP500 supercomputers

Percent share in top500

- AMD
- IBM
- Intel
- NVIDIA
- Other
- PEZY

Codeplay Software Ltd.
**arm** COMPUTE LIBRARY

Intel MKL-DNN

AMD MiOpen

cuDNN
Royalty free open standards from the Khronos Group.
SYCL Ecosystem

Applications → C++ Libraries → SYCL for OpenCL → OpenCL → OpenCL enabled accelerators
SYCL Ecosystem

Applications

C++ Libraries

SYCL for OpenCL

OpenCL

OpenCL enabled accelerators

Machine learning model

TensorFlow, PyTorch, …

SYCL-DNN, SYCL-BLAS, Eigen, …
The SYCL-DNN neural network acceleration library.

<table>
<thead>
<tr>
<th>sycl</th>
<th>cpp</th>
<th>cpp11</th>
<th>cplusplus</th>
<th>opengl</th>
<th>gpgpu</th>
<th>machine-learning</th>
<th>neural-network</th>
</tr>
</thead>
</table>

- 239 commits
- 1 branch
- 5 releases
- 3 contributors
- Apache-2.0
Supported operations

2D convolutions
2D depthwise convolutions
2D max & average pooling
Relu & tanh activations
2D Convolution
Highly parameterized kernels

```cpp
template <typename T, typename Index, typename ConvType, int TileRows, int TileCols, int ChannelVectorWidth, int FeatureVectorWidth, bool UseFastDiv, int WindowRows, int WindowCols, int Stride>
struct TiledConv2D;

template <typename T, typename Index, int ChannelVector, int M, int N, int R, int S, typename ConvType>
struct WinogradInputTiles;
```
Different algorithms  First layers in ResNet

Run on AMD R9 Nano, batch size 32.
Performance on ARM HiKey 960

ARM NEON and OpenCL from ARM Compute Library v18.11.
Performance on Intel i7-6700K

MKL-DNN v1.0-pc running on CPU.
Still to come

Performance improvements
Quantized integer support
Integration of third party libraries
More operations

…