

5/11/2022

@qualcomm

IWOCL 2022

Qualcomm

Machine Learning on Mobile with OpenCL

Balaji Calidas

Qualcomm Technologies Inc.



Overview

- Machine Learning on Mobile is a rapidly growing area.
 - Increase in number of use cases.
 - Increase in research references.
- GPUs remain a popular option for ML on mobile
 - A large number of ML frameworks and compilers use OpenCL
 - TFLite, SNPE, MNN, Mace, Paddle-Lite and TVM
- Currently most ML use cases are focused on inference.
 - We see training at the edge as an emerging area.
- Important Considerations for ML on mobile
 - Low Power consumption, low latency dispatch and synchronization
 - Zero Copy Data Import/Export (Android Hardware Buffer, DMA-BUF)
 - Reduced memory footprint (reuse memory across layers when possible)

cl_qcom_ml_ops

- Extension for Accelerating Machine Learning on Adreno GPUs at the Op (Metacommand) Level.
 - Shipping from Snapdragon 888 onwards with inference support.
 - Training support added on Snapdragon 8 Gen 1
 - Integrated with TVM BYOC (See our poster at IWOCCL 2022)
- Uses existing OpenCL constructs such as command queues, events and buffers
 - Adds new constructs such as ops and tensors.
- Fully interoperable with other OpenCL kernels
 - Inline execution, synchronization and data sharing.
 - Vendor provided Ops can deliver a significant performance advantage.
- Samples and documentation available from <https://developer.qualcomm.com/software/adreno-gpu-sdk/tools>

Training with `cl_qcom_ml_ops`

- Use cases for edge training include Transfer Learning, Personalization and Federated Learning
- Memory footprint is a major consideration for training on mobile devices
 - The memory needed for all of the weights, gradients and activations for Mobilenet can be around 2 GB.
- `cl_qcom_ml_ops` leverages the Tensor Batch 1 approach for significantly reduced memory footprint.
 - Activations and gradients have a tensor batch size of 1. Gradients are accumulated, then applied at the end of the batch.
 - If Batch Normalization layers are present, the statistics are frozen.
 - This approach works well for Transfer Learning/Personalization use cases.
 - Additional training approaches will be supported going forward.

Qualcomm Extensions for ML

- `cl_qcom_dot_product8`
 - OpenCL C builtins for 8 bit dot product with saturating accumulate
 - Useful for accelerating 8 bit quantized DNNs.
- `cl_qcom_recordable_queues`
 - Record a sequence of `EnqueueNDRangeKernel` commands
 - Replay recording with optional updates to kernel arguments
 - Significant improvements in dispatch latency and CPU power consumption for enqueue of Machine Learning models.
 - Especially useful for streaming mode ML use cases.
- Extensions related to zero copy (DMA-BUF/AHB import), subgroup operations, subgroup size control.

Upcoming ML related Extensions from Khronos

- `cl_khr_integer_dot_product`
- Command Buffer Recording and Replay (provisional)
 - `cl_khr_command_buffer`
 - `cl_khr_mutable_dispatch`
- `cl_ext_float_atomics` (roadmap)
- Generalized Image from buffer
 - `cl_ext_image_from_buffer`
- Extended Vectors (roadmap)
- Semaphores (provisional)
 - `cl_khr_semaphore`, `cl_khr_external_semaphore`, `cl_khr_external_semaphore_sync_fd`

Summary

- Machine Learning on mobile is an important and growing technology area
- Qualcomm will continue to invest in extensions that accelerate Machine Learning with OpenCL.



Thank you

Follow us on: [f](#) [🐦](#) [in](#) [@](#)

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark of Qualcomm Incorporated, registered in the United States and other countries. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.