Deep Learning Optimization for Whole Slide Image Analysis in Low-Resource Environments

Siddhesh Thakur

Data Engineer

Indiana University School of Medicine

Geeks-Sid
thakursp@iu.edu

Disclosures

Speaker Name: Siddhesh P. Thakur

I have nothing to disclose

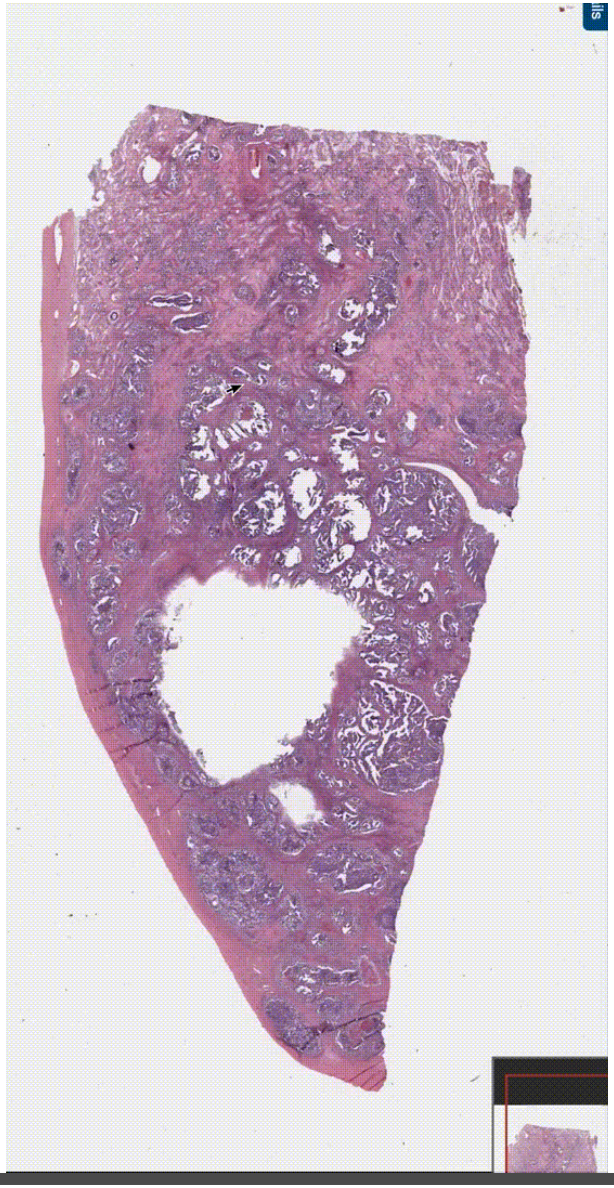
Geeks-Sid
thakursp@iu.edu

Motivation

- Computational Whole Slide Image (WSI) Analysis is demanding
  - High compute requirement
  - Hardware dependent
- Clinical environments are considered low-resource environments
  - Consumer grade CPUs workstations
  - Not considering GPUs, AI accelerators == unnecessary expense
  - Countries representing underserved population cannot afford
- Focusing on Delineation
  - A tedious manual task
  - Clinical experts are already overworked by assessing health system generated data

Geeks-Sid
thakursp@iu.edu

Optimization of DL for Clinical Environments



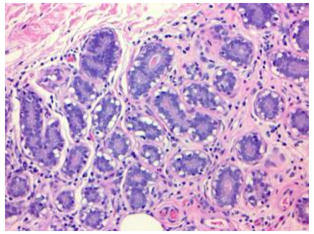
Deep Learning is expensive $$$

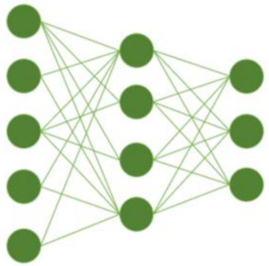
Price tag : $40000 only

But what if it wasn't?

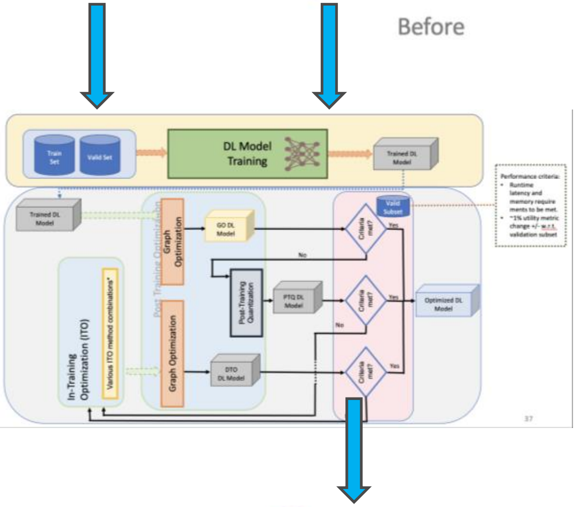
Price tag : $800 only

Proposed Solution





Before



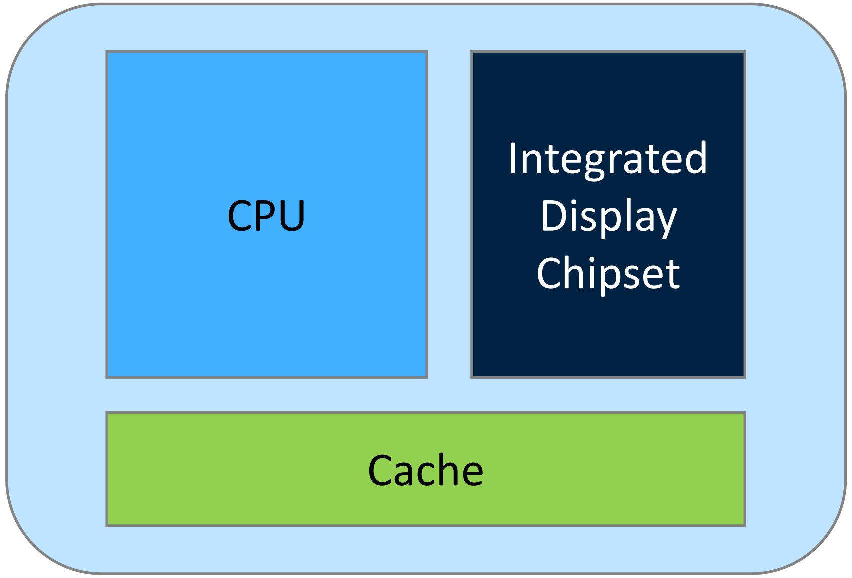
Smaller Model

Geeks-Sid
thakursp@iu.edu

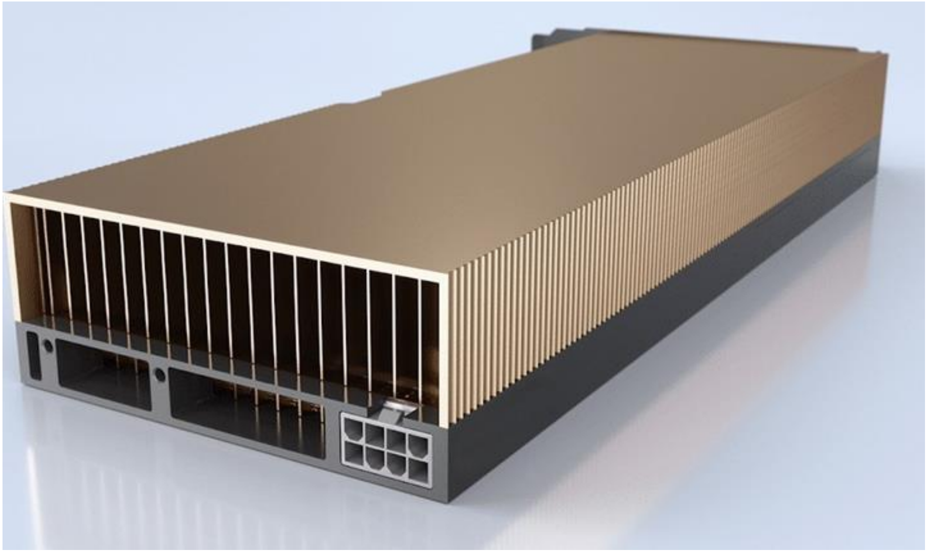
Integrated vs Discrete Graphics

Integrated
- within the CPU chipset
- requires no additional purchase

Discrete
- separate hardware component.
- requires additional purchase ($$$$)



CPU-integrated display chipset (**iGPU**)

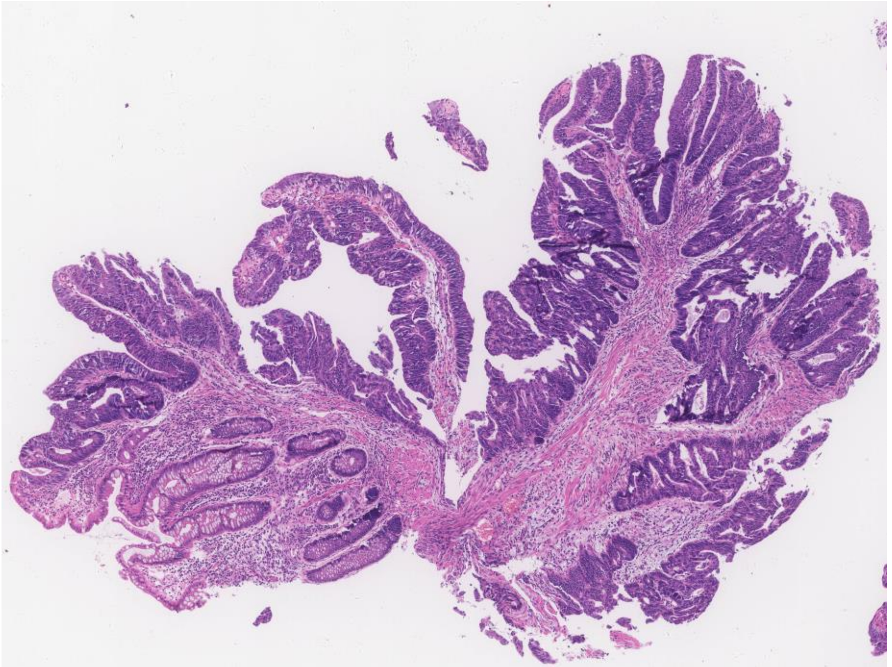


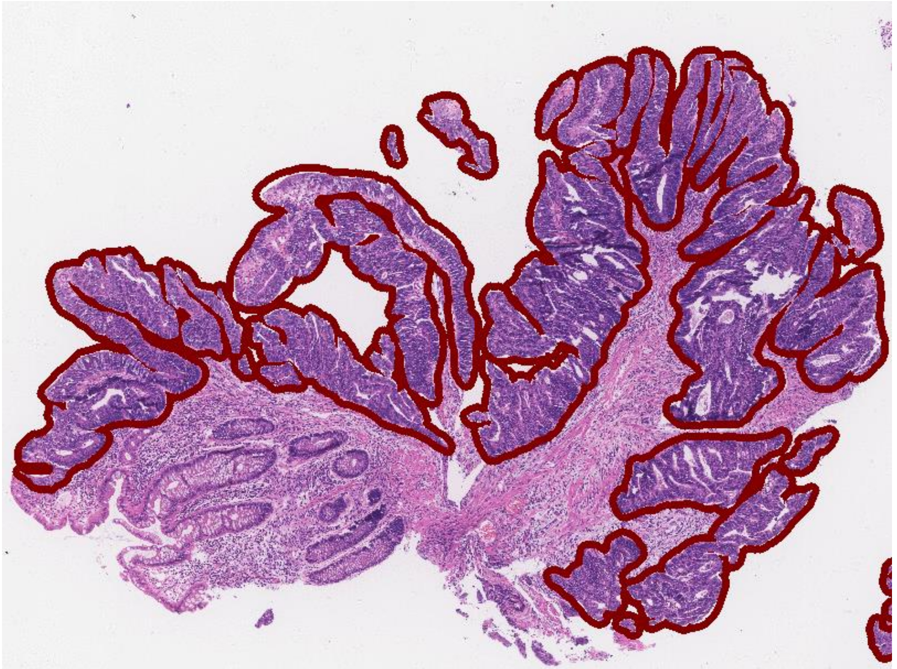
*PyTorch Models are supported on CPUs or discrete GPUs.

Data

- DigestPath[1] 2019 Dataset –
  - Multi-Institutional Data
  - Public competition
  - Colorectal adenocarcinoma (H&E)

Dataset	Number of slides	Number of patches
Train	200	100000
Test(Hold-out)	50	25000
Total	250	125000





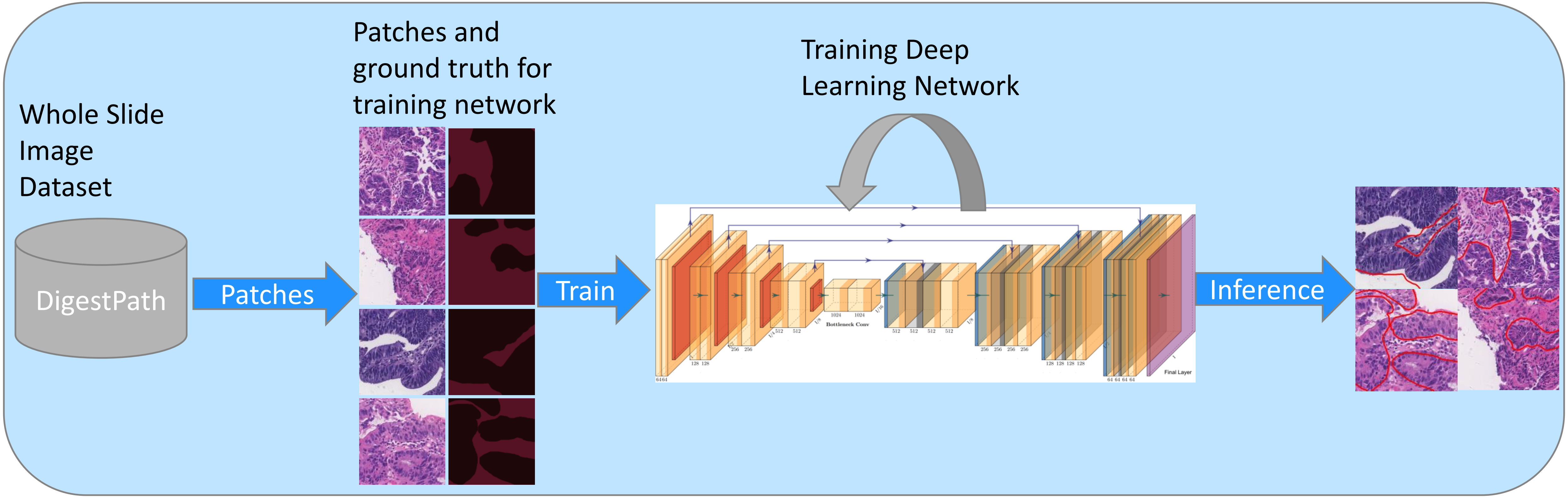
Example of Digitized Tissue section on the left and cancer delineation in red on the right

Geeks-Sid
thakursp@iu.edu

[1] Li, J., Yang et al. International Conference on Information Processing in Medical Imaging, Springer, Cham, (842-854)2019.

Methodology

- 2D UNet with residual connections
- Zero-code use, through the Generally Nuanced Deep Learning Framework (GaNDLF)[2]
- Models were rigorously trained for multiple hyperparameters
- Optimal picked following 5-fold cross-validation

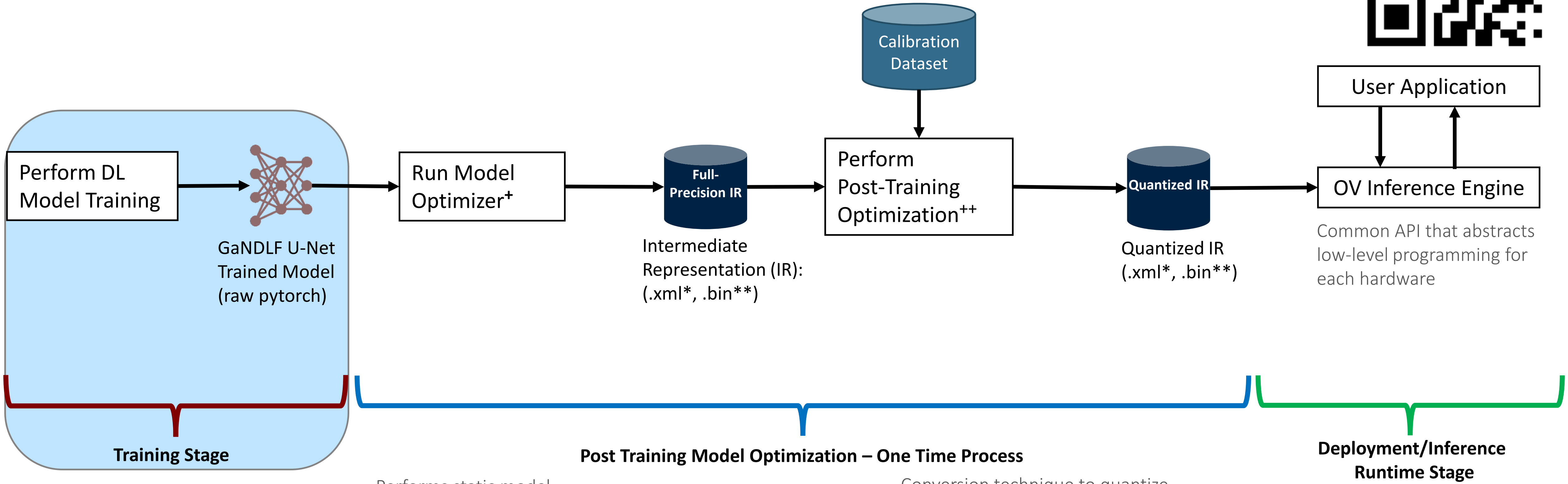


Pipeline for training a Deep Learning Model for Whole Slide images through GaNDLF

Geeks-Sid
thakursp@iu.edu

[2] Pati et al, GaNDLF: A Generally Nuanced Deep Learning Framework for Scalable End-to-End Clinical Workflows in Medical Imaging

Optimization via GaNDLF



Performs static model analysis and adjusts deep learning models for optimal execution on end-point target devices.

Conversion technique to quantize models to low precision for high performance. Reduces model size while also improving latency, with little degradation in model accuracy and <u>without model re-training.</u>

Geeks-Sid
thakursp@iu.edu

+ https://docs.openvinotoolkit.org/latest/openvino_docs_MO_DG_Deep_Learning_Model_Optimizer_DevGuide.html
*.xml – describes the network topology **.bin – describes the weights and biases (binary data)
++ https://docs.openvinotoolkit.org/latest/pot_README.html

Details about Optimization

- Optimization strategies
  - Optimizing topologies which include
    - Node Merging
    - Layer Fusion

    (Conv + BN + ReLU → ConvBNReLU)
    - Horizontal Fusion
  - Optimized kernels
    - -CPU Inststruction set specific kernels
  - Operator fusion
  - Folds constant paths in graph
  - Residual optimization
- Quantization
  - INT8 quantization
  - FP16 quantization

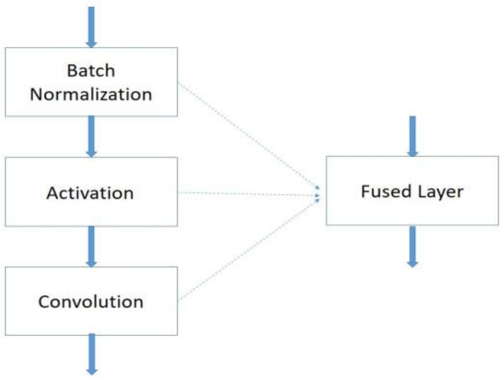


Figure F: Example of Layer Fusion

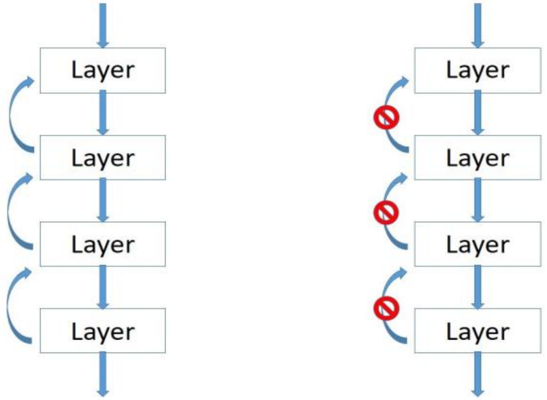


Figure G: Layer freezing for Residual optimization

Details about Quantization

QUANTIZATION

A technique to reduce memory consumption and computation time of deep neural networks by lowering the precision of parameters

Increases energy efficiency (Watts)

Can accelerate workloads, but caution is essential to preserve model accuracy

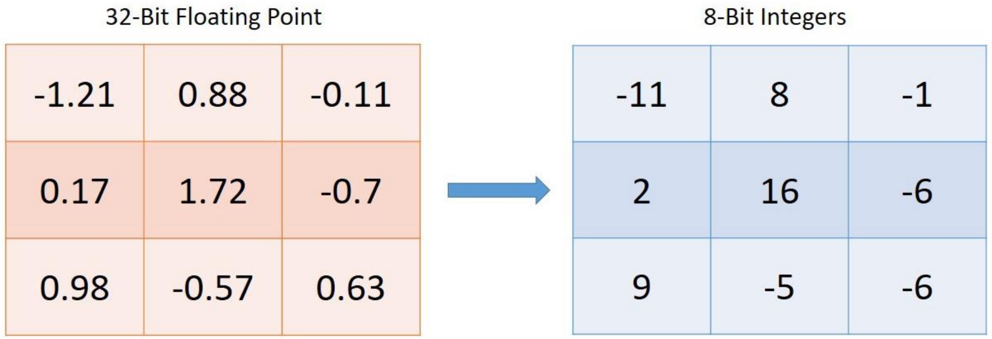


Figure I : Example of Range mapping in quantization

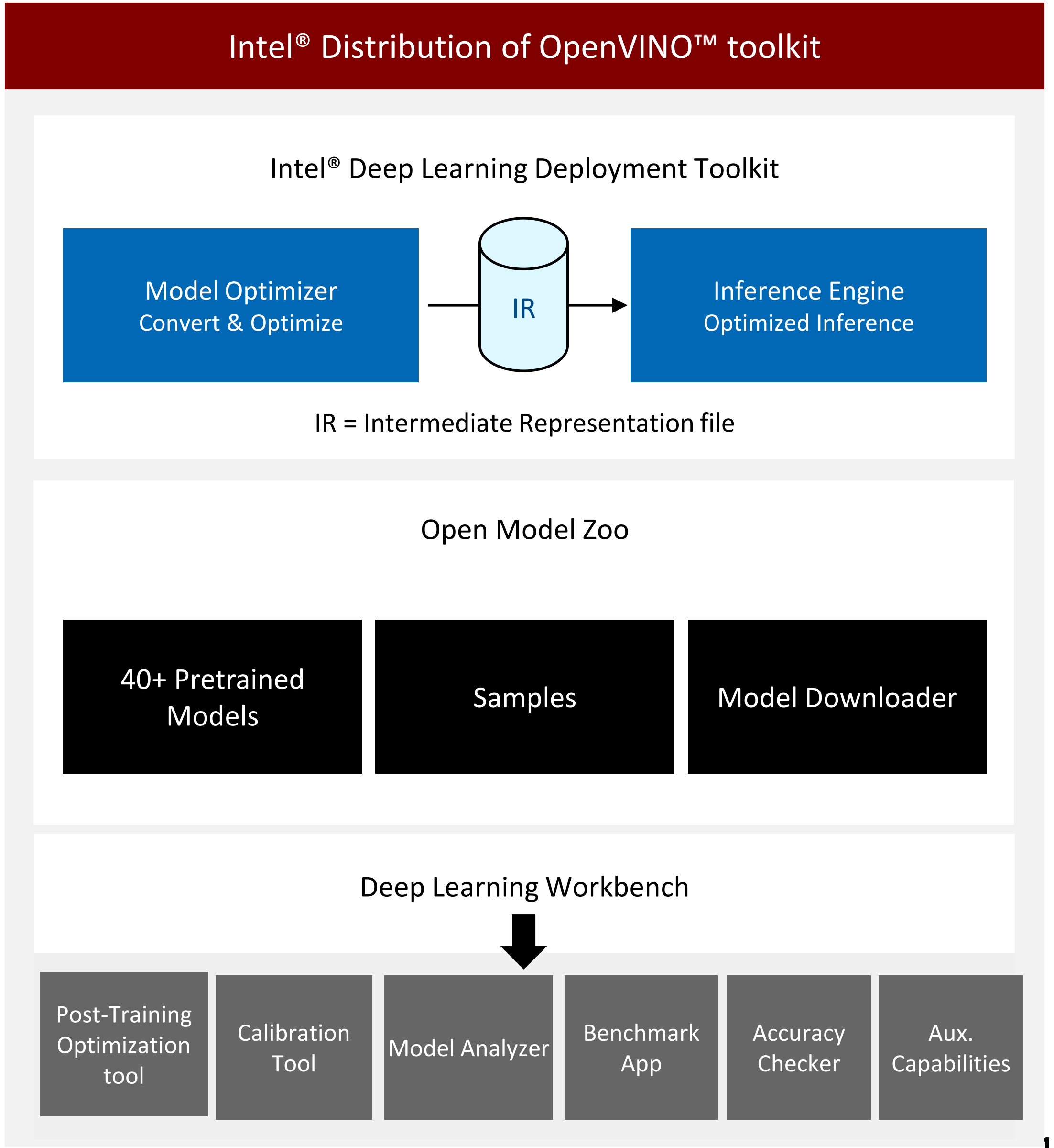
Figure H : Full Precision to INT8 conversion

Operation	Multiplication (Factor)	Addition (Factor)
INT8	1x	1x
INT32	12x	3.33x
FP16	5x	13.33x
FP32	16x	30.0x

Table 1 : Energy utilization(watts) of operations*

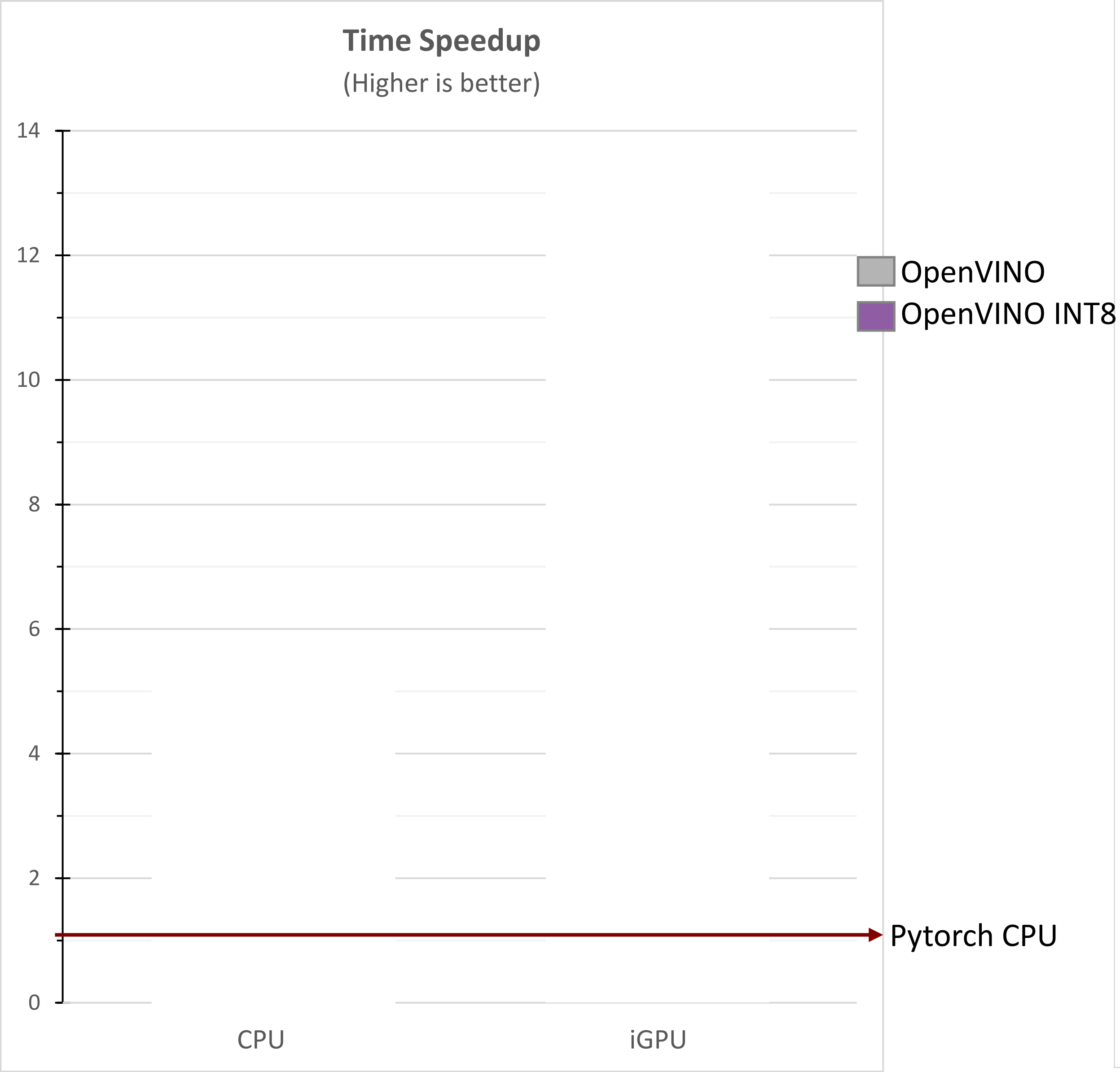
OpenVINO™ toolkit

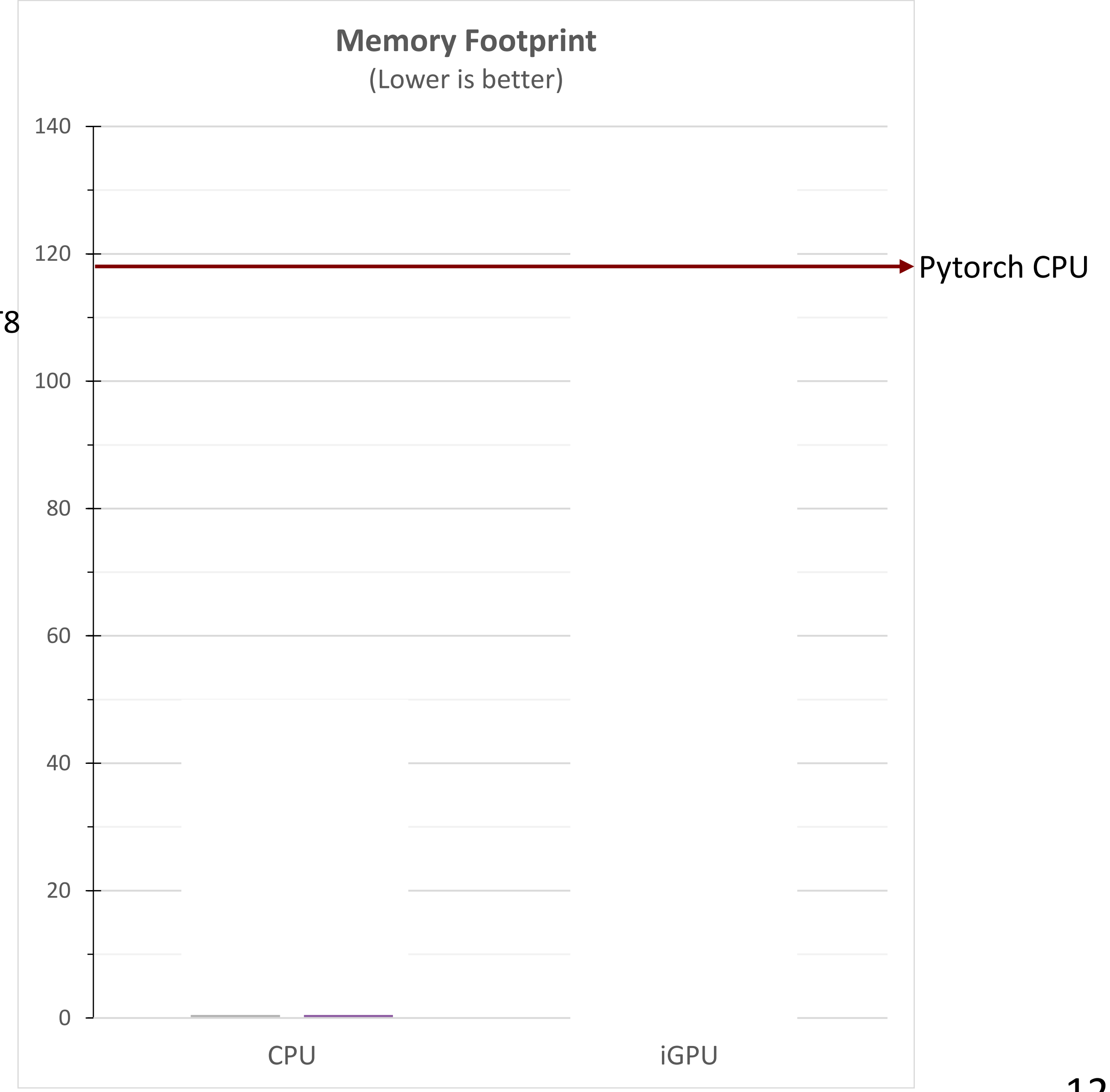
- All optimizations of this study are incorporated in OpenVINO
- OpenVINO v2022.1.0 is integrated in GaNDLF
- Several features are available:
  - Model Optimizer
    - Model conversion
    - Optimization (device agnostic)
  - Post-training Optimization Tool
    - Quantization
  - Extensibility
    - C++
    - OpenCL



Geeks-Sid
thakursp@iu.edu

Results





Geeks-Sid
thakursp@iu.edu

CPU: Intel® Core™ i7-1185G7 @ 3.00GHz w/ Iris® Xe Graphics (8 threads, 16GB)

Results

CONSUMER-GRADE LAPTOP

CPU: Intel® Core™ i7-1185G7 @ 3.00GHz w/ Iris® Xe Graphics
8 threads, 16 GB RAM

		CPU Intel® Core™ i7-1185G7		iGPU Iris® Xe Graphics processor	
	Dice	Execution Time Speedup (X) on CPU	Peak Memory Footprint (MB)	Execution Time Speedup (X) on iGPU (Iris® Xe)	Peak Memory Footprint (MB)
PyTorch FP32	0.79190	1.00	118.04	N/A	N/A
OpenVINO FP32	0.79190	1.48	41.77	3.97	123.92
OpenVINO FP16	0.79180	1.49	43.23	7.03	77.84
OpenVINO PTQ INT8	0.79190	5.16	22.25	11.95	58.38
OpenVINO PTQ INT8	0.79190	5.16	22.25	11.95	58.38

Geeks-Sid
thakursp@iu.edu

Take home message

- Post training optimization improves throughput without affecting DL model quality
- Enabling algorithm execution in low resource environments
- Paving the way for easier/direct clinical translation of AI models

github.com/openvinotoolkit/openvino

Where do we go from here?

- Publishing plan
- Evaluate the method for WSI-based classification
- Explore additional improvements through further optimization methods
  - e.g., distillation, parameter pruning
- Extend to alternate hardware configurations – FPGA, GPU, compute sticks

github.com/mlcommons/GaNDLF

Thank you for your attention!

siddhesh@iu.edu

ACKNOWLEDGEMENTS

IU:
Sarthak Pati
Bhakti Baheti
Ujjwal Baid
Michael Feldman
Spyridon Bakas

Intel:
Ravi Panchumarthy
Junwen Wu
Prashant Shah
Dmitry Kurtaev
Alexander Kozlov
Vasily Shamporov

Kingston University:
Dimitrios Makris

Geeks-Sid
thakursp@iu.edu

FUNDING: NIH/NCI/ITCR: U01CA242871, NIH/NINDS: R01NS042645

Feedback?