Training Deep Neural Networks Directly on Hundred-million-pixel Histopathology Images on a Large-scale GPU cluster

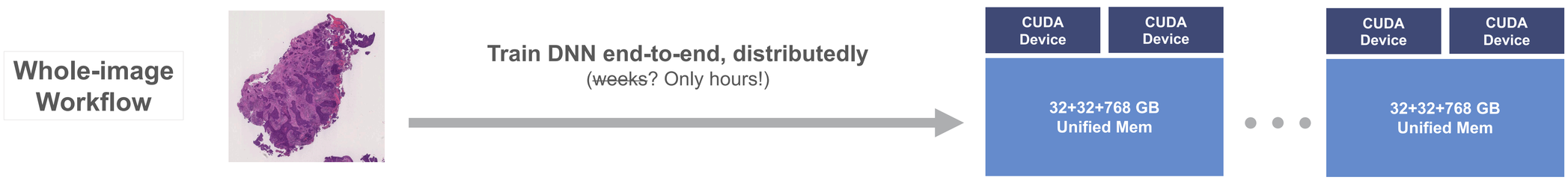
Chi-Chung Chen*†, Wen-Yu Chuang*‡, Wei-Hsiang Yu†, Hsi-Ching Lin§, Shuen-Tai Wang§, Fang-An Kuo§, Chao-Chun Chuang§ and Chao-Yuan Yeh†

*Both authors contributed equally. †aetherAI, Taiwan. ‡Department of Pathology, Chang Gung Memorial Hospital and Chang Gung University, Taoyuan, Taiwan. §National Center for High-Performance Computing, Taiwan.

Whole-image Training Pipeline

Deep learning for digital pathology has been a challenging task because pixel resolution of its major image format, digital whole slide image (WSI), is extremely high, often in the range of billions. The most common approach, patch-based method, crops images to prevent out-of-memory error.

Our Whole-image training pipeline leverages CUDA Unified Memory to fulfill end-to-end training. This scheme skips laborious detailed annotation and preserves inter-patch features. However, using Unified Memory w/o tuning is inefficient due to the limited bandwidth of PCIe interconnect.

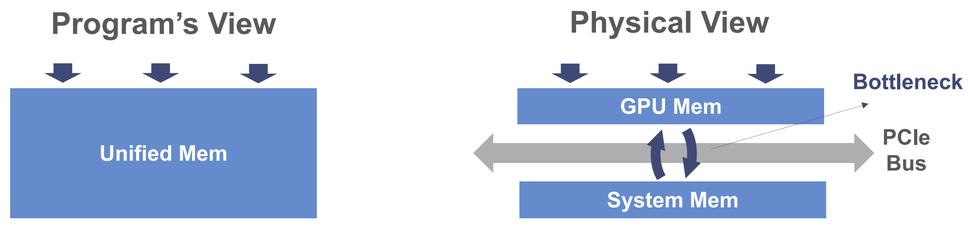


Contributions

- **Our proposed memory optimization methods along with mixed precision training speed up by 411%.**
- **We deploy whole-image pipeline on TAIWANIA 2, a multi-GPU, multi-node supercomputing environment, and achieve 146.28 speedup on 32 GPUs, shortening the total training time from several months (estimated) to 33.1 hours.**
- **CNN trained with our new approach achieves similar level of performance on slide-level prediction with patch-based method and grad-CAM visualization revealed high level of consistency between two methods.**

Memory Optimization

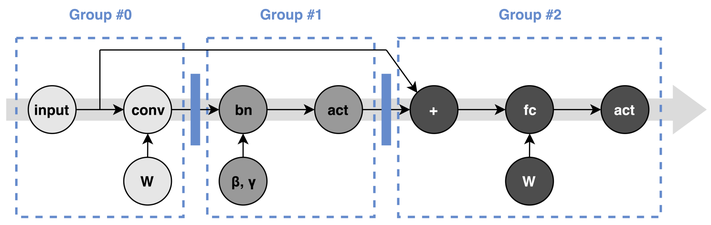
The performance of training DNN with Unified Memory is bounded by the massive access amount on system memory via PCIe.



Group Execution. Deep learning framework, e.g. Tensorflow, tends to execute multiple operations in parallel that cause thrashing in our use case. Curbing some parallelism by placing barriers between operation groups can reduce memory access.

Group Prefetch. During the computation of one group, the data required by the next group can be prefetched in parallel.

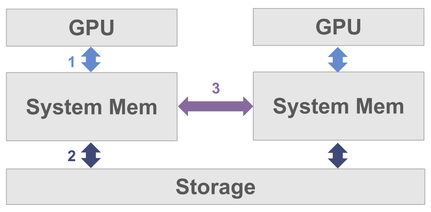
Mixed Precision Training*. Mixed precision training stores and computes most data in half-precision floating point numbers (FP16). FP16 requires half amount of memory space and thus lessen data swapping.



* Micikevicius et al. Mixed Precision Training. ICLR 2018.

Scalable Training

To further shorten the training period, we deploy the training pipeline on distributed computing platform by Horovod.



3 datapaths:

1. Memory swapping for compute (bottleneck)
2. Storage access for images
3. Inter-worker comm. for weight sync.

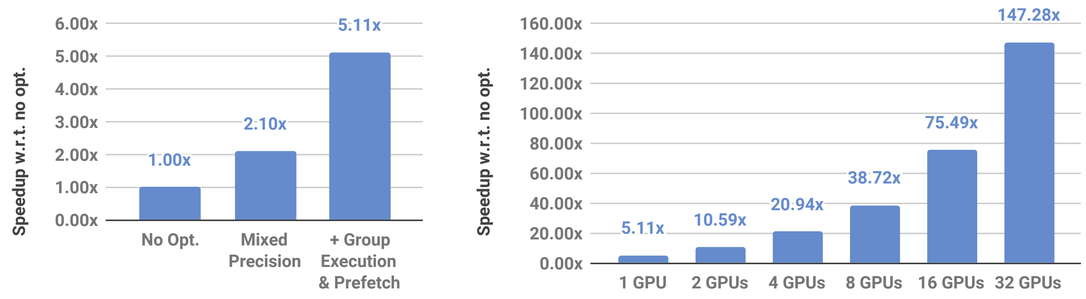
NUMA-aware Binding. To keep efficiency of memory swapping (1), we bind memory space in the same NUMA node with each GPU.

Overlapped Storage Access (2). Training images are prepared in the background threads, hiding the access latency.

Inter-worker communication (3) **cost can be ignored** because model parameters (weights) does not scale with enlarged resolution.

Efficiency Evaluation

We deploy our training pipeline on **TAIWANIA 2**, a GPU supercomputing cluster, achieving **146.28x** speedup on totally 32 GPUs (8 nodes).



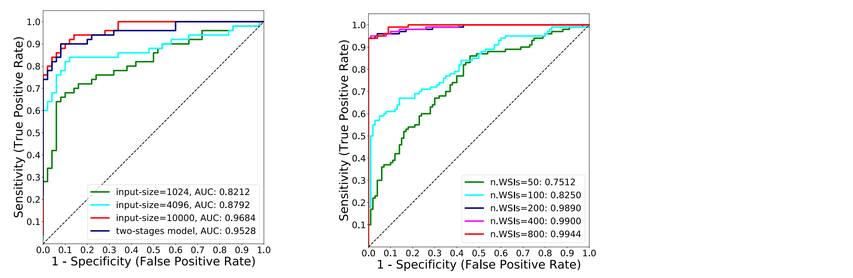
Classification Performance

Here we present the slide level results on both NPC and CRCLN datasets.

ROC Curves.

(left) Performing on the NPC dataset. We compares whole-image method with different downsampled image resolutions and the patch-based method. Whole image approach with 10,000 x 10,000 (limited by cuDNN) input can achieve the same level of performance of patch-based method without any annotation.

(right) Performing on CRCLN dataset with increasing number of slides trained. The whole-image approach achieved an outstanding performance (0.994 AUC).



Visualization. The grad-CAM of whole-image and prediction map of patch-based model shows highly alignment of lesion contours.