Sentinel: Efficient Tensor Migration and Allocation on Heterogeneous Memory Systems for Deep Learning

Jie Ren¹, Jiaolin Luo¹, Kai Wu¹, Minjia Zhang², Hyeran Jeon¹, Dong Li¹
¹University of California Merced ²Microsoft

Motivation

Memory capacity becomes a major bottleneck for DL training
- Heterogenous Memory (HM) is Promising to Address Memory Capacity Limitation
- State-of-the-art Tensor Management Solutions on HM

Characterization of Main Memory Access in DNN

Methodology

Observations
- There are a large number of small tensors (i.e., smaller than 4KB) with short lifetime (i.e., lifetime smaller than one layer) in DNN training workloads.
- The uneven distribution of hot and cold tensors provides opportunities for tensor management.
- Page-level false sharing exists in DNN.
  - The page-level profiling (not tensor-level) for tensor management can be misleading.

Dynamic Profiling
- Profiling overhead is easily amortized over training steps
  - Domain knowledge: exploiting workload repetatability of deep learning training.
- Memory overhead is ignorable.

Lazy (No) Migration for Short-lived Tensors
- Reserve space in fast memory for short-lived tensors
  - Avoid unnecessary data movement
  - Reuse the same memory space to save fast memory space

Proactive Migration for Long-lived Tensors
- The tensor migration is overlapped with DNN training computation as much as possible
  - Domain knowledge: The deep learning model topology (i.e., layers) and its depth to decide the optimal migration interval and trigger data migration.

Experiment Results

Sentinel consistently outperforms five state-of-the-art solutions on both Optane-based and GPU based training platform.

Testing bed
- Optane-based HM
  - Software platform: Linux v4.19, TensorFlow v1.14
  - Hardware platform: Q22 mainboard with four Intel Xeon Gold 6252 CPU at 2.3GHz; DRAM: 32GB x2, Optane: 756GB x2
  - The size of DRAM is equal to 20% of peak memory consumption of DNN models

Overall Performance on Optane-based HM
- Sentinel consistently outperforms five state-of-the-art solutions on Optane-based HM training platform.
- Performance difference between Sentinel and the DRAM-only system is very small

Overall Performance on GPU-based HM
- Sentinel consistently outperforms five state-of-the-art solutions on GPU for models with various training batch size.

References