nicis
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM

Contribution ID: **198**

Type: **not specified**

® Enable Fundamental Cacheability for Distributed Deep Learning Training

Wednesday, 6 December 2023 13:30 (20 minutes)

Deep learning training (DLT) applications exhibit unique I/O workload behaviors that pose new challenges for storage system design. DLT is I/O intensive since data samples need to be fetched continuously from a remote storage. Accelerators such as GPUs have been extensively used to support these applications. As accelerators become more powerful and more data-hungry, the I/O performance lags behind. This creates a crucial performance bottleneck, especially in distributed DLT. At the same time, the exponentially growing dataset sizes make it impossible to store these datasets entirely in memory. While today's DLT frameworks typically use a random sampling policy that treat all samples uniformly equally, recent findings indicate that not all samples are equally important and different data samples contribute differently towards improving the accuracy of a model. This observation creates an opportunity for DLT I/O optimizations by exploiting the data locality enabled by importance sampling.

In this talk, I'll present the design of SHADE, a new DLT-aware caching system that detects fine-grained importance variations at per-sample level and leverages the variance to make informed caching decisions for a distributed DLT job. SHADE adopts a novel, rank-based approach, which captures the relative importance of data samples across different mini-batches. SHADE then dynamically updates the importance scores of all samples during training. With these techniques, SHADE manages to significantly improve the cache hit ratio of the DLT job, and thus, improves the job's training performance.

Student or Postdoc?

Primary author: BUTT, Ali (Virginia Tech)

Presenter: BUTT, Ali (Virginia Tech)

Session Classification: Special