Optane PMem as an Enabler for Large DNN Models with Homomorphic Encryption

Guillermo Lloret-Talavera (Barcelona Supercomputing Center)

Abstract

The proliferation of machine learning services in the last few years has raised privacy concerns. Homomorphic encryption (HE) enables inference using encrypted data but it incurs 100x-10,000x memory and runtime overhead. Secure deep neural network (DNN) inference using HE is currently limited by computing and memory requirements, with frameworks requiring hundreds of gigabytes of DRAM to evaluate small models. To overcome these limitations, we explore the feasibility of leveraging hybrid memory systems comprised of DRAM and persistent memory subsystems. In particular, we explore the recently-released Intel Optane PMem to run large DNNs such as MobileNetV2 (in its largest variant) and ResNet-50 for the first time ever. We present an in-depth analysis of the efficiency of the executions with different hardware and software configurations. Our results conclude that DNN inference using HE incurs on friendly access patterns for this memory configuration, yielding efficient executions.

Bio

Guillermo Lloret-Talavera is a Jr. Research Engineer at the Barcelona Supercomputing Center (BSC), working in the Accelerators and Communications for HPC team. His interests include performance tuning for heterogeneous memory systems and deep learning frameworks.