Introduction

Introduction#

oneDNN Graph API extends oneDNN with a unified graph API for multiple AI hardware classes (CPU, GPU, accelerators). With a flexible graph interface, it maximizes the optimization opportunity for generating efficient code across a variety of Intel and non-Intel HW, and can be closely integrated with ecosystem framework and inference engines. oneDNN Graph API accepts a deep learning computation graph as input and performs graph partitioning, where nodes that are candidates for fusion are grouped together. oneDNN Graph compiles and executes a group of deep learning operations in a graph partition as a fused operation.

With the graph as input, oneDNN Graph implementation can perform target-specific optimization and code generation on a larger scope, which allows it to map the operation to hardware resources and improve execution efficiency and data locality with a global view of the computation graph. With the rapid introduction of hardware support for dense compute, the deep learning workload characteristic changed significantly from a few hot spots on compute-intensive operations to a broad number of operations scattering across the applications. Accelerating a few compute-intensive operations using primitive API has diminishing returns and limits the performance potential. It is critical to have a graph API to better exploit hardware compute capacity.

oneDNN Graph API provides graph partition as a unified graph interface for different types of AI hardware classes. Users construct a graph with operations and logical tensors and pass it to oneDNN Graph implementation to get partitions. oneDNN Graph implementation has a chance to receive a full graph and decides the best way to partition, with the consideration of maximizing performance and coordinating with the application's control of hardware resources. As the partition size can range from single op to the full graph, it satisfies the different needs of graph size for compilation and execution on different AI hardware.

This specification provides high-level descriptions for oneDNN Graph programming model and operation set. More implementation-specific details can be found at open source implementation.