Optimized Code Generation for Deep Learning Networks using LATTE and SWIRL

Janaan Lake
University of Utah

UUCS-20-003

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

22 April 2020

Abstract

As Deep Neural Networks (DNNs) become more widely used in a variety of applications, the need for performance and portability on a number of different architectures, including CPUs, becomes increasingly important. Traditionally, many DNN frameworks resort to statically-tuned libraries to get performance on specific platforms. This approach is limited by the library performance which can vary greatly across different data sizes and layouts, memory hierarchies and hardware features. Compiler-based methods are getting increased attention because they offer opportunities for performance gains by exploiting data reuse and parallelism, efficient memory access, and vectorization for specific backends with the use of abstraction.

Training DNNs can be challenging, and the Batch Normalization (BN) operator has become a popular technique for accelerating training and making networks more robust. Most DNN frameworks include an optimized implementation of this operator, but the computation efficiency decreases dramatically when this operator does not fit the preoptimized version of library functions.

LATTE is a domain-specific language for DNNs, and *SWIRL* is a compiler that can