FTDL: An FPGA-tailored Architecture for Deep Learning Systems

Runbin Shi, *The University of Hong Kong*
Yuhao Ding, *The University of Hong Kong*
Xuechao Wei, *Peking University*
Hang Liu, *Stevens Institute of Technology*
Hayden So, *The University of Hong Kong*
Caiwen Ding, *University of Connecticut*
Contact: rbshi@eee.hku.hk

Hardware acceleration of deep learning (DL) systems has been increasingly studied to achieve desirable performance and energy efficiency. The FPGA strikes a balance between high energy efficiency and fast development cycle and therefore is widely used as a DNN accelerator. However, there exists an architecture-layout mismatch in the current designs, which introduces scalability and flexibility issues, leading to irregular routing and resource imbalance problems. To address these limitations, in this work, we propose FTDL, an FPGA-tailored architecture with a parameterized and hierarchical hardware that is adaptive to different FPGA devices. FTDL has the following novelties: (i) At the architecture level, FTDL consists of Tiled Processing Elements (TPE) and super blocks, to achieve a near-to-theoretical digital signal processing (DSP) operating-frequency of 650 MHz. More importantly, FTDL is configurable and delivers good scalability, i.e., the timing is stabilized even when the design is scaled-up to 100% resource utilization for different deep learning systems. (ii) In workload compilation, FTDL provides a compiler that manages to map the DL workloads to the architecture level in an optimal manner. Experimental results show that for most benchmark layers in MLPerf, FTDL achieves an over 80% hardware efficiency.

Keywords: FPGA primitives; deep-learning architecture; high-performance design

DOI: https://doi.org/10.1145/3373087.3375384