Alibaba Group
阿里巴巴集团

Ultra Low Latency and High Performance Deep Learning Processor with FPGA

Yang Kong, Jun Xu, Xiuyu Sun, Zesheng Dou, Lixin Zhang, Hesen Chen,
Xulin Yu, Hao Li, Yangming Zhang, Xu Hao
Alibaba Inc,

Abstract

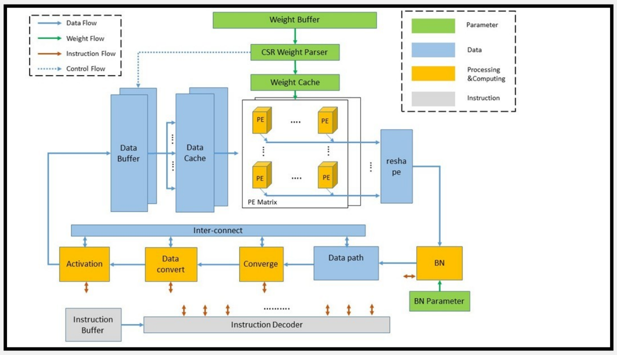
Image recognition and analysis are widely used in Alibaba which has many workload with strict requirement of service quality. However, current solutions such as GPU cannot balance low latency and high performance at the same time.

In order to achieve a good user experience and apply deep learning in some special scenarios, Alibaba infrastructure service group and algorithm team from iDST have architected an ultra low latency and high performance DLP(deep learning processor) on FPGA.

The DLP can support sparse convolution and low precision data computing at some time, meanwhile a customized ISA(instruction Set Architecture) was defined to meet flexibility and user experience. Latency test result with Resnet18(sparse kernel) is 0.174ms,

Architecture

There are mainly 4 types of module function.
- Computing: Convolution, Batch Normalization, Activation and other calculation.
- Data Path: data store, movement and reshape
- Parameter: store weight and other parameters, decoding
- Instruction: Instruction unit and global control

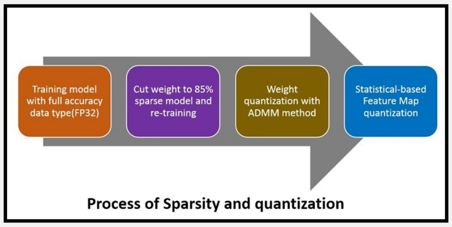


PE can support,
- Int4 data type input.
- Int32 data type output.
- Int16 quantization
- >90% efficiency

Weight loading can support,
- CSR decoder.
- Data pre-fetch

Training

Re-training is needed to achieve excellent model accuracy. There are 4 main steps illustrated below to get both sparse weight and low precision data feature map.


Training model with full accuracy data type(FP32)
Cut weight to 85% sparse model and re-training
Weight quantization with ADMM method
Statistical-based Feature Map quantization

Process of Sparsity and quantization

An effective method is used to train the Resnet18 model to sparse and low precision (1707.09870). We focus on compressing and accelerating deep models with network weights represented by very small numbers of bits, referred to as extremely low bit neural network. We model this problem as a discretely constrained optimization problem. Borrowing the idea from Alternating Direction Method of Multipliers (ADMM), we decouple the continuous parameters from the discrete constraints of network, and cast the original hard problem into several subproblems. We propose to solve these subproblems using extragradient and iterative quantization algorithms that lead to considerably faster convergency compared to conventional optimization methods. Extensive experiments on image recognition and object detection verify that the proposed algorithm is more effective than state-of-the-art approaches when coming to extremely low bit neural network.

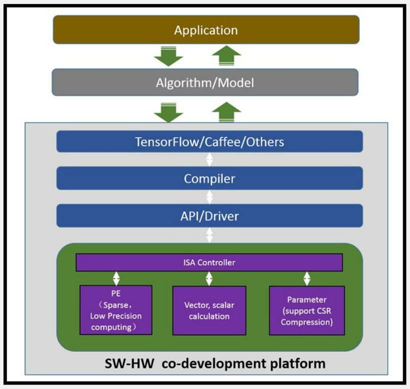
Sparse Rate	0.0	0.5	0.6	0.7	0.8	0.85	0.85-QW	0.85-QW-QFM
Accuracy	0.691	0.691	0.691	0.686	0.681	0.675	0.670	0.663

ISA/Compiler

Only have low latency is not enough for online service and usage scenario since the algorithm model will change frequently. As we know, FPGA development cycle is very long, it usually takes a few weeks or months time to finish a customized design. In order to solve this challenge, we design ISA and compiler to reduce model upgrade time to a few minutes.

Compiler was deployed between NN frameworks and FPGA API/Drivers. It analyze algorithm model graph, then optimize it and translate it to FPGA instructions. There is a Instruction decoder which serve as a ISA controller inside FPGA. It has multi-thread function to enable each NN function modules works in parallel(it also depend on the capability of compiler).

.- Compiler: model graph analysis and instruction generation
- API/Driver: CPU-FPGA DMA picture reshape, weight compression.
- ISA Controller: Instruction decoding, task scheduling, multi-thread pipeline management.


Application
Algorithm/Model
TensorFlow/Caffee/Others
Compiler
API/Driver
ISA Controller
PE （Sparse，Low Precision computing）
Vector, scalar calculation
Parameter (support CSR Compression)
SW-HW co-development platform

Hardware Card

The DLP was implemented on Alibaba designed FPGA Card which has PCIe and DDR4 memory. The DLA can benefit application scenarios such as paying face and Pai-Li-Tao(online searching by picture)in Alibaba.

Resource Type	Utilization
LUTs	50%
DSP	16%
RAM	75%

FPGA （VU13P）Resource Utilization

Function Logic	Clock Rate
DPS Matrix	600 Mhz
Data Plane	300 Mhz

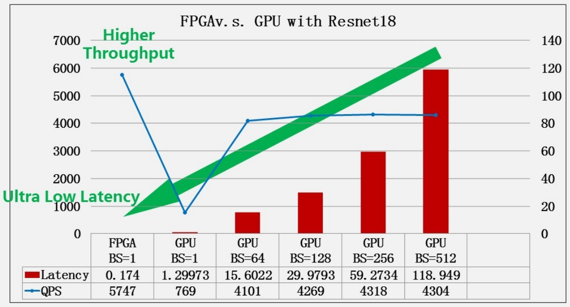
FPGA （VU13P）Clock Rate

Result

FPGA test result with Resnet18 shows that our design achieved ultra-low level latency meanwhile maintaining very high performance with less than 70W chip power.

Hardware	Batch Size	Latency(ms)	QPS
FPGA(DLP)	1	0.174	5800
GPU	1	1.3	769
GPU	64	29.98	4318

Performance Comparison - Resnet18(85% Sparse)



	FPGA BS=1	GPU BS=1	GPU BS=64	GPU BS=128	GPU BS=256	GPU BS=512
🟥Latency	0. 174	1. 29973	15. 6022	29. 9793	59. 2734	118. 949
—●—QPS	5747	769	4101	4269	4318	4304