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

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 modality, Get weight to 80% after re-training.
Weight Quantization with DMM Method.
Statistical-based Sparse 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 weight 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 convergence 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 network.

<table>
<thead>
<tr>
<th>Function Logic</th>
<th>Clock Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH Memory</td>
<td>600 MHz</td>
</tr>
<tr>
<td>Data Plane</td>
<td>300 MHz</td>
</tr>
<tr>
<td>FPGA VCCSP</td>
<td>300 MHz</td>
</tr>
</tbody>
</table>

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.

![Image of FPGA test results with Resnet18](image-url)

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 pictures in Alibaba.

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 an 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).

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Batch Size</th>
<th>Latency(ns)</th>
<th>OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPGA/IPL</td>
<td>1</td>
<td>0.174</td>
<td>5800</td>
</tr>
<tr>
<td>GPU</td>
<td>1</td>
<td>1.3</td>
<td>769</td>
</tr>
<tr>
<td>GPU</td>
<td>64</td>
<td>23.98</td>
<td>4318</td>
</tr>
</tbody>
</table>

Performance Comparison: Resnet50(NVIDIA Apex)

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 pictures in Alibaba.

![Image of FPGA Card](image-url)