Machine Learning on the Latest Intel® Architecture - Three Case Studies

Stephen Blair-Chappell
Abstract

Until recently, if you asked most AI aficionados what their preferred compute platform was for ML/DL, the answer would likely have included the word 'GPU'. In this session we show how three of the latest Intel technologies are being used in the field of Artificial Intelligence to provide compelling alternatives to the GPU.

Firstly, the advantages of the latest Intel Xeon processor - Intel Skylake - with its AVX512 instruction set is explored in the context of AI; Secondly, we show how the Intel FPGA can be used to accelerate the next-generation of Deep Neural Networks; and finally we show how the Intel Movidius Neural Computer Stick can be used to prototype, validate and deploy inference applications offline.
Three ingredients to success

1. Intel Silicon
   - Datacenter
   - Gateway
   - Edge

2. Optimised Frameworks
   - Caffe
   - TensorFlow

3. Intel S/W & tools
   - Intel Parallel Studio XE
   - Intel Nervana

Download resources:
- Intel Nervana: https://www.intelnervana.com/
Intel® Scalable Processors and AI
1: Manufacturing package fault detection

Using the Movidius Compute Stick
2: Hands-on experience

Accelerating CNN with Intel FPGA
3: Using the Intel Computer Vision SDK

Intel® Scalable Processors and AI

1: Manufacturing package fault detection
PROOF OF CONCEPT CASE STUDY


Manufacturing Package Fault Detection Using Deep Learning

By Yiqiang Z. (Intel), JIONG G. (Intel), Updated August 30, 2017
From Sand to package

Sand/Ingots

Photolithography

Wafer Sort/Singulation

Ingot/Wafer
TYPICAL SILICON DEFECTS

- Silicon
- Lithographic
- Mask contamination
- Process Variation
- Defective Oxides
- Electrical defects (shorts/opens)
- Logical Defects
- Cosmetic (fingerprints/scratches /stains)
MANUAL INSPECTION TAKES TIME

• More than 11 criteria rules are used
• Inspection process is subjective
• Human error
• Inconsistent results
“Training should take no longer than 15 hours”
“Accuracy of results needs to be 95%”
THE PLATFORM

Model

Modified GoogleNet

GoogLeNet

Framework

Intel Optimised Caffe

Platform

Intel® Xeon® Scalable Processor

Intel® HPC Developer Conference 2017
• “The required top-1 accuracy is much stricter than that for the standard GoogLeNet V1 on ImageNet-1k (approximately 68.7 percent), but we need 95%+”
THE MODEL - CNN PRODUCES TOO MANY FALSE POSITIVES

The MODEL - CNN produces too many false positives. Standard GoogLeNet V1 is only 69% accurate.
THE MODEL – ADDING RCNN IMPROVES RESULT ACCURACY

Input → Preprocessing → CNN → Postprocessing

Strong passed result → Passed result → Human review

Weak results passed → RCNN

Weak results passed to second stage

Rejected result → Rejected result

Human review
“The topology change is not much. The secret sauce is actually how we do pre-processing and post-processing.”
Two other optimisations

- Input
- Preprocessing
- CNN
- Postprocessing
- Passed result
- Rejected result
- Human review

All pictures are green, so only one channel from RGB is used.

Only 2 classifications are needed.

Passed result

Human review
CODE CHANGE TO GOOGLENET - PASS/FAIL CLASSIFICATION

layer {
  name: "loss3/classifier"
  type: "InnerProduct"
  bottom: "pool5/7x7_s1"
  top: "loss3/classifier"
  param {
    lr_mult: 1
    decay_mult: 1
  }
  param {
    lr_mult: 2
    decay_mult: 0
  }
}

inner_product_param {
  num_output: 2
  weight_filler {
    type: "xavier"
  }
  bias_filler {
    type: "constant"
    value: 0
  }
}
PREPROCESSING TRICK

Pass
Pass
Fail

4,000 images

Rotate 36 times

144,000 images

All pictures are green, so only one channel from RGB is used
Pass rate has less than 5% false negatives, so results can be accepted without review.

Human review

Reject rate has 30% false positives hence review

Human review
**PREPROCESSING**

- Extract region of interest using OpenCV
- Augment the extracted images with rotations

**POSTPROCESSING**

- Ensemble the results from rotated images
- Feed results into RCNN
# RESULTS: 2-SOCKET-BROADWELL VS 2-SOCKET-SKYLAKE

<table>
<thead>
<tr>
<th>Hardware</th>
<th>TFlops</th>
<th>Software</th>
<th>Batch Size</th>
<th>Images per Second</th>
<th>Time to Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel® Xeon® processor E5-2699</td>
<td>3.1</td>
<td>Berkeley Vision and Learning Center Caffe with CPU mode</td>
<td>36</td>
<td>1.5</td>
<td>55 days (estimated)</td>
</tr>
<tr>
<td>Intel Xeon processor E5-2699</td>
<td>3.1</td>
<td>Caffe optimized for Intel® architecture with the Intel® Math Kernel Library (Intel® MKL)</td>
<td>36</td>
<td>173</td>
<td>11.5 hours</td>
</tr>
<tr>
<td>Intel® Xeon® Platinum 8180 processor (2.5 GHz 8180)</td>
<td>8.2</td>
<td>Caffe optimized for Intel architecture with the Intel MKL</td>
<td>36</td>
<td>266</td>
<td>7.5 hours</td>
</tr>
</tbody>
</table>

E5-2699: “Broadwell”, 22 core, 256bit vector registers, 2 FMA

Platinum 8180: “Skylake”, 28 cores, 512bit vector registers, 2 FMA
Reminder – Difference between frameworks

- Improved optimisations
- Vectorisation AVX512
- Threading
- Using MKL ‘under the hood’
“we can filter out more than 70% of the images as “Pass” images, so the remaining images are only 30% of the original volume so 70 percent of the manual inspection effort was saved”

APPLYING AI reduces manual effort by 70%
REFERENCES


Manufacturing Package Fault Detection Using Deep Learning

By Yiqiang Z. (Intel), JIONG G. (Intel), Updated August 30, 2017
Using the Movidius Compute Stick

2: Hands-on experience
MYRIAD 2 VISION PROCESSING UNIT (VPU)

Myriad 2 Vision Processor Unit (VPU)

Myriad 2 MA2x5x Block Diagram
VVPU Example Use - The DJI Spark Drone

- Face Aware
- Gesture Mode
- Safe Landing

See: https://www.dji.com/spark
Example: Scaling inference performance with multiple sticks
Movidius Neural Compute Stick
Redefining the AI Developer Kit

- Neural Network Accelerator in USB Stick Form Factor
- No additional heat-sink, no fan, no cables, no additional power supply
- Prototype, tune, validate and deploy deep neural networks at the edge
- Features the same Movidius vision processing unit (VPU) used in drones, surveillance cameras, VR headsets, and other low-power intelligent and autonomous products
MOVIDIUS NEURAL COMPUTE STICK
REDEFINING THE AI DEVELOPER KIT

NC SDK
Free download @ developer.movidius.com
NC SDK WORKFLOW

TRAINING

Model

Trained model

PROFILING, TUNING, AND COMPILING

Neural Compute Stick

Development Host

Movidius™ NC Toolkit

Compiled model

PROTOTYPING

Neural Compute Stick

Prototyping Host

Movidius™ NC API
WHAT CAN I DO WITH THE NCS?

Profiler

A tool that provides a detailed stage-by-stage breakdown of where the bottlenecks are in your system.

Checker

 Runs a single inference on the NCS using the provided model, allowing for the calculation of classification correctness.

Compiler

The compiler is used to create a graph which is an optimized binary file that can be processed by the NCS.

C API

GetDeviceName
OpenDevice
AllocateGraph
DeallocateGraph
LoadTensor
SetGraphOption
CloseDevice
...

Python bindings

Status
GlobalOption
DeviceOption
GraphOption
EnumerateDevices
SetGlobalOption
LoadTensor
...

DNN architect / data scientist

Applications developer
Accelerating CNN with Intel FPGA

3: Using the Intel Computer Vision SDK
INTEL COMPUTER VISION SDK

INTEL® COMPUTER VISION SDK BETA

Deep Learning Deployment Toolkit
Cross hardware approach to deep learning inference
Model Optimizer
Convert & optimize trained models
Inference Engine
Run optimized inferences

OpenCV*
Optimized functions for Intel® Processors

OpenVX™
Enhanced, graphical development using Vision Algorithm Designer

OpenVX and the OpenVX logo are trademarks of the Khronos Group Inc.
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos
*Other names and brands may be claimed as the property of others.
Supported Topologies

- AlexNet
- GoogleNet v1
- VGG-16
- SqueezeNet 1.0 and 1.1
- ResNet-18
- SqueezeNet-based variant of the SSD*
- GoogLeNet- based variant of SSD
Computer Vision Pipeline Application

Input Image → Vision nodes / functions → CNN Nodes → Vision nodes / functions → Inference Results

Pre-trained Optimized Deep Learning Models

Intermediate Representation (IR) xml binary file that contains weights & biases

Model Optimizer → Inference Engine

Intel® Computer Vision SDK Beta – Deep Learning Deployment Toolkit

Runs on FPGA

Runs on Host
For developers looking to run deep learning models on the edge

1. Imports trained models from popular DL framework regardless of training HW
2. Enhances model for improved execution, storage & transmission
   - Optimizes Inference execution for target hardware (computational graph analysis, scheduling, model compression, quantization)
3. Enables seamless integration with application logic
4. Delivers embedded friendly Inference solution

Ease of use + Embedded friendly + Extra performance boost
Closing Comment
Three ingredients to success

1. **Intel Silicon**
   - Datacenter
   - gateway
   - Edge

2. **Optimised Frameworks**
   - Caffe
   - TensorFlow

3. **Intel S/W & tools**
   - Intel® Parallel Studio XE
   - Intel Nervana
     - https://www.intelnervana.com/