BRING INTELLIGENCE TO THE EDGE WITH INTEL® MOVIDIUS™ NEURAL COMPUTE STICK

Darren Crews
Principal Engineer, Lead System Architect, Intel NTG
LEGAL DISCLAIMER

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENCE IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product or order.

Code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.

*Other names and brands may be claimed as the property of others.

Copyright © 2017, Intel Corporation. All rights reserved.
Agenda

• Motivation to move intelligence to the edge
• Edge compute use cases
• Barriers to moving intelligence to the edge
  • Deep learning algorithms – can they run on an edge device?
• Movidius Neural Compute Stick (arch, usage, etc)
• Code walkthrough and demo
Let's look at a larger scale...

20 billion connected devices by 2020¹

generating billions of petabytes of data traffic between devices & the cloud

¹ Source: [http://www.gartner.com/newsroom/id/3598917](http://www.gartner.com/newsroom/id/3598917)
Why move intelligence from the cloud to the edge?

What are you? I am asking the cloud if I should vacuum you, too.

I'll scratch you down to your motors if you come any closer!

Reduce BW  
Reduce Latency  
Privacy
Movidius
an Intel company

Computer vision and AI at the edge
But what about Developers?

Complete HW + SW solution for developing Deep Learning application on the edge

NC SDK
Free download @ developer.movidius.com
Use Case with the Neural Compute Stick

Little Ripper Lifesaver* UAV

https://www.youtube.com/watch?v=QJOMfDyhUyo
https://newsroom.intel.com/editorials/data-most-important-force-society-today/
Key Capabilities

Typical use cases could be:
• Robot
• Security camera
• Smart-home assistant

Key capabilities:
• Object detection
• Object classification
• Facial recognition
• Natural language processing
Barriers to Moving Compute to the Edge

Move compute from the cloud to the edge:

• Accuracy
• Available compute
• Model efficiency
• Model size
Image Classification: Getting more accurate and every year
Efficiency Key for Edge Devices

Alfredo Canziani, Eugenio Culurciello, Adam Paszke “AN ANALYSIS OF DEEP NEURAL NETWORK MODELS FOR PRACTICAL APPLICATIONS”
MobileNet performance on NCS

Real time image classification on NCS

https://github.com/tensorflow/models/blob/master/research/slim/nets/mobilenet_v1.md
Object Detection Benchmarks


Ross Girshick, IEEE International Conference on Computer Vision (ICCV), 2015
**Object Detectors – Speed matters!**

Significant improvements in Object Detectors in the last two years enable Object Detection on edge devices like Intel Movidius NCS.

<table>
<thead>
<tr>
<th></th>
<th>Pascal 2007 MaP</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN</td>
<td>66.0</td>
<td>0.05FPS</td>
</tr>
<tr>
<td>Fast R-CNN</td>
<td>70.0</td>
<td>0.5FPS</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>73.2</td>
<td>7FPS</td>
</tr>
<tr>
<td>YOLO</td>
<td>69.0</td>
<td>45FPS</td>
</tr>
<tr>
<td>SSD-Mobilenet</td>
<td>72.7</td>
<td>11FPS</td>
</tr>
</tbody>
</table>

Jonathan Huang, et al, Speed/accuracy trade-offs for modern convolutional object detectors
Object Detectors – lots of choices, choose wisely

Jonathan Huang, et al, Speed/accuracy trade-offs for modern convolutional object detectors
Machine Translation

Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]
Intel® Movidius™ Neural Compute Stick
Redefining the AI developer kit

• Neural network accelerator in USB stick form factor

• TensorFlow™ and Caffe frameworks supported, along with many popular networks

• Source is available for the SDK, which allows you to compile for other platforms

• Features the same Intel Movidius vision processing unit (Intel Movidius VPU) used in drones, surveillance cameras, VR headsets, and other low-power intelligent and autonomous products
Demo: Scaling inference performance with multiple sticks
Object Detection and Classification

Downscale → Crop → Draw box and class

Object detection
Bird classification

https://github.com/movidius/ncappzoo/tree/master/apps/birds
Architecture

Intel Movidius NCS contains the Intel® Movidius™ Myriad™ 2 vision processing unit, including 4 Gbit of LPDDR.

Intel Movidius NCS is connected to an application processor (AP), such as a Raspberry Pi or UP Squared board.

Execution is controlled by the LEON microprocessor, and the calculations are done on the SHAVE processors.
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

Intel® Movidius™ Neural Compute Stick Program
NC SDK Workflow

TRAINING

Model + Trained model

PROFILING, TUNING, AND COMPILING

Neural Compute Stick + Development Host

Compiled model

Movidius™ NC Toolkit

PROTOTYPING

Neural Compute Stick + Prototyping Host

Movidius™ NC API
Development Flow

Step 1: Convert the model
Trained model can be profiled on NCS for performance:

Summary

- 12 SHV Processors
- Inference time 91.25 ms
- Bandwidth 1821.01 MB/sec

This network is Compute bound
NC SDK Workflow
Development Flow

Step 2: Load the model and the FW
Development Flow

Step 3: Perform inference
- Load the image
- Run the model
- Return the results
Demo
Let's Review

• Moving AI to the edge is important for a number of reasons (lots of data, latency, and privacy)

• Deep Learning has progressed in many areas where these use cases can be run on the edge (object detection, classification, etc.)

• Intel Movidius NCS is an easy-to-use prototyping vehicle for developing your edge devices
Explore developer.movidius.com

Try out the following pages:

- Main page
- Getting started
- Downloads
- Docs
- Forums
- Where to buy
Questions?