Introducing NetsPresso by Nota: For Really Fast Inference on Cortex-A devices

Nota Inc.

Tae-Ho Kim, CTO&Co-Founder
22nd June, 2021
Short Bio

Tae-Ho Kim, CTO, Nota Inc.

- B.S. Bio and Brain Engineering, KAIST
- M.S. Electrical Engineering, KAIST
- Research Intern, Universite de Montreal (P.I. Yoshua Bengio)
- Research Intern, Chinese University of Hong Kong (P.I. Xiaogang Wang)
- Senior Research Scientist, KAIST Institute
- CTO / Co-Founder, Nota

Research interests: Deep learning architecture, its application to CV, NLP, and Speech
Contents

• Deep Learning? How big?
• Compression Methods
• Nota’s Solution
• Q&A
Achievements of Deep Learning

Google DeepMind's AlphaFold 2

AI Breakthrough in Biology

Future is coming

https://openai.com/blog/dall-e/
Deep Learning? How big?

GPT-2: 1.5B Parameters

GPT-3: 175B Parameters

It’s going to be bigger...

Next?
Deep Learning? How big?

Training the Model

GPT-3 is trained using next word prediction, just the same as its GPT-2 predecessor. To train models of different sizes, the batch size is increased according to number of parameters, while the learning rate is decreased accordingly. For example, GPT-3 125M use batch size 0.5M and learning rate of $6.0 \times 10^{-4}$, where GPT-3 175B uses batch size 3.2M and learning rate of $0.6 \times 10^{-4}$.

We are waiting for OpenAI to reveal more details about the training infrastructure and model implementation. But to put things into perspective, GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run. Similarly, a single RTX 8000, assuming 15 TFLOPS, would take 665 years to run.

Time is not the only enemy. The 175 Billion parameters needs $175 \times 4 \rightarrow 700\text{GB}$ memory to store in FP32 (each parameter needs 4 Bytes). This is one order of magnitude larger than the maximum memory in a single GPU (48 GB of Quadro RTX 8000). To train the larger models without running out of memory, the OpenAI team uses a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on V100 GPUs on the part of a high-bandwidth cluster provided by Microsoft.

In fact, the size of SOTA language model increases by at least a factor of 10 every year: BERT-Large (2018) has 355M parameters, GPT-2 (early 2019) reaches 1.5B, T5 (late 2019) further stretches to 11B, GPT-3 (mid-2020) finally gets to 175B. The progress of the sizes of language models clearly outpace the growth of GPU memory. This implies that for NLP, the days of “embarrassingly parallel” is coming to an end, and model parallelization is going to be indispensable for researching SOTA language models.

355 GPU-years
$4.6M$
700GB

Super-huge!
Challenges of Cloud-based AI

- High Cost
- Privacy Concerns
- Network Connectivity Issues

Cloud-based AI becomes Unsustainable
AI Model Compression

Solution: Nota’s AI Model Compression Technology
How can we use deep learning at the edge?

- NAS (Neural Architecture Search)
- Popular models
- Filter Decomposition
- Knowledge Distillation
- Network Pruning
- Quantization
- Joint Approach
- Kernel Optimization
- Graph Optimization
- HW improvement
Model Training

- Neural Architecture Search (NAS)
  - Gradient based NAS
  - Reinforcement learning based NAS

* Benchmark Analysis of Representative Deep Neural Network Architectures
Model Training

- Neural Architecture Search (NAS)
  - Gradient based NAS
  - Reinforcement learning based NAS
- Popular models
  - Residual Networks
  - EfficientNetB0
  - MobileNetv2

* Benchmark Analysis of Representative Deep Neural Network Architectures
Deployment

Various lineup for edge AI is in progress

- Cortex-A, Mali and Ethos-N78
- Cortex-M and Ethos-U55/U65
- Cortex-M55
- Cortex-M today

Data throughput

http://linuxgizmos.com/arm-unveils-two-lightweight-npus-for-edge-ai/

- Vibration detection
- Sensor fusion
- Keyword detection
- Anomaly detection
- Object detection
- Gesture detection
- Biometric awareness
- Speech recognition
- Object classification
- Real-time recognition
Ethos-N78 NPU: Performance, Efficiency and Configurability

2nd Generation NPU

**Standalone**
Fully autonomous ML graph processor

1-10
TOP/s (scalable)

**Int8**
Optimised for int8 weights and activations

>100
Unique configurations
How can we use deep learning at the edge?

Model Training
- NAS (Neural Architecture Search)
- Popular models

Model Compression
- Filter Decomposition
- Knowledge Distillation
- Network Pruning
- Quantization
- Joint Approach

Compile
- Kernel Optimization
- Graph Optimization

Deployment
- HW improvement
Filter Decomposition

Cost saving

\[
\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2}
\]

Types of filter decomposition

- Tucker Decomposition (Z. Zhong, 2019)
- Depthwise Separable Convolution (A. Howard, 2017)
- Network Decoupling (J. Guo, 2018)
- Truncated SVD
- ...
Knowledge Distillation

Types of Knowledge Distillation
- Features Distillation
- Softlabel
- Attention Distillation

Techniques
- KD (G. Hinton, 2015)
- FitNets (A Romero, 2014)
- OverHaul KD (B Heo, 2019)
- Relational KD (W Park, 2019)
- ...
Toward Compact Deep Neural Networks via Energy-Aware Pruning

Pruning weight with nuclear norm threshold.
Quantization

- Bit? 0 or 1
- 2 bit variable can represent 4 numbers
- 32 bit variable can represent $2^{32}$ numbers

- DoReFa (S Zhou, 2016)
- PACT (J Choi, 2018)
- QAT (Quantization Aware Training)
- PTQ (Post Training Quantization)
- ...

Slide credit: Song Han, 2015
Model compression for AI everywhere
About Us

Nota
To make the world convenient

Nota Inc. provides edge-device AI solutions that remove the need for any server or cloud computing.

Our mission is to democratize AI-based solutions to every aspect of society.

- Founded (KAIST E5)
- Seed Funding: Naver (1st Portfolio)
- Pre-Series A ($1.2M): Stonebridge Ventures
- New branch in San Jose, US
- Successfully ended, Intel PoC Prj.
- Item pivoting to on-device AI
- Bridge Funding: Naver & Blueprint partners
- Partnership with Intel, Nvidia, SK Telecom, Samsung SDS, LG CNS, Axis
- 40+ clients/partners globally
- Total 24 publications
- New subsidiary in Berlin, Germany
- Series A ($7Mil.): SBV, LBI, Samsung Venture Investment, LG CNS
- 40+ clients/partners globally
- Total 24 publications
- New subsidiary in Berlin, Germany
- Series A ($7Mil.): SBV, LBI, Samsung Venture Investment, LG CNS
Team Nota

Organization

<table>
<thead>
<tr>
<th>Department</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>R&amp;D</td>
<td>41 Members</td>
</tr>
<tr>
<td>BD</td>
<td>9 Members</td>
</tr>
<tr>
<td>Operation</td>
<td>6 Members</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>56+ Members</strong></td>
</tr>
</tbody>
</table>
## Business

### Clients

- 20+
- Manufacturer
- Radar company
- Medical device company
- Chipset Vendor

### References

- 30+
- Radar company
- Medical device company
- Chipset Vendor

### Publications

- 24+
- 21 patents

<table>
<thead>
<tr>
<th>Revenue Jul. 2020</th>
<th>Potential Revenue 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>€600K</td>
<td>€1.3M</td>
</tr>
</tbody>
</table>

---

© 2020 Arm Limited (or its affiliates)
Business Portfolio

AI Model Development

- DMS / IMS / Hand Gesture Recognition
- Face Recognition / PPE
- Inventory Management / Customer Journey Analysis
- Video Analytics

AI Model Compression

- Classification / Detection / Tracking / Segmentation
# Nota’s Compression Technology

<table>
<thead>
<tr>
<th>Same Accuracy</th>
<th>Efficient</th>
<th>Faster</th>
<th>Low Power</th>
<th>Cost-effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td># of Computation</td>
<td>Inference Time</td>
<td>Power Consumption</td>
<td>Cost for AI systems</td>
</tr>
<tr>
<td>Original 97%</td>
<td>Original 14.4B</td>
<td>Original 0.38sec</td>
<td>Original 100%</td>
<td>Original 100%</td>
</tr>
</tbody>
</table>

**SAME**
- 97%
- 2.3B
- 0.08sec
- 60%
- 15%

*Computed by a current client (Manufacturer).*

Face recognition in Raspberry Pi 3+.
Competitive Advantage

- Conventional AI model compression

Pretrained Model → Compression Technique → Compressed Model

Compression Methods:
- Pruning
- Quantization
- Knowledge Distillation
- NAS

Problems of current network compression

- DL engineers manually compress the model
- Compression methods are developed in different places and forms
- Hard to know which compression method or combination to use
- Compression metric does not fit to practical metric

https://www.youtube.com/watch?v=ZIPjHmAUyJ0
Competitive Advantage

- Nota’s NetsPresso (Automatic Model Compression Platform)

- Problem Solving
  - Automatic compression without manpower
  - Combination of multiple compression methods
  - Fitted metric for practical usage

Nota’s Automatic AI Model Compression Platform: NetsPresso

- Optimum compression platform for:
  - Target task
  - Target dataset
  - Target device
  - Target accuracy / latency / model size
Competitive Advantage

Competitors

Demand → Big Model / Target HW → Supply

Companies

Compressed Model

Human customizing > 2 weeks

NetsPresso

Demand → Big Model / Target HW → Supply

Companies

Compressed Model

Automatically compressed ≤ a week

© 2020 Arm Limited (or its affiliates)
Structure

**INPUT**

- TensorFlow
- PyTorch
- Keras
- ONNX
- MXNet

- NVIDIA Jetson
- Coral board
- RPi
- Mobile Server

- Data Label

- Toolkit (OpenVINO/ TF Lite/ TensorRT)
  - Model size
  - Latency

**OUTPUT**

- Pre-trained Model

- Target H/W

- Dataset

- Target Performance

- Compression methods
  - L1 Ch P
  - L1 W P
  - G-Pruning
  - S-Conv
  - BLKD
  - Post Q
  - NAS
  - QAT
  - Etc.
  - O-KD
  - LQ
  - HAQ

- Optimized model
  - H/W
  - Toolkit
  - Size
  - Latency

- Hyper-Parameter Optimization

- Compressed Model

- Edge device Pool
### Benchmark Table

<table>
<thead>
<tr>
<th></th>
<th>EfficientNetB0</th>
<th>ShuffleNetV2</th>
<th>MBv3 + SSDlite</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>84.79% +2.71%</td>
<td>85.51% +0.26%</td>
<td>mAP 15% +0% 15%</td>
</tr>
<tr>
<td><strong>Number of Parameters</strong></td>
<td>4.06M x81% 3.31M</td>
<td>4.03M x52% 2.09M</td>
<td>FLOPs 330M x65% 215M</td>
</tr>
<tr>
<td><strong>Inference Time</strong></td>
<td>54ms x84% 45ms</td>
<td>112ms x86% 97ms</td>
<td>FPS 3.3 x151% 5</td>
</tr>
<tr>
<td>Dataset</td>
<td>Network</td>
<td>Acc (%)</td>
<td>FLOPs (M)</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>CIFAR 100</td>
<td>ResNet18</td>
<td>76.92</td>
<td>1110.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>76.02 (-0.9)</td>
<td>254.49 (4.37x)</td>
</tr>
<tr>
<td>CIFAR 100</td>
<td>VGG19</td>
<td>72.28</td>
<td>796.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.77 (-1.51)</td>
<td>361.74 (2.2x)</td>
</tr>
<tr>
<td>CIFAR 100</td>
<td>ResNet50</td>
<td>78.03</td>
<td>2596.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>76.84 (-1.19)</td>
<td>359.43 (7.22x)</td>
</tr>
<tr>
<td>CIFAR 100</td>
<td>MobileNet V1</td>
<td>66.68</td>
<td>92.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>65.73 (-0.95)</td>
<td>18.95 (4.9x)</td>
</tr>
</tbody>
</table>
Use Case 1. Facial Recognition

- Built facial recognition algorithms with 99.9% accuracy by losslessly optimizing AI model that requires no HW or design upgrades.

**SPEED**
- < 0.3 sec in total
  - Detection: 0.05 sec
  - Extraction: 0.2 sec
  - Matching: 0.01 sec

**ACCURACY**
- 99.9% Accuracy
  - TPIR: 99.9%
  - FAR: 0.01%

**SECURITY**
- Advanced Security
  - Anti-Spoofing (liveness detection)
  - No privacy infringement

**LIGHTWEIGHT**
- Compressed Model
  - Run independently on small edge devices without GPU

© 2020 Arm Limited (or its affiliates)
Use Case 2. Intelligent Transportation System

- The first company in the world to deliver an on-device AI solution to monitor vehicles and control traffic signals in real time to increase vehicle speed by 3x during rush hours.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Size</td>
<td>150MB</td>
<td>1.5MB</td>
</tr>
<tr>
<td>Frame Rate</td>
<td>7fps</td>
<td>30fps</td>
</tr>
<tr>
<td>Vehicle Speed (Rush Hour)</td>
<td>~10km/h</td>
<td>~30km/h</td>
</tr>
<tr>
<td>On-Device AI</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
NetsPresso Inside

- NetsPresso Demo Video

netspresso
Wanna Try NetsPresso?

Contact us for trial!
### AI Virtual Tech Talks Series

<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 11th</td>
<td>Bringing PyTorch Models to Arm Cortex-M Processors</td>
<td>AITS</td>
</tr>
<tr>
<td>May 25th</td>
<td>Optimized C Code Generation for Ultra-efficient tinyML Applications</td>
<td>Imagimob</td>
</tr>
<tr>
<td>June 8th</td>
<td>Leveraging DarwinAI’s Deep Learning Solution to Improve Production Efficiency in Manufacturing</td>
<td>DarwinAI</td>
</tr>
<tr>
<td>June 22nd</td>
<td>Introducing NetsPresso by Nota: For Really Fast Inference on Cortex-A devices</td>
<td>Nota.AI</td>
</tr>
<tr>
<td>July 13th</td>
<td>Bringing Edge AI to Life - from PoC to Production</td>
<td>Arcturus &amp; NXP</td>
</tr>
<tr>
<td>July 20th</td>
<td>Easy TinyML with Arduino: taking advantage of machine learning right where things are happening</td>
<td>Arduino</td>
</tr>
</tbody>
</table>

Visit: [developer.arm.com/techtalks](http://developer.arm.com/techtalks)
Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board
Thank You
Danke
Merci
谢谢
ありがとうございます
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
תודה
Model Training

• Neural Architecture Search (NAS)
  • Gradient based NAS
  • Reinforcement learning based NAS

* Benchmark Analysis of Representative Deep Neural Network Architectures
Model Training

- Neural Architecture Search (NAS)
  - Gradient based NAS
  - Reinforcement learning based NAS
- Popular models
  - Residual Networks
  - EfficientNetB0
  - MobileNetv2

* Benchmark Analysis of Representative Deep Neural Network Architectures
Compile (Runtime Optimization)

How to ‘fit’ SW to HW is an important problem
Network Pruning

Weight? Filter? Channel?
- Structured Pruning
- Unstructured Pruning

Metrics
- L1 / L2 (S. Han, 2015)
- GM Pruning (Y He, 2018)
- BN Pruning (Y Liu, 2019)

Comparison scope
- Local Pruning
- Global Pruning

Slide credit: Song Han, 2015
Automatic Network Adaptation for Ultra-Low Uniform-Precision Quantization

Algorithm 1: Neural Channel Expansion

Input:
- Split the training set into two disjoint sets: \( D_{\text{weight}} \) and \( D_{\text{arch}} \) \( n(D_{\text{weight}}) = n(D_{\text{arch}}) \)
- Search Parameter: \( \{ \alpha_1^l, \alpha_2^l, \ldots, \alpha_i^l \} \in A_i \)
- \( \{ A_1, A_2, \ldots, A_i \} \subset A_i \), \( L \) = number of layer
- Expand Threshold: \( T \)

For Warm-up Epoch do
1. Sample batch data \( D_{\text{w}} \) from \( D_{\text{weight}} \) and network from \( \Lambda \sim U(0, 1) \)
2. Calculate \( \mathcal{L}_{\text{weight}} \) on \( D_{\text{w}} \) to update network weights
3. End for

For Search Epoch do
4. Sample batch data \( D_{\text{w}} \) from \( D_{\text{weight}} \) and network from \( \mathcal{S}_{\text{softmax}}(\Lambda) \)
5. Calculate \( \mathcal{L}_{\text{weight}} \) on \( D_{\text{w}} \) to update network weights
6. Sample batch data \( D_{\text{a}} \) from \( D_{\text{arch}} \) and network from \( \mathcal{S}_{\text{softmax}}(\Lambda) \)
7. Calculate \( \mathcal{L}_{\text{arch}} \) on \( D_{\text{a}} \) to update \( \Lambda \)
8. For layer do
9. \( j \leftarrow \#A^l \)
10. If \( \mathcal{S}_{\text{softmax}}(\alpha^l_j) \geq T \) do
11. Expand search space \( (\alpha^l_{j+1}) \)
12. End if
13. End for
14. Derive the searched network from \( \Lambda \)
15. Randomly initialize the searched network and optimize it on the training set

<table>
<thead>
<tr>
<th>Network</th>
<th>Method</th>
<th>Top-1 Acc</th>
<th>Top-5 Acc</th>
<th>FLOPs</th>
<th>PARAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>LSQ</td>
<td>67.6%</td>
<td>87.6%</td>
<td>1.814G</td>
<td>11.69M</td>
</tr>
<tr>
<td></td>
<td>QIL</td>
<td>65.7%</td>
<td>85.9%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LQ-Nets</td>
<td>64.4%</td>
<td>85.6%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PACT</td>
<td>65.9%</td>
<td>86.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>EdIMIPS</td>
<td></td>
<td></td>
<td>4.089G</td>
<td>25.56M</td>
</tr>
<tr>
<td>ResNet50</td>
<td>LSQ</td>
<td>73.7%</td>
<td>91.5%</td>
<td>4.089G</td>
<td>25.56M</td>
</tr>
<tr>
<td></td>
<td>LQ-Nets</td>
<td>71.5%</td>
<td>90.3%</td>
<td>4.089G</td>
<td>25.56M</td>
</tr>
<tr>
<td></td>
<td>PACT</td>
<td>72.2%</td>
<td>90.5%</td>
<td>4.089G</td>
<td>25.56M</td>
</tr>
<tr>
<td></td>
<td>EdIMIPS</td>
<td>72.1%</td>
<td>90.6%</td>
<td>4.089G</td>
<td>25.56M</td>
</tr>
</tbody>
</table>

Uniform precision quantization with channel expansion
Joint Approach

- Once-for-all: Considering pruning, KD, kernel size, and number of layers (ICLR 2020)
- APQ: Joint Search for Network Architecture, Pruning and Quantization Policy (CVPR 2020)

Slide credit: Song Han
Deployment

Various lineup for edge AI is in progress

Deployment

Various lineup for edge AI is in progress

Cortex-A, Mali and Ethos-N78

Cortex-M and Ethos-U55/U65

Cortex-M55

Cortex-M today

Vibration detection
Sensor fusion
Keyword detection
Anomaly detection
Object detection
Gesture detection
Biometric awareness
Speech recognition
Object classification
Real-time recognition

Data throughput

http://linuxgizmos.com/arm-unveils-two-lightweight-npus-for-edge-ai/
Ethos-N78 NPU: Performance, Efficiency and Configurability

2nd Generation NPU

Standalone
Fully autonomous ML graph processor

1-10 TOP/s (scalable)

Int8
Optimised for int8 weights and activations

>100 Unique configurations
Solutions

Mobility
- DMS
- ADAS
- Hand Gesture Recognition
- ITS

Security & Surveillance
- Smart Access Control
- Safety Monitoring
- CCTV Monitoring

Retail
- Smart Inventory Management
- Customer Journey Analysis
- Smart Kiosk with Hand Gesture
Quantization (Activation Change)

- Unbound activation function such as ReLU or Swish could lose information when quantization is applied.
- It is better to change unbound activation to bound activation such as ReLU6.
Use Case 2. Cloud Usage Optimization

- Customers can minimize the inferencing cloud cost while maximize the performance (up to 5x performance improvement or 85% cost saving) through AI model compression.

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Original</td>
<td>Original</td>
<td>Original</td>
<td>Original</td>
</tr>
<tr>
<td>P100</td>
<td>X 3</td>
<td>24FPS</td>
<td>97.14%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>5.3X</td>
<td>5% faster</td>
<td>- 0.1%</td>
<td>Same</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Original</th>
<th>Original</th>
<th>Original</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100</td>
<td>X 8</td>
<td>22FPS</td>
<td>97.14%</td>
<td>$2,203</td>
</tr>
<tr>
<td>Lower</td>
<td>14% faster</td>
<td>- 0.1%</td>
<td>85% save</td>
<td>$391</td>
</tr>
</tbody>
</table>

netspresso

<table>
<thead>
<tr>
<th>Instance</th>
<th>Channel</th>
<th>Speed</th>
<th>Accuracy</th>
<th>Fee (Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P100</td>
<td>X 16</td>
<td>25FPS</td>
<td>97.03%</td>
<td>$1,540</td>
</tr>
<tr>
<td>Same</td>
<td>2X</td>
<td>14% faster</td>
<td>97.03%</td>
<td>Same</td>
</tr>
<tr>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td>$1,540</td>
<td></td>
</tr>
</tbody>
</table>
Use Case 4. Inventory Management Solution

- Need to reduce resources checking inventory levels in a large market.
- Collaboration with 2nd biggest retailer in KR.