AI FOR EVERYONE

Presented to: American Indian Science and Engineering Society
Presenter: Meghana Rao, Intel Corporation
25th March, 2020
NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit [http://www.intel.com/benchmarks](http://www.intel.com/benchmarks).

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit [http://www.intel.com/benchmarks](http://www.intel.com/benchmarks).

Intel® Advanced Vector Extensions (Intel® AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at [http://www.intel.com/go/turbo](http://www.intel.com/go/turbo).

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.
Agenda

• Introduction to Artificial Intelligence
• AI in the past and present day
• Intel and AI
• AI Journey
• Introduction to Machine Learning
• Introduction to Deep Learning
• Challenges in solving problems through AI
• Community Support
• QnA
INTRODUCTION TO AI
What is AI?

Training:
- Human tagged data
- Forward "Strawberry" error
- Backward "Bicycle"

Inference:
- Forward "Bicycle"?
- Model weights
- Many different approaches to AI
HISTORY AND REASONS FOR CURRENT MOMENTUM
AI has experienced several hype cycles, where it has oscillated between periods of excitement and disappointment.
WHY AI NOW? ACCESS TO DATA

DATA DELUGE (2019)

- 25 GB¹ per month
  Internet User
- 50 GB² per day
  Smart Car
- 3 TB² per day
  Smart Hospital
- 40 TB² per day
  Airplane Data
- 1 PB² per day
  Smart Factory
- 50 PB² per day
  City Safety


ANALYTICS CURVE

- FORESIGHT
  Predictive Analytics
- INSIGHT
  Diagnostic Analytics
- HINDSIGHT
  Descriptive Analytics
- FORECAST
  Prescriptive Analytics
- ACT/ADAPT
  Cognitive Analytics


INSIGHTS

- BUSINESS
- OPERATIONAL
- SECURITY
## AI Transformation Across Industries

### Consumer
- Smart Assistants
- Chatbots
- Search
- Personalization
- Augmented Reality
- Robots

### Health
- Enhanced Diagnostics
- Drug Discovery
- Patient Care
- Research
- Personalization
- Sensory Aids

### Finance
- Algorithmic Trading
- Fraud Detection
- Research
- Merchandising
- Loyalty
- Supply Chain
- Risk Mitigation

### Retail
- Support
- Experience
- Marketing
- Data Insights
- Safety & Security
- Resident Engagement
- Smarter Cities

### Government
- Defense
- Oil & Gas Exploration
- Smart Grid
- Operational Improvement
- Conservation

### Energy
- In-Vehicle Experience
- Automated Driving
- Aerospace
- Shipping
- Search & Rescue

### Transport
- Factory Automation
- Predictive Maintenance
- Precision Agriculture
- Field Automation

### Industrial
- Advertising
- Education
- Gaming
- Professional & IT Services
- Telco/Media
- Sports

### Other
- Source: Intel forecast
User-touch end point devices with lower power requirements such as laptops, tablets, smart home devices, drones

Small scale data centers, small business IT infrastructure, to few on-premise server racks and workstations

Large scale data centers such as public cloud or comms service providers, gov’t and academia, large enterprise IT

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
INTEL AI SOFTWARE
AI = ML + DL + MT

DEVELOPER TOOLS
App Developers
SW Platform Developer

TOPOLOGIES & MODELS
Data Scientist

FRAMEWORKS
Data Scientist

GRAPH
ML Performance Engineer

KERNEL
ML Performance Engineer

MACHINE LEARNING

- Intel Distribution for Python (SKlearn, Pandas)
- MLib and Mahout on Spark

DEEP LEARNING

- TensorFlow
- PyTorch
- PaddlePaddle
- mxnet
- Caffe
- BigDL
- ONNX

- nGraph
- PlaidML
- Intel Machine Learning Scaling Library (Intel MLSL)
- Deep Learning Boost
- Intel® Deep Neural Network Library (DNNL)

MANAGEMENT TOOLS

- ModelZoo
- OpenVINO

- DL Tools
  - NN Distiller, NLP Architect, RL Coach, CVAT

- Architect & DevOps
  - Containers
  - Data Analytics Reference Stack
  - Deep Learning Reference Stack

CPU

CPU = GPU = FPGA = CUSTOM

Red font products are the most broadly applicable SW products for AI users.
THE AI JOURNEY
MACHINE LEARNING
MACHINES LEARN IN TWO WAYS

Supervised Learning & Unsupervised Learning
**SUPERVISED LEARNING**

We train the model. We feed the model with correct answers. Model Learns and finally predicts.

We feed the model with “ground truth”.
EXAMPLES OF SUPERVISED LEARNING - CLASSIFICATION

Predict a label for an entity with a given set of features.

PREDICTION

SENTIMENT ANALYSIS
There are many metrics available* to measure performance, such as:

- **Accuracy**: how well predictions match true values.

- **Mean Squared Error**: average square distance between prediction and true value.

\[
\min_{\beta_0, \beta_1} \frac{1}{m} \sum_{i=1}^{m} \left( (\beta_0 + \beta_1 x_{obs}^{(i)}) - y_{obs}^{(i)} \right)^2
\]

*The wrong metric can be misleading or not capture the real problem.*
UNSUPERVISED LEARNING

Data is given to the model. Right answers are not provided to the model. The model makes sense of the data given to it.

Can teach you something you were probably not aware of in the given dataset.
EXAMPLE OF UNSUPERVISED LEARNING - CLUSTERING

Group entities with similar features

Market Segmentation

Play time in hours

Serious Gamers

Causal Gamers

Non Gamers

Age
ADDITIONAL MACHINE LEARNING EXAMPLES

Fraud Detection

Movie Recommendation

Recommending Similar news articles

Other brand names can be claimed as the property of others
WHAT IS THE LIMITATION WITH LINEAR CLASSIFIERS?

XOR
The counter example to all models
We need non-linear functions

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

WE NEED LAYERS USUALLY LOTS WITH NON-LINEAR TRANSFORMATIONS

XOR = (X1 and not X2) OR (Not X1 and X2)

Threshold to 0 or 1

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(1 x 1) + (0 x 1) < 1.5 = 0
(1x1) + (0x-2) + (0x1) = 1 > 0.5 = 1
How do deep learning networks learn? Each layer learns something.

Layer 1 ➔ Layer 2 ➔ Layer N ➔ Prediction

Zeieler & Fergus ‘13
HOW CAN I BUILD A NEURAL NETWORK?
**MOTIVATION FOR NEURAL NETS**

- Use biology as inspiration for mathematical model
- Get signals from previous neurons
- Generate signals (or not) according to inputs
- A neuron fires when it’s output > threshold
- Pass signals on to next neurons
- By layering many neurons, can create complex model
FULLY CONNECTED NEURAL NETWORK

- Multiple layers of stacked neurons forming a network (topology)
- Each neuron is connected to every neuron in subsequent layers
- Network topologies are constantly evolving based on complexity of problems being solved by AI
WHAT IS AN ACTIVATION FUNCTION?

• The output of a neuron could range from \(-\infty\) to \(+\infty\)
• How does it know when to fire?
• An activation function establishes a boundary for the output
• Many types of activation functions exist

![Graph showing different activation functions](image)
$z^{(2)} = xW^{(1)}$

$a^{(2)} = \sigma(z^{(2)})$

$W^{(1)}$ is a 3x4 matrix

$z^{(2)}$ is a 4-vector

$a^{(2)}$ is a 4-vector
CONTINUING THE COMPUTATION

For a single training instance (data point)

Input: vector \( x \) (a row vector of length 3)

Output: vector \( \hat{y} \) (a row vector of length 3)

\[
\begin{align*}
    z^{(2)} &= xW^{(1)} \\
    a^{(2)} &= \sigma(z^{(2)}) \\
    z^{(3)} &= a^{(2)}W^{(2)} \\
    a^{(3)} &= \sigma(z^{(3)}) \\
    z^{(4)} &= a^{(3)}W^{(3)} \\
    \hat{y} &= \text{softmax}(z^{(4)})
\end{align*}
\]
A FULLY CONNECTED NEURAL NETWORK WITH ACTIVATIONS

$\hat{y}$, the actual output may not be the expected output. The Network needs to be trained to get better accuracy.
HOW CAN I TRAIN A NEURAL NETWORK?
FORWARD PROPAGATION

• Calculate the Loss Function – compare the predictions to the ground truth

• How far the “actual output” is from “Ground Truth” determines how much more the network needs to learn to adjust it's output to minimize loss

Evaluate: $J(y_i, \hat{y}_i)$
BACKPROPAGATION

Want: \( \frac{\partial J(y_i, \hat{y}_i)}{\partial W_k} \)
APPLY GRADIENTS TO EVERY WEIGHT IN THE NETWORK

\[
\frac{\partial J}{\partial W^{(3)}} = (\hat{y} - y) \cdot a^{(3)}
\]

\[
\frac{\partial J}{\partial W^{(2)}} = (\hat{y} - y) \cdot W^{(3)} \cdot \sigma'(z^{(3)}) \cdot a^{(2)}
\]

\[
\frac{\partial J}{\partial W^{(1)}} = (\hat{y} - y) \cdot W^{(3)} \cdot \sigma'(z^{(3)}) \cdot W^{(2)} \cdot \sigma'(z^{(2)}) \cdot X
\]

- Recall that: \( \sigma'(z) = \sigma(z)(1-\sigma(z)) \) (Sigmoid activation function)
- Though they appear complex, above are easy to compute!
FULLY CONNECTED NEURAL NETWORK

VS.

CONVOLUTIONED NEURAL NETWORKS
More complicated problems can be solved by connecting multiple neurons together and using more complicated activation functions.

- Organized into layers of neurons.
- Each neuron is connected to every neuron in the previous layer.
- Each layer transforms the output of the previous layer and then passes it on to the next.
- Every connection has a separate weight
CONVOLUTIONAL NEURAL NETWORK
Convolutional neural networks reduce the required computation and are good for detecting features.

- Each neuron is connected to a small set of nearby neurons in the previous layer
- The same set of weights are used for each neuron
- Ideal for spatial feature recognition, Ex: Image recognition
- Cheaper on resources due to fewer connections
CNN FOR RECOGNIZING DIGITS
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.
IDENTIFYING FEATURES

Convolution

- Each element in the output is the result of a dot product between two vectors.

Detected the pattern!
CHALLENGES
CHALLENGES

• Availability of data
  • Data Sources
  • Data Shapes
  • Amount of Data
  • Data Preprocessing
  • Labelling the data
• Reducing possibilities for overfitting and under-fitting
• Human error in data labelling
• Human Bias
HOW CAN YOU LEARN MORE?
LEARN MORE AT THE INTEL® AI ACADEMY
For developers, students, instructors and startups

Get smarter using online tutorials, webinars, student kits and support forums

Educate others using available course materials, hands-on labs, and more

Get 4-weeks FREE access to the Intel® AI DevCloud, use your existing Intel® Xeon® Processor-based cluster, or use a public cloud service

Showcase your innovation at industry & academic events and online via the Intel AI community forum

software.intel.com/ai
RESOURCES

• Intel Developer Zone
  • https://software.intel.com

• Intel® AI Academy
  https://software.intel.com/ai-academy

• Intel® AI Student Kits
  https://software.intel.com/ai-academy/students/kits/
Q&A