Emerging AI Technologies on Intel® Client Platforms

Manuj Sabharwal, Anand Bodas, Neelay Pandit, Vivek Rane, Sherine Abdelhak
Client Computing Group, Artificial Intelligence Group, and Intel Software and Systems Group
Relevancy of Client* in AI space
Emerging AI Technologies on Client & Active Deployments
Federated Learning & Concept Examples
Performance & Optimization Guidance
Recap & Conclusion

* Client: your PC, Notebook, Clamshell, etc...
WHY AI ON CLIENT

TRUST & PRIVACY

NETWORK BANDWIDTH

RESPONSIVENESS

SERVICE COST

30%  20%  5%
**Examples of Emerging AI Technologies – Generative Adversarial Networks (GANs)**

**Functionality:**
- **Discriminator’s** goal is to correctly discover fake images (Fake (0))
- **Generator’s** goal is to fool the discriminator (Real (1))

---

1 Concept revived by Ian Goodfellow (2015)
USAGES OF GANS

IMAGE TRANSLATION - CYCLEGAN
Input Image → Van Gogh Paintings → Generated Image

SUPER-RESOLUTION - SRGAN
Input Low Res. Photo → Bicubic Upscaling → Super Resolution (using GANs)

IMAGE SYNTHESIS FROM TEXT - STACKGAN
“A small yellow bird with a black crown and a short black pointed beak”

Intel® HPC Developer Conference 2017
Examples of Emerging AI Technologies – Deep Reinforcement Learning (DRL)\(^1\)

- **Supervised learning**: training data is provided and is labeled
- **Unsupervised learning**: training data is provided and is not labeled
- **Reinforcement learning**: training data is in real-time and is in the form of a reward / punishment

RL Agent observes the state of an environment, performs an action, and optimizes its action based on the *reward/punishment* it receives from the environment; in *realtime*, in an *unsupervised* way

\(^1\) Concept made popular by adoption from Google Deepmind (2015)
USAGES OF DEEP REINFORCEMENT LEARNING

GAMING

DIALOGUE SYSTEMS

PERSONAL ASSISTANTS*

Amazon*, Google*, Cortana* and Siri* are 3rd party products.
Federated learning or Collaborative learning is where **multiple devices participate** in the machine learning process (training or inferencing).

Federated learning **decouples storage** from the machine learning process.

*Concept introduced by Google (April 2017)*
**FEDERATED GANS**

- For media streaming, super-resolution GAN can help video playback on client devices retain high quality when network access is diminished.

- Generator networks run on Client & periodically send their outputs to train a Global GAN (Discriminator / Generator Network).

---

**RX Content**  
**Decode**  
**SR-GAN**  
**TX Output**
• Personalized Gaming:
  • Actor-Critic algorithms such as A3C, A2C, ACKTR are run on single system
  • Actors can be split across systems and implement federated learning
  • Fast learner on Client for faster personalization
  • Slow learner on Server for better generalization
ACTIVE DEPLOYMENTS

ADOBE* SKETCH & SCRIBBLE PREVIEW¹

Sketch & Scribble Preview employing texture synthesis generative adversarial networks & running local on client

UNITY* REINFORCEMENT LEARNING FEATURE²

Employing Q-Learning for virtual agents to learn new skills in new environments

**PERFORMANCE OF GANS ON CLIENT**

**DCGAN Performance on IA**

Data on Intel 7th Generation Core Platform
MKL is Intel® Math Kernel Library

- **Vanilla TF**: 160 FPS
- **MKL NHWC Format**: 80 FPS
- **MKL NCHW Format**: 330 FPS
- **MKL NCHW & Fused BN**: 715 FPS

**RELATIVE PERFORMANCE GAIN**

- **2.2X** with MKL NCHW
- **4.5X** With Fused BN

---

The performance optimization shown is based on the 7th Generation Intel® Core™ i7 Mobile Processor Ubuntu® Linux® 16.04. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.
**PERFORMANCE OF RL ON CLIENT**

Data on Intel 7th Generation Core Platform
MKL is Intel® Math Kernel Library

The performance optimization shown is based on the 7th Generation Intel® Core™ i7 Mobile Processor Ubuntu® Linux® 16.04. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.
Perfomance Optimization of Background Segmentation on Client 3D Camera

- Optimizing high performance compute workload for CPU
  - Multi-core scalability
  - Intel AVX2 Instruction Set

The performance optimization shown is based on the 7th Generation Intel® Core™ i7 Mobile Processor Ubuntu® Linux® 16.04. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

*Personify https://www.personify.com/*
# Performance Optimization on Modern Platforms

## Hierarchical Parallelism

### Coarse-Grained
- Multi-node domain decomposition

### Fine-Grained Parallelism / Within Node
- Multi-level domain decomposition (ex. across layers)
- Data decomposition (layer parallelism)

## Scale Workload
- Improve load balancing
- Reduce synchronization events, all-to-all comms

## Utilize All Cores
- OpenMP, MPI, TBB...
- Reduce synchronization events, serial code
- Improve load balancing

## Vectorize / SIMD
- Unit strided access per SIMD lane
- High vector efficiency
- Data alignment

## Efficient Memory & Cache Use
- Blocking
- Data reuse
- Prefetching
- Memory allocation
DATA LAYOUT IMPACTS PERFORMANCE SIGNIFICANTLY

- Sequential access to avoid gather/scatter
- Have iterations in inner most loop to ensure high vector utilization
- Maximize data reuse; e.g. weights in a convolution layer

VARIOUS DATA FORMATS TO CHOOSE FROM

- The most popular formats are NCHW and NHWC (Image Number/Height/Width/Color)
- Intel® MKL is optimized for both formats. Do make sure to try both to see which one provides higher performance.

Better optimized for some operations vs.

Converting to/from optimized Layout is sometimes less expensive than operating on unoptimized Layout
EXAMPLE 2: CONVERSIONS

CONVERSIONS IMPACTS PERFORMANCE SIGNIFICANTLY

- End to end optimization can reduce conversions
- Staying in optimized layout as long as possible becomes one of the tuning goals
- Minimize the number of back and forth conversions
- Use of graph optimization techniques

Conversions impacts performance significantly.
SUMMARY FOR DEVELOPERS

WHERE TO GET HELP?
https://www.intelnervana.com/tensorflow/
https://www.tensorflow.org/performance/performance_guide#tensorflow_with_intel_mkl_dnn

OPTIMIZATION
Avoid Data format conversions between TensorFlow and MKL layouts
Maximize Parallelism with threads and SIMD instructions via MKL

DON'T BE FOOLED!
Significant performance headroom on Intel CPUs, GPUs, etc...
Close to 300x speedup in certain topologies
CONCLUSION

- Emerging AI being deployed on client
- Intel HW IPS meet inference needs
- Tuning of AI algorithms is crucial
- HPC & Collaborative learning are a future trend on client
Legal 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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

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/performance.

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.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described 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.

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.

© 2016 Intel Corporation. Intel, the Intel logo Intel®, Core™ and others are trademarks of Intel Corporation in the U.S. and/or other countries.

OpenCL™ has been licensed to Intel. The 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.
Optimization Notice

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 SSE3 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.

Notice revision #20110804