EFFICIENT INERENCE WITH TENSORRT

Han Vanholder
AI INFERENCING IS EXPLODING

2 Trillion Messages Per Day On LinkedIn

500M Daily active users of iFlyTek

140 Billion Words Per Day Translated by Google

60 Billion Video frames/day uploaded on Youtube
NETWORKS ARE GROWING
Bigger, Better, Exponential Compute

Speed/accuracy trade-offs for modern convolutional object detectors
April 2017, Jonathan Huang et al
NETWORKS ARE GROWING
Bigger, Better, Exponential Compute

OUTRAGEOUSLY LARGE NEURAL NETWORKS: THEMIXTURE-OF-EXPERTS SPARSELY-GATED LAYER
ICLR 2017, Noam Shazeer et all
DL FLOW
Pivot: Research to Production

Source Dataset
Import
Format...

Curated Dataset
Preprocess
Clean, clip, label, normalize, ...

Visualization

Model Zoo
Train
Score + Optimize, Visualization

Deploy
tune, compile + runtime

Inference & Microservices
Result*
inference, prediction

Rest API
DL FLOW
Pivot: Research to Production

Source Dataset

IMPORT
Format...

Curated Dataset

PREPROCESS
clean, clip,
label, Normalize, ...

VISUALIZATION

MODEL ZOO

TRAIN

SCORE + OPTIMIZE,
VISUALIZATION

DEPLOY
tune, compile + runtime

INFEERENCE & MICROSERVICES

RESULT *
inference, prediction

REST API

Format...

clean, clip,
label, Normalize, ...

VISUALIZATION

DEPLOY
tune, compile + runtime

INFEERENCE & MICROSERVICES

RESULT *
inference, prediction

REST API
DL FLOW
Pivot: Research to Production

Source Dataset

IMPORT
Format...

Curated Dataset

PREPROCESS
clean, clip, label, Normalize, ...

MODEL ZOO

TRAIN

DEPLOY
tune, compile + runtime

INFERENCEx MICROSERVICES

REST API

RESULT *
inference, prediction

Inference needs
Latency
Accuracy
Efficiency
Throughput
NVIDIA TENSORRT PROGRAMMABLE INFERENCING PLATFORM
NVIDIA TensorRT
Programmable Inference Accelerator

Maximize throughput and minimize latency
Deploy reduced precision without retraining and without accuracy loss
Train in any framework, deploy in TensorRT without overhead

developer.nvidia.com/tensorrt
# TENSORRT 3 INFERENCE PERFORMANCE

## Up to 40x Faster CNNs vs. CPU

ResNet50, under 7ms latency

<table>
<thead>
<tr>
<th>Throughput (Images/Sec)</th>
<th>CPU-Only</th>
<th>V100 + TensorFlow</th>
<th>P4 + TensorRT</th>
<th>V100 + TensorRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>140</td>
<td>305</td>
<td>1400</td>
<td>5700</td>
</tr>
</tbody>
</table>

Inference throughput (images/sec) on ResNet50. V100 + TensorFlow: NVIDIA TensorRT (FP16), batch size 39, Tesla V100-8xM2-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. P100 + TensorRT: NVIDIA TensorRT (FP16), batch size 10, Tesla P100-PCIE-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. V100 + TensorFlow: Preview of volta optimized TensorFlow (FP16), batch size 2, Tesla V100-PCIE-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. CPU-Only: Intel Xeon-D 1587 Broadwell-E CPU and Intel DL SDK. Score doubled to comprehend Intel’s stated claim of 2x performance improvement on Skylake with AVX512.

## Up to 140x Faster RNNs vs CPU

OpenNMT Neural Translation

<table>
<thead>
<tr>
<th>Throughput (Sentences/Sec)</th>
<th>CPU-Only + Torch</th>
<th>V100 + Torch</th>
<th>P4 + TensorRT</th>
<th>V100 + TensorRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>25</td>
<td>65</td>
<td>550</td>
</tr>
</tbody>
</table>

Inference throughput (sentences/sec) on OpenNMT 692M. V100 + TensorRT: NVIDIA TensorRT (FP32), batch size 64, Tesla V100-PCIE-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. P100 + TensorRT: NVIDIA TensorRT (FP32), batch size 64, Tesla P100-PCIE-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. V100 + Torch: Torch (FP32), batch size 4, Tesla V100-PCIE-16GB, ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On. CPU-Only: Torch (FP32), batch size 1, Intel ES-2690 @2.60GHz 3.5GHz Turbo (Broadwell) HT On.
TENSORRT 3 INFECTION PERFORMANCE

Up to 40x Faster CNNs vs. CPU

ResNet50, under 7ms latency

Up to 140x Faster RNNs vs CPU

OpenNMT Neural Translation

Inference throughput (images/sec) on ResNet50. V100 + TensorRT: NVIDIA TensorRT (FP16), batch size 39, Tesla V100-8X32GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. P100 + TensorRT: NVIDIA TensorRT (FP16), batch size 10, Tesla P100-PCIE-16GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. V100 + TensorFlow: Preview of volta optimized TensorFlow (FP16), batch size 2, Tesla V100-PCIE-16GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. CPU-Only: Intel Xeon-D 1587 Broadwell-EP CPU and Intel DL SDK. Score doubled to comprehend Intel’s stated claim of 2x performance improvement on Skylake with AVX512.

Inference throughput (sentences/sec) on OpenNMT 692M. V100 + TensorRT: NVIDIA TensorRT (FP32), batch size 64, Tesla V100-PCIE-16GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. P100 + TensorRT: NVIDIA TensorRT (FP32), batch size 64, Tesla P100-PCIE-16GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. V100 + Torch: Torch (FP32), batch size 4, Tesla V100-PCIE-16GB, ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. CPU-Only: Torch (FP32), batch size 1, Intel ES-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On.
NVIDIA TENSORRT 3
Volta TensorCore ● TensorFlow Importer ● Python API

Volta TensorCore Support

3.7x faster inference on Tesla V100 vs. Tesla P100 under 7ms real-time latency

Import TensorFlow Models

Optimize and deploy TensorFlow models up to 18x faster vs. TensorFlow framework

Python API

Improved productivity with easy to use Python API for data science workflows

TensorRT 3 RC is now available as a free download to members of NVIDIA Developer Program

developer.nvidia.com/tensorrt
Un-optimized network

TensorRT Optimized Network

- Vertical fusion of relu, bias and 1x1 conv
- Horizontal fusion of 1x1 CBR layers
- Eliminate concatenation layers

developer.nvidia.com/tensorrt
**TENSORRT LAYERS**

**Built-in**
- Activation: ReLU, tanh, sigmoid
- Concatenation
- Convolution
- Deconvolution
- Element wise operations
- Fully-connected
- LRN
- Pooling: max and average
- RNN
- Scaling
- SoftMax

**Custom Layer API**
- Application
- TensorRT
- CUDA Runtime

developer.nvidia.com/tensorrt
TENSORRT 3: TENSORFLOW IMPORTER AND PYTHON API

- Optimize and deploy TensorFlow models that are up to 18x faster vs. TensorFlow framework
- Improved productivity with easy to use Python API for Data Science workflows

AI Researchers
Data Scientists

Caffe ➔ TensorFlow ➔ Other Frameworks ➔ Python API ➔ Model Importer ➔ Optimized Model ➔ TensorRT Runtime

- Data center
- Embedded
- Automotive

developer.nvidia.com/tensorrt
import tensorrt

# create the inference engine object
mnist_engine = tensorrt.lite.Engine(framework="tf",
    path=DATA_DIR + "/mnist/lenet5_mnist_frozen.pb",
    max_batch_size=10,
    input_nodes="in":(1,28,28),
    output_nodes="out",
    preprocessors="in":normalize,
    postprocessors="out":argmax)

results = mnist_engine.infer(data)[0]  # run the inference
WEIGHT AND ACTIVATION PRECISION CALIBRATION

Maintain accuracy without retraining

<table>
<thead>
<tr>
<th>Model</th>
<th>FP32 TOP 1</th>
<th>INT8 TOP 1</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexnet</td>
<td>57.22%</td>
<td>56.96%</td>
<td>0.26%</td>
</tr>
<tr>
<td>Googlenet</td>
<td>68.87%</td>
<td>68.49%</td>
<td>0.38%</td>
</tr>
<tr>
<td>VGG</td>
<td>68.56%</td>
<td>68.45%</td>
<td>0.11%</td>
</tr>
<tr>
<td>Resnet-50</td>
<td>73.11%</td>
<td>72.54%</td>
<td>0.57%</td>
</tr>
<tr>
<td>Resnet-101</td>
<td>74.58%</td>
<td>74.14%</td>
<td>0.44%</td>
</tr>
<tr>
<td>Resnet-152</td>
<td>75.18%</td>
<td>74.56%</td>
<td>0.61%</td>
</tr>
</tbody>
</table>
NVIDIA PLATFORM SAVES DATA CENTERS

Game Changing Inference Performance

1 NVIDIA HGX with 8 Tesla V100 GPUs
3 KWatts

160 Xeon Scalable Processor CPU servers
65 KWatts
**INFERENCIE: THROUGHPUT & LATENCY**

Real Time, High Throughput: P4, TensorRT3

---

**ResNet50** Inf/s vs Batch (bigger is better)

- E5-2690v4 Inf/s
- P4 Inf/s
- P4 INT8 Inf/s

- **Latency (log ms)** vs Batch (smaller is better)

- E5-2690v4 (ms)
- P4 (ms)
- P4 INT8 (ms)

---

Inference throughput (images/sec) on ResNet50, NVIDIA benchmarking. P4 measured with TensorRT 3.0 RC. E5-2690v4 measured with Intel DL SDK.
**INFEINCE: THROUGHPUT & LATENCY**

Real Time, High Throughput: V100, TensorRT3

Inference throughput (sentences/sec) on OpenNMT 692M, NVIDIA benchmarking. V100 measured with TensorRT (FP32), E5-2690v4 measured with Torch (FP32).
DL FLOW
Pivot: Research to Production

Source Dataset
Curated Dataset

IMPORT
Format...

PREPROCESS
clean, clip, label, Normalize, ...

VISUALIZATION

MODEL ZOO

TRAIN

DEPLOY
tune, compile + runtime

Automated
with TensorRT

Rapid Deployment, High Productivity

A/B Testing,
Usage data

REST API

INFERENCE & MICROSERVICES

RESULT *

inference, prediction

MODEL ZOO

SCORE + OPTIMIZE, VISUALIZATION

Curated Dataset

PREPROCESS
clean, clip, label, Normalize, ..

IMPORT
Format...

Source Dataset

VISUALIZATION
“We measured that the new Azure instances running NVIDIA cards accelerated the inference on detection network by 3x! [...] Ultimately with Azure, ObjectStore, and NVIDIA technology we built an object detection system which can handle full Bing index worth of data under peak traffic [and] provide great experience to users.”

Source: Microsoft Bing Blog

“TensorRT is a real game changer. Not only does TensorRT make model deployment a snap but the resulting speed up is incredible: out of the box, BodySLAM™, our human pose estimation engine, now runs over two times faster than using CAFFE GPU inferencing.”

Source: Paul Kruszewski, CEO - WRNCH
# Datacenter GPUs for DL

<table>
<thead>
<tr>
<th></th>
<th>V100 DGX-1, HGX-1</th>
<th>V100</th>
<th>V100 FHHL</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form Factor</strong></td>
<td>SXM2</td>
<td>PCIe Double Wide</td>
<td>PCIe FHHL Single Wide</td>
<td>PCIe Low Profile (HHHL)</td>
</tr>
<tr>
<td><strong>Max Power</strong></td>
<td>300W</td>
<td>250W</td>
<td>150W</td>
<td>75W / 50W</td>
</tr>
<tr>
<td><strong>FP32</strong></td>
<td>15.7 TFLOP</td>
<td>13.3 TFLOP</td>
<td>10 TFLOP</td>
<td>5.5 TFLOPs</td>
</tr>
<tr>
<td><strong>TensorCore</strong></td>
<td>125 TFLOP</td>
<td>106 TFLOP</td>
<td>80 TFLOP</td>
<td>—</td>
</tr>
<tr>
<td><strong>Integer 8</strong></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>22 TOPs</td>
</tr>
</tbody>
</table>
AI revolutionizes every industry

Inferencing demands acceleration.

TensorRT Programmable Inference platform: throughput, latency, accuracy, efficiency

Download now

developer.nvidia.com/tensorrt