DEV DAY

Optimize your Machine Learning workloads

Julien Simon
Global Evangelist, AI & Machine Learning
@julsimon

Amazon SageMaker

Build

Pre-built notebooks for common problems

Built-in, high-performance algorithms

Train

One-click training

Hyperparameter optimization

P3DN, C5N
TensorFlow on 256 GPUs
Dynamic Training on MXNet
Automatic Model Tuning

Deploy

One-click deployment

Fully managed hosting with auto-scaling

Model compilation
Elastic inference
Inference pipelines

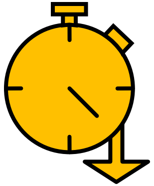
DEV DAY

Optimizing infrastructure

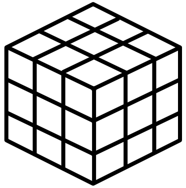
aws

Amazon EC2 P3dn

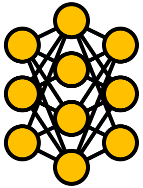
https://aws.amazon.com/blogs/aws/new-ec2-p3dn-gpu-instances-with-100-gbps-networking-local-nvme-storage-for-faster-machine-learning-p3-price-reduction/



Reduce machine learning training time



Better GPU utilization



Support larger, more complex models

KEY FEATURES

100Gbps of networking bandwidth

8 NVIDIA Tesla V100 GPUs

32GB of memory per GPU (2x more P3)

96 Intel Skylake vCPUs (50% more than P3) with AVX-512

Amazon EC2 C5n

https://aws.amazon.com/blogs/aws/new-c5n-instances-with-100-gbps-networking/

Intel Xeon Platinum 8000

Up to 3.5GHz single core speed

Up to 100Gbit networking

Based on Nitro hypervisor for bare metal-like performance

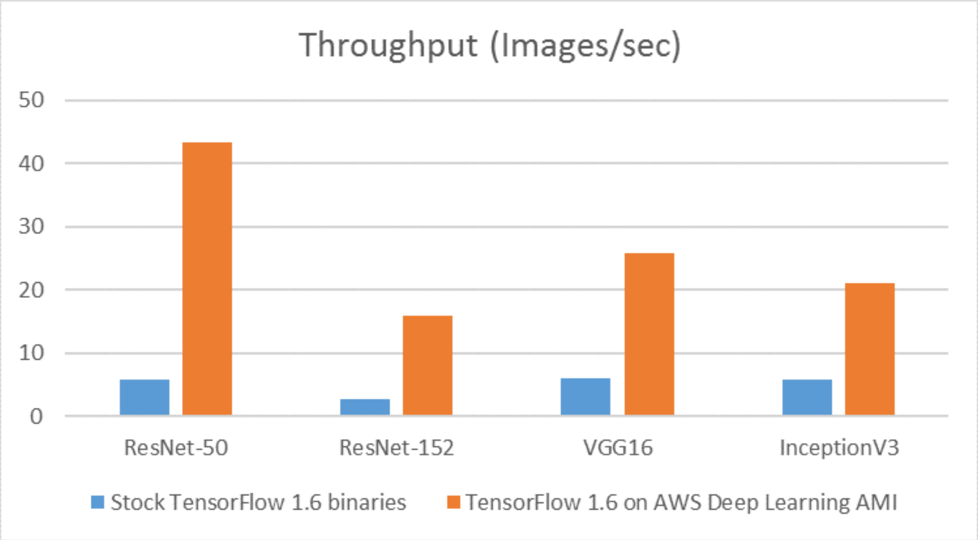
Instance Name	vCPUs	RAM	EBS Bandwidth	Network Bandwidth
c5n.large	2	5.25 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.xlarge	4	10.5 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.2xlarge	8	21 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.4xlarge	16	42 GiB	3.5 Gbps	Up to 25 Gbps
c5n.9xlarge	36	96 GiB	7 Gbps	50 Gbps
c5n.18xlarge	72	192 GiB	14 Gbps	100 Gbps

DEV DAY

Optimizing frameworks

aws

Making TensorFlow faster



https://aws.amazon.com/blogs/machine-learning/faster-training-with-optimized-tensorflow-1-6-on-amazon-ec2-c5-and-p3-instances/ (March 2018)

Training a ResNet-50 benchmark with the synthetic ImageNet dataset using our optimized build of TensorFlow 1.11 on a c5.18xlarge instance type is **11x faster** than training on the stock binaries.

https://aws.amazon.com/about-aws/whats-new/2018/10/chainer4-4_theano_1-0-2_launch_deep_learning_ami/ (October 2018)

Scaling TensorFlow near-linearly to 256 GPUs

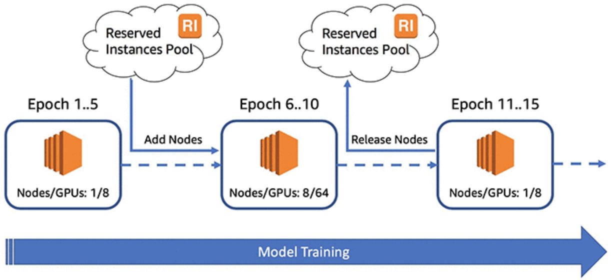
https://aws.amazon.com/about-aws/whats-new/2018/11/tensorflow-scalability-to-256-gpus/

Stock TensorFlow		AWS-Optimized TensorFlow	
65% scaling efficiency with 256 GPUs	→	90% scaling efficiency with 256 GPUs	Available with Amazon SageMaker and the AWS Deep Learning AMIs
30m training time	→	14m training time	

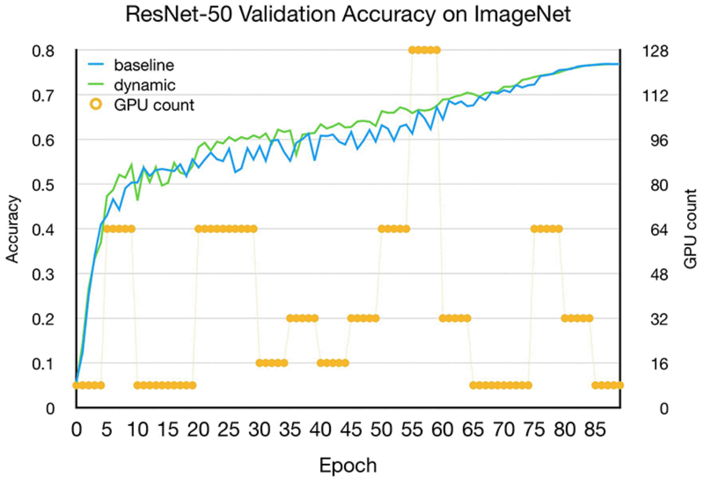
Dynamic training with Apache MXNet and RIs

https://aws.amazon.com/blogs/machine-learning/introducing-dynamic-training-for-deep-learning-with-amazon-ec2/

Use a variable number of instances for distributed training



No loss of accuracy



Coming soon
spot instances,
additional frameworks

DEV DAY

Optimizing models

Examples of hyperparameters

XGBoost

Tree depth
Max leaf nodes
Gamma
Eta
Lambda
Alpha
…

Neural Networks

Number of layers
Hidden layer width
Learning rate
Embedding dimensions
Dropout
…

Automatic Model Tuning

Finding the optimal set of hyper parameters

1. Manual Search ("I know what I'm doing")
2. Grid Search ("X marks the spot")
   - Typically training hundreds of models
   - Slow and expensive
3. Random Search ("Spray and pray")
   - Works better and faster than Grid Search
   - But… but… but… it's random!
4. HPO: use Machine Learning
   - Training fewer models
   - Gaussian Process Regression and Bayesian Optimization
   - You can now resume from a previous tuning job

Demo: HPO with Apache MXNet

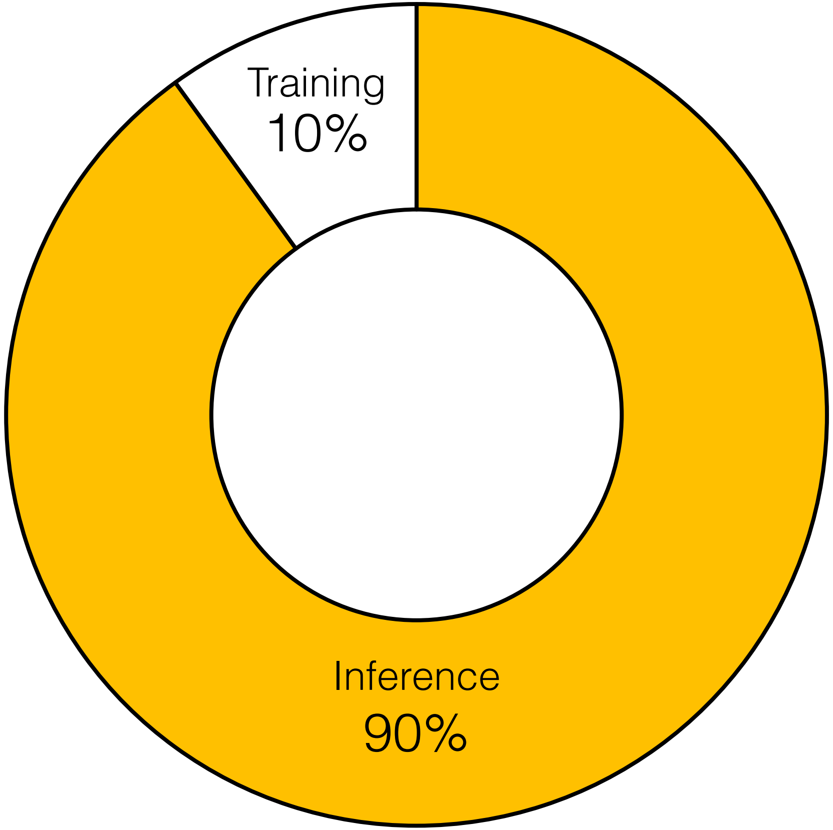
https://github.com/awslabs/amazon-sagemaker-examples/tree/master/hyperparameter_tuning/hpo_mxnet_mnist.ipynb

DEV DAY

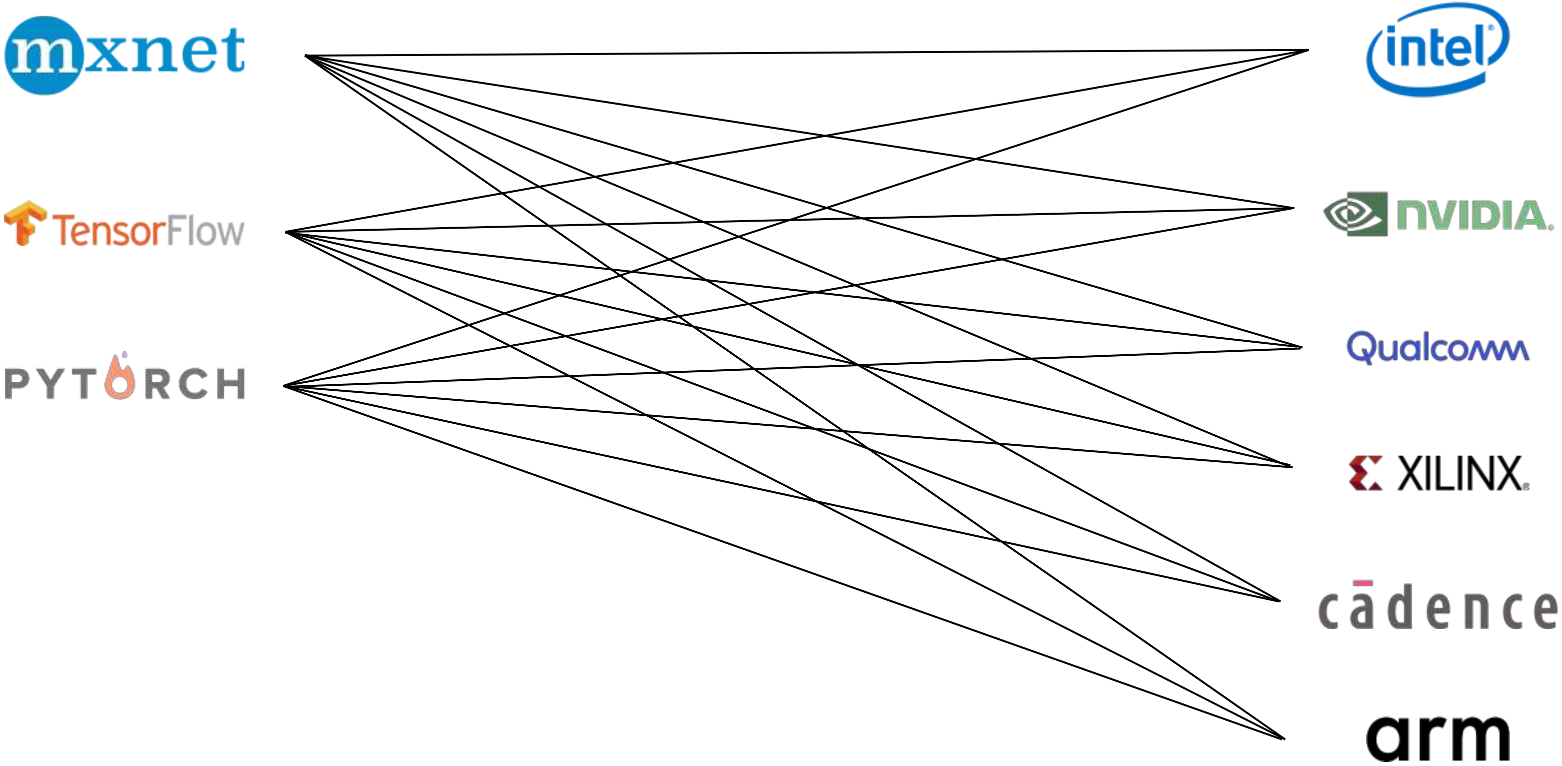
Optimizing inference

aws

Predictions drive complexity and cost in production



Model optimization is extremely complex



Amazon Neo: compiling models

https://aws.amazon.com/blogs/aws/amazon-sagemaker-neo-train-your-machine-learning-models-once-run-them-anywhere/

- Train once, run anywhere
- Frameworks and algorithms
    - TensorFlow, Apache MXNet, PyTorch, ONNX, and XGBoost
- Hardware architectures
    - ARM, Intel, and NVIDIA starting today
    - Cadence, Qualcomm, and Xilinx hardware coming soon
- Amazon SageMaker Neo is open source, enabling hardware vendors to customize it for their processors and devices: https://github.com/neo-ai/

Compiling ResNet-50 for the Raspberry Pi

Configure the compilation job

```
{
  "RoleArn":$ROLE_ARN,
  "InputConfig": {
   "S3Uri":"s3://jsimon-neo/model.tar.gz",
   "DataInputConfig": "{\"data\": [1, 3, 224, 224]}",
   "Framework": "MXNET"
  },
 "OutputConfig": {
  "S3OutputLocation": "s3://jsimon-neo/",
  "TargetDevice": "rasp3b"
 },
 "StoppingCondition": {
 "MaxRuntimeInSeconds": 300
 }
}
```

Compile the model

```
$ aws sagemaker create-compilation-job
--cli-input-json file://config.json
--compilation-job-name resnet50-mxnet-pi

$ aws s3 cp s3://jsimon-neo/model-
rasp3b.tar.gz .

$ gtar tfz model-rasp3b.tar.gz
compiled.params
compiled_model.json
compiled.so
```

Predict with the compiled model

```
from dlr import DLRModel
model = DLRModel('resnet50', input_shape,
output_shape, device)
out = model.run(input_data)
```

Demo: compiling a pre-trained PyTorch model with Neo

https://github.com/awslabs/amazon-sagemaker-examples/blob/master/advanced_functionality/pytorch_torchvision_neo/pytorch_torchvision_neo.ipynb

Amazon Elastic Inference

https://aws.amazon.com/blogs/aws/amazon-elastic-inference-gpu-powered-deep-learning-inference-acceleration/

Lower inference costs up to 75%

Match capacity to demand

Available between 1 to 32 TFLOPS

KEY FEATURES

Integrated with Amazon EC2, Amazon SageMaker, and Amazon DL AMIs

Support for TensorFlow, Apache MXNet, and ONNX with PyTorch coming soon

Single and mixed-precision operations

Demo:
Elastic Inference with TensorFlow

https://github.com/awslabs/amazon-sagemaker-examples/blob/master/sagemaker-python-sdk/tensorflow_iris_dnn_classifier_using_estimators/

Train & predict faster

Save time

Save money

Save your sanity (no plumbing!)

Getting started

https://ml.aws

https://aws.amazon.com/sagemaker

https://github.com/awslabs/amazon-sagemaker-examples

DEV DAY

Thank you!

Julien Simon
Global Evangelist, AI & Machine Learning
@julsimon
https://medium.com/@julsimon