Work with Machine Learning in Amazon SageMaker

Randy Ridgley, AWS Solutions Architect

Dec 2018
Solving Some Of The Hardest Problems In Computer Science

- Learning
- Language
- Perception
- Problem Solving
- Reasoning
Machine Learning at Amazon: A long heritage

- Personalized Recommendations
- Fulfillment automation & Inventory Management
- Drones
- Voice driven Interactions
- Inventing entirely new Customer experiences
ML @ AWS

OUR MISSION

Put machine learning in the hands of every developer and data scientist
The AWS Machine Learning Stack

APPLICATION SERVICES
- REKOGNITION
- REKOGNITION VIDEO
- POLLY
- TRANSCRIBE
- TRANSLATE
- COMPREHEND
- LEX

PLATFORMS
- Amazon SageMaker
- AWS DeepLens

FRAMEWORKS
- Caffe2
- CNTK
- mxnet
- PyTorch
- TensorFlow
- KERAS
- GLUON
Complete Control over Frameworks & Infrastructure

*For the data scientist, AI researcher, or advanced ML practitioner*

### Frameworks and Interfaces

<table>
<thead>
<tr>
<th>Frameworks</th>
<th>Interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe2</td>
<td>KERAS</td>
</tr>
<tr>
<td>CNTK</td>
<td>GLUON</td>
</tr>
<tr>
<td>mxnet</td>
<td></td>
</tr>
<tr>
<td>PyTorch</td>
<td></td>
</tr>
<tr>
<td>TensorFlow</td>
<td></td>
</tr>
</tbody>
</table>

### Infrastructure (GPU)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td></td>
</tr>
<tr>
<td>NVIDIA</td>
<td></td>
</tr>
<tr>
<td>Tesla V100</td>
<td></td>
</tr>
<tr>
<td>GPUs</td>
<td></td>
</tr>
<tr>
<td>(14x faster than P2)</td>
<td></td>
</tr>
<tr>
<td>5,120 Tensor cores</td>
<td>1 Petaflop of compute</td>
</tr>
<tr>
<td>128GB of memory</td>
<td>NVLink 2.0</td>
</tr>
<tr>
<td></td>
<td>Machine Learning AMIs</td>
</tr>
</tbody>
</table>

### Infrastructure (CPU)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C5</td>
<td></td>
</tr>
<tr>
<td>Intel Xeon</td>
<td></td>
</tr>
<tr>
<td>3.0 GHz Skylake CPU</td>
<td></td>
</tr>
<tr>
<td>(25% better perf/price than C4)</td>
<td></td>
</tr>
<tr>
<td>72 vCPUs</td>
<td>AVX 512</td>
</tr>
<tr>
<td>144 GB of memory</td>
<td>Nitro Hypervisor</td>
</tr>
<tr>
<td></td>
<td>Machine Learning AMIs</td>
</tr>
</tbody>
</table>
Machine Learning Process - Review
The Machine Learning Process

Business Problem –

ML problem framing

Data Collection

Data Integration

Data Preparation & Cleaning

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Are Business Goals met?

No

Yes

Re-training

Monitoring & Debugging

Model Deployment

Predictions

Data Augmentation

Feature Augmentation

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Data Visualization & Analysis

Business Problem – ML problem framing

- Help formulate the right questions
  - Domain Knowledge

Data Collection → Data Integration → Data Preparation & Cleaning → Data Visualization & Analysis

Feature Augmentation

Feature Engineering → Model Training & Parameter Tuning → Model Evaluation

Are Business Goals met?

No → Re-training

Yes → Model Deployment

Monitoring & Debugging

Predictions
Integration: The Data Architecture

Business Problem –

ML problem framing

- Data Collection
- Data Integration
- Data Preparation & Cleaning
- Data Visualization & Analysis
- Feature Engineering
- Model Training & Parameter Tuning
- Model Evaluation
- Are Business Goals met?
- Predictions
- Model Deployment
- Monitoring & Debugging

Retraining

Data Augmentation

• Build the data platform:
  • Amazon S3
  • AWS Glue
  • Amazon Athena
  • Amazon EMR
  • Amazon Redshift Spectrum

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Why We built Amazon SageMaker: The Model Training Undifferentiated Heavy Lifting

- Setup and manage Notebook Environments
- Setup and manage Training Clusters
- Write Data Connectors
- Scale ML algorithms to large datasets
- Distribute ML training algorithm to multiple machines
- Secure Model artifacts

Data Collection

Data Integration

Data Preparation & Cleaning

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Retraining

Monitoring & Debugging

Model Deployment

Predictions

Are Business Goals met?

No

Yes

ML problem framing
Where should you spend your time?

*Design training experiments*
Where should you spend your time?

Run scalable training
Where should you spend your time?

Deploy and operate

Build → Train → Deploy
Amazon SageMaker

Build, train, tune, and host your own models

Pre-built notebooks for common problems
Built-in, high performance algorithms

Build, train, tune, and deploy your own models.
Amazon SageMaker

Build, train, tune, and host your own models

BUILD

- Pre-built notebooks for common problems
- Built-in, high performance algorithms

TRAIN & TUNE

- One-click training
- Hyperparameter Tuning

DEPLOY
Amazon SageMaker

Build, train, tune, and host your own models

BUILD

- Pre-built notebooks for common problems
- Built-in, high performance algorithms

TRAIN & TUNE

- One-click training
- Hyperparameter Tuning

DEPLOY

- One-click deployment
- Fully managed hosting with auto-scaling

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Amazon SageMaker

Build, train, tune, and host your own models

Compliance and audit capabilities

Pay as you go

End-to-end encryption with KMS

BUILD

Pre-built notebooks for common problems

Built-in, high performance algorithms

TRAIN & TUNE

One-click training

Hyperparameter Tuning

DEPLOY

One-click deployment

Fully managed hosting with auto-scaling

End-to-end VPC support

Metadata and experiment management capabilities

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Key Advantages of Amazon SageMaker
Zero Setup For Exploratory Data Analysis

- Recommendations/Personalization
- Fraud Detection
- Forecasting
- Image Classification
- Churn Prediction
- Marketing Email/Campaign Targeting
- Log processing and anomaly detection
- Speech to Text
- More…

“Just add data”

Notebook Instances

Access to S3 Data Lake

ETL Access to AWS Database services

Authoring & Notebooks

EBS

VPC

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Amazon SageMaker: 10x better algorithms

Algorithms

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- And More!

Amazon provided Algorithms

Bring Your Own Script (SM builds the Container)

Bring Your Own Algorithm (You build the Container)

SM Estimators in Apache Spark
Managed Distributed Training with Flexibility

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- And More!

Amazon provided Algorithms

**Bring Your Own Script (SM builds the Container)**

**Bring Your Own Algorithm (You build the Container)**

**SM Estimators in Apache Spark**

**CPU**

**GPU**

**Hyperparameter Tuning**

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Easy Model Deployment to Amazon SageMaker

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!
Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Inference Code

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

Amazon ECR

Create a Model

ModelName: prod

Amazon SageMaker
Amazon ECR

InstanceType: c3.4xlarge
InitialInstanceCount: 3
maxInstanceCount: 10
ModelName: prod
VariantName: primary
InitialVariantWeight: 50

ProductionVariant

Create weighted ProductionVariants

Version of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!
Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Amazon ECR

Inference Image

Model Artifacts

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

Amazon SageMaker

ProductionVariant

A/B Testing Models
Create an EndpointConfiguration from one or many ProductionVariant(s)

InstanceType: c3.4xlarge
InitialInstanceCount: 3
maxInstanceCount: 10
ModelName: prod
VariantName: primary
InitialVariantWeight: 50

Easy Model Deployment to Amazon SageMaker

Model versions

EndpointConfiguration

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Amazon ECR

Model Artifacts

Inference Image

Model versions

EndpointConfiguration

Inference Endpoint

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

ProductionVariant

One-Click!

InstanceType: c3.4xlarge
InitialInstanceCount: 3
maxInstanceCount: 10
ModelName: prod
VariantName: primary
InitialVariantWeight: 50

Amazon SageMaker

Amazon Provided Algorithms

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Easy Model Deployment to Amazon SageMaker

- Auto-Scaling Inference APIs
- A/B Testing (more to come)
- Low Latency & High Throughput
- Bring Your Own Model
- Python SDK
Amazon SageMaker: Algorithm and Framework Support

Data → Built-In Algorithms → Deep Learning Frameworks → Custom/Docker → Amazon SageMaker → Training (single or distributed cluster)
Sample your data...

... explore and refine models in a single notebook instance

Use the same code to train on the full dataset in a cluster of GPU instances

... deploy to production
Bring Your Own Algorithm

Choose your own framework

... add algorithm code to a Docker container ...

... publish to Amazon ECR
Amazon SageMaker Local Mode

Enabling experimentation speed

Train with local notebooks
Train on notebook instances
PetaFLOP training on p3.16xl
Go distributed with one line of code

Same containers
Amazon SageMaker Automatic Model Tuning

Hyperparameter Optimizer

<table>
<thead>
<tr>
<th>Decision Trees</th>
<th>Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree depth</td>
<td>Number of layers</td>
</tr>
<tr>
<td>Max leaf nodes</td>
<td>Hidden layer width</td>
</tr>
<tr>
<td>Gamma</td>
<td>Learning rate</td>
</tr>
<tr>
<td>Eta</td>
<td>Embedding dimensions</td>
</tr>
<tr>
<td>Lambda</td>
<td>Dropout</td>
</tr>
<tr>
<td>Alpha</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

“Hyperparameters”
(algorithm parameters that significantly affect model quality)
Amazon SageMaker Automatic Model Tuning

Hyperparameter Optimization

<table>
<thead>
<tr>
<th>model name</th>
<th>objective metric</th>
<th>eta</th>
<th>max_depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>model1</td>
<td>0.08</td>
<td>0.07</td>
<td>6</td>
</tr>
<tr>
<td>model2</td>
<td>0.75</td>
<td>0.09</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
from sagemaker.tuner import IntegerParameter, CategoricalParameter, ContinuousParameter, HyperparameterTuner

hyperparameter_ranges = {
    'batch_size': CategoricalParameter([16, 32, 64, 128]),
    'learning_rate': ContinuousParameter(0.0009, 0.01),
    'epochs': IntegerParameter(10, 100)}

objective_metric_name = 'Validation-accuracy'
metric_definitions = [{'Name': 'Validation-accuracy',
                       'Regex': 'Validation-accuracy=(\[0-9\].+)?'}]

tuner = HyperparameterTuner(estimator,
                            objective_metric_name,
                            hyperparameter_ranges,
                            metric_definitions,
                            max_jobs=20,
                            max_parallel_jobs=4)

tuner.fit(dataset_location, job_name=job_name)
SageMaker Sample End-to-End Architecture: Style Transfer

**Build**
- SageMaker Notebooks
- Training Algorithm
- SageMaker Training
- Amazon ECR
- Code Commit
- Code Pipeline

**Train**
- Coco dataset
- static website hosted on S3
- Web assets on Cloudfront

**Deploy**
- Inference requests
- AWS Lambda
- API Gateway

- End-to-end VPC support
- End-to-end encryption with KMS
Recent Launches
Amazon SageMaker Ground Truth

Build highly accurate training datasets and reduce data labeling costs by up to 70% using machine learning.
How it works

Raw Data
How it works

Raw Data -> Human Annotations
How it works

Raw Data → Human Annotations → Active Learning Model → Automatic Annotations → Training Data
How it works

Raw Data → Human Annotations → Active Learning Model → Automatic Annotations → Human Annotations → Training Data
How it works

Raw Data → Human Annotations → Active Learning Model → Automatic Annotations → Training Data

Human Annotations → Human Annotations
Creating training data

- Mechanical turk workers
- Private labeling workforce
- Third-party vendors
Amazon SageMaker ground truth
Label machine learning training data easily and accurately

Quickly label training data
Easily integrate human labelers
Get accurate results

KEY FEATURES

Automatic labeling via machine learning
Ready-made and custom workflows for image bounding box, segmentation, and text
Private and public human workforce
Label management
AWS Marketplace for machine learning
ML algorithms and models available instantly

Browse or search AWS Marketplace

Subscribe in a single click

Available in Amazon SageMaker

KEY FEATURES

SELLERS
Automatic labeling via machine learning
IP protection
Automated billing and metering

BUYERS
Broad selection of paid, free, and open-source algorithms and models
Data protection

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Over 150 models and algorithms available

SELECTED VENDORS

SOME OF THE AVAILABLE ALGORITHMS AND MODELS

Natural Language Processing  Grammar & Parsing  Text OCR  Computer Vision  Named Entity Recognition  Video Classification
Speech Recognition  Text-to-Speech  Speaker Identification  Text Classification  3D Images  Anomaly Detection
Text Generation  Object Detection  Regression  Text Clustering  Handwriting Recognition  Ranking

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Amazon SageMaker Neo

A deep learning model compiler that lets customers train models once, and run them anywhere with up to 2X improvement in performance.
Train once, run anywhere
Amazon SageMaker Neo
Train once, run anywhere with 2x the performance

Key Features

- Get accuracy and performance
- Automatic optimization
- Broad framework support
- Broad hardware support

Compiler & run-time are open source | 1/10th the size of original models

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Amazon SageMaker RL

New machine learning capabilities in Amazon SageMaker to build, train and deploy with reinforcement learning
Amazon SageMaker RL
Reinforcement learning for every developer and data scientist

**KEY FEATURES**

- TensorFlow, Apache MXNet, Intel Coach, and Ray RL support
- 2D & 3D physics environments and OpenAI Gym support
- Supports Amazon Sumerian and Amazon RoboMaker
- Example notebooks and tutorials
Useful Amazon SageMaker Resources

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

https://github.com/aws-samples/aws-ml-vision-end2end

https://github.com/juliensimon

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

https://github.com/aws/sagemaker-spark
Thank you!

Workshop: