Developing Deep Learning Models for Computer Vision with Amazon EC2 P3 Instances

Wenming Ye
Sr. Solutions Architect AI/ML
Amazon Web Services
Agenda

• Deep Learning
• Computer Vision Scenarios and DL Architecture
• P3 Instance Details
• Learning Resources
Deep Learning in Context

AI

ML

DL
Generalization Performance

`rank \approx human\ effort`

- SVM
- Medium NN
- Random Forest
- Bayes Net
- Ruleset
- Genetic Agents

![Graph showing model accuracy vs dataset size for different models.](image)
Scenarios and DL Architecture

• **Architecture**
• **Convolutional Neural Network (CNN)**

• **Scenarios**
• Image Classification
• Object Detection
• Image Segmentation
• Visual Search
• GANs for Item Generation
Convolution Neural Network
Deep Learning in Computer Vision

Convolutional neural network

Explore spatial information with convolution layers

Layer 1

Layer 2

Output

\[ p(\text{cat}) \]

\[ p(\text{dog}) \]

Img src: https://leonardoaraujo.gitbooks.io/artificial-intelligence
Demo: Convolution Neural Network
Object Detection
Object Detection
Single Shot Detector
Image Segmentation
Image Segmentation
Kitti
Visual Search
Visual Search
Architecture
Pipeline Stages

Image Query Processing
  data normalization / augmentation

Embedding
  DNN model (s)

kNN + Ranking
  post-processing, de-dup
Demo: Amazon Shopping
## Domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generic</strong></td>
<td>Optimized for a broad range of image classification tasks. If none of the other domains are appropriate, or you are unsure of which domain to choose, select the Generic domain.</td>
</tr>
<tr>
<td><strong>Food</strong></td>
<td>Optimized for photographs of dishes as you would see them on a restaurant menu. If you want to classify photographs of individual fruits or vegetables, use the Food domain.</td>
</tr>
<tr>
<td><strong>Landmarks</strong></td>
<td>Optimized for recognizable landmarks, both natural and artificial. This domain works best when the landmark is clearly visible in the photograph. This domain works even if the landmark is slightly obstructed by people in front of it.</td>
</tr>
<tr>
<td><strong>Retail</strong></td>
<td>Optimized for images that are found in a shopping catalog or shopping website. If you want high precision classifying between dresses, pants, and shirts, use this domain.</td>
</tr>
<tr>
<td><strong>Adult</strong></td>
<td>Optimized to better define adult content and non-adult content. For example, if you want to block images of people in bathing suits, this domain allows you to build a custom classifier to do that.</td>
</tr>
<tr>
<td><strong>Compact domains</strong></td>
<td>Optimized for the constraints of real-time classification on mobile devices. The models generated by compact domains can be exported to run locally.</td>
</tr>
</tbody>
</table>
Embedding = Learned Representation Space
Demo Notebook + Tensorboard

https://github.com/miroenev/teach_DL/blob/master/data_buckets_fashion_mnist_cnn_ae.ipynb
Generative Adversarial Networks (GANs)
GAN Overview

Random Noise → Generative Model → Real/Fake Label

Fake Img → Discriminative Model → Real Img

MNIST Examples
GANs @ Celebrity Faces
Helping ShopBop to design new shoes using Progressive GANs
Video: ShopBop Progressive GAN Demo
Amazon EC2 P3 and P3dn Instances

One of the fastest, most powerful GPU instances in the cloud

• Up to eight NVIDIA Tesla V100 GPUs

• 1 PetaFLOPs of computational performance – *Up to 14x better than P2*

• 300 GB/s GPU-to-GPU communication (NVLink) – *9X better than P2*

• 16GB GPU memory with 900 GB/sec peak GPU memory bandwidth
Use-Cases for P3 Instances

<table>
<thead>
<tr>
<th>Machine Learning/AI</th>
<th>High Performance Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Language Processing</td>
<td>Computational Fluid Dynamics</td>
</tr>
<tr>
<td>Image and Video recognition</td>
<td>Financial and Data Analytics</td>
</tr>
<tr>
<td>Autonomous vehicle systems</td>
<td>Weather Simulation</td>
</tr>
<tr>
<td>Recommendation Systems</td>
<td>Computational Chemistry</td>
</tr>
</tbody>
</table>

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
# P3 Instances Details

<table>
<thead>
<tr>
<th>Instance Size</th>
<th>GPUs</th>
<th>GPU Peer to Peer</th>
<th>vCPUs</th>
<th>Memory (GB)</th>
<th>Network Bandwidth</th>
<th>EBS Bandwidth</th>
<th>On-Demand Price/hr*</th>
<th>1-yr RI Effective Hourly*</th>
<th>3-yr RI Effective Hourly*</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3.2xlarge</td>
<td>1</td>
<td>No</td>
<td>8</td>
<td>61</td>
<td>Up to 10Gbps</td>
<td>1.7Gbps</td>
<td>$3.06</td>
<td>$1.99 (35% Disc.)</td>
<td>$1.23 (60% Disc.)</td>
</tr>
<tr>
<td>P3.8xlarge</td>
<td>4</td>
<td>NVLink</td>
<td>32</td>
<td>244</td>
<td>10Gbps</td>
<td>7Gbps</td>
<td>$12.24</td>
<td>$7.96 (35% Disc.)</td>
<td>$4.93 (60% Disc.)</td>
</tr>
<tr>
<td>P3.16xlarge</td>
<td>8</td>
<td>NVLink</td>
<td>64</td>
<td>488</td>
<td>25Gbps</td>
<td>14Gbps</td>
<td>$24.48</td>
<td>$15.91 (35% Disc.)</td>
<td>$9.87 (60% Disc.)</td>
</tr>
</tbody>
</table>

Note: P3 instances are also supported via Spot Instances with up to 70% discount

## Regional Availability

P3 instances are generally available in AWS US East (Northern Virginia), US East (Ohio), US West (Oregon), EU (Ireland), Asia Pacific (Seoul), Asia Pacific (Tokyo), AWS GovCloud (US) and China (Beijing) Regions.

## Framework Support

P3 instances and their V100 GPUs supported across all major frameworks (such as TensorFlow, MXNet, PyTorch, Caffe2 and CNTK)
New larger P3 size – P3dn.24xlarge

Optimized for distributed ML training

- One of the most powerful GPU instances available in the cloud
- 100 Gbps of networking throughput
- 96 vCPU using AWS customer Skylake CPUs and 768 GB of system memory
- Based on NVIDIA’s latest GPU Tesla V100 with 32 GB of memory

<table>
<thead>
<tr>
<th>Instance size</th>
<th>GPUs</th>
<th>GPU memory</th>
<th>GPU peer to peer</th>
<th>vCPUs</th>
<th>CPU type</th>
<th>Memory (GB)</th>
<th>Network Bandwidth</th>
<th>Amazon EBS Bandwidth</th>
<th>Local instance storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3.2xlarge</td>
<td>1 x V100</td>
<td>16 GB/GPU</td>
<td>No</td>
<td>8</td>
<td>Broadwell</td>
<td>61</td>
<td>Up to 10 Gbps</td>
<td>1.7 Gbps</td>
<td>NA</td>
</tr>
<tr>
<td>P3.8xlarge</td>
<td>4 x V100</td>
<td>16 GB/GPU</td>
<td>NVLink</td>
<td>32</td>
<td>Broadwell</td>
<td>244</td>
<td>10 Gbps</td>
<td>7 Gbps</td>
<td>NA</td>
</tr>
<tr>
<td>P3.16xlarge</td>
<td>8 x V100</td>
<td>16 GB/GPU</td>
<td>NVLink</td>
<td>64</td>
<td>Broadwell</td>
<td>488</td>
<td>25 Gbps</td>
<td>14 Gbps</td>
<td>NA</td>
</tr>
<tr>
<td>P3dn.24xlarge</td>
<td>8 x V100</td>
<td>32 GB/GPU</td>
<td>NVLink</td>
<td>96</td>
<td>Skylake</td>
<td>768</td>
<td>100 Gbps</td>
<td>14 Gbps</td>
<td>2 TB NVMe</td>
</tr>
</tbody>
</table>

Latest NVIDIA V100 GPU with 32 GB memory for large models and higher batch sizes
96 Skylake vCPUs with support for AVX-512 instructions for pre-processing of training data
100 Gbps of networking throughput for large-scale distributed training & fast data access

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Scaling Performance using Distributed Training

• Using single P3 instances, with Volta GPUs, customers can cut down training times of their machine learning models from days to a few hours.

• Using distributed training via multiple P3 instances, high performance networking and storage solutions, customers can further cut down their time-to-train from hours in to minutes.

• Example – We been able to train ResNet-50 to Top1 validation accuracy of 76% in 47 mins using 8 P3.16xlarge instances.

Toyota Research Institute's (TRI) mission is to improve the quality of human life through advances in artificial intelligence, automated driving, and robotics. The core compute capability needed by TRI to accelerate the training of their machine learning models is powered by multiple Amazon EC2 P3 instances.

With P3 instances, TRI is seeing a 4X faster time-to-train than the P2 instances they had used previously. This gives them significant agility to optimize and retrain their models quickly and deploy them in their test cars or simulations environment for further testing. Furthermore, the significant performance improvement in P3s over P2s, coupled with pay-as-you-go model translates to lower operating costs.
P3 Instance Demo
A fully managed service that enables data scientists and developers to quickly and easily build machine-learning based models into production smart applications.
Amazon SageMaker

Build

- Highly-optimized machine learning algorithms

Train

- One-click training for ML, DL, and custom algorithms

Deploy

- Fully-managed hosting at scale
- Deployment without engineering effort

- Easier training with hyperparameter optimization

Amazon SageMaker

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Amazon SageMaker: Launch Customers

- intuit
- DigitalGlobe
- ZipRecruiter
- Hotels.com
- THOMSON REUTERS
Amazon SageMaker

1. Notebook Instances
2. Algorithms
3. ML Training Service
4. ML Hosting Service

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
SageMaker Demo
Hands on Labs to Try

• Image classification
• Object detection
• URL: https://bit.ly/2TNdE10
Where to go from here?

- Other Topics (coming soon)
  - Other DL architectures [Auto Encoder, RNNs, and more]
  - DL Inference
    - SageMaker hosting
    - Elastic Inference
    - Inference at the Edge [Deeplens, Green Grass Inference]
- Learning Resources:
  - New Book: https://en.diveintodeeplearning.org/chapter_deep-learning-computation/index.html
Thank you!

Wenming Ye
twitter: @wenmingye