- Format the EBS volume, mount it on `/data`, and then change the owner to `ec2-user:ec2-user`. You may refer to https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-using-volumes.html if you do not know how to mount it.
- Download MXNet repository and TensorFlow models repository.

```
$ cd /data $ git clone https://github.com/tensorflow/models.git
$ git clone https://github.com/apache/incubator-mxnet.git # or,
you can just type `pip install mxnet`
```

  - [Optional] For your convenience, use symbolic link such that:

```
[ec2-user@ip-172-31-34-246 data]$ ls -l
합계 13864
drwxrwxr-x 4 ec2-user ec2-user       37  9월  16 02:20 im2rec
lrwxrwxrwx 1 ec2-user ec2-user       41  9월  10 02:38 imagenet -> /data/models/research/inception/inception
drwxrwxr-x 7 ec2-user ec2-user      249  9월  10 02:22 models
lrwxrwxrwx 1 ec2-user ec2-user       44  9월  16 00:29 mxnet -> /usr/local/lib/python2.7/site-packages/mxnet
-rw------- 1 ec2-user ec2-user 14193041  9월  17 23:46 nohup.out
drwxrwxr-x 4 ec2-user ec2-user       98  9월  16 00:52 opencv
[ec2-user@ip-172-31-34-246 data]$
```

- [Important Step] You need to install OpenCV also. (Both 3.x and 4.x work well). If you do not install OpenCV, then you cannot convert ImageNet raw data to RecordIO files since `im2rec.py` utilizes some OpenCV functions. You may refer to https://www.pyimagesearch.com/2018/08/15/how-to-install-opencv-4-on-ubuntu/.
- [Caution] I strongly recommend to use Python2 instead of Python3 because many codes of Tensorflow models repository does not work on Python3. Please refer to https://stackoverflow.com/questions/38546672/inception-build-imagenet-data-py-typeerror-rgb-has-type-class-str-but-ex.

Downloading ImageNet

Please note that ImageNet server is sometimes unstable so download speed is not fast, taking 4 to 5 days.

Method 1

- Go to http://www.image-net.org/, sign up, and get your own username and access key.

- Extract the training set

```
$ mkdir train $ mv ILSVRC2012_img_train.tar train $ cd train $ tar xf ILSVRC2012_img_train.tar $ find . -name "*.tar" | while read NAME ; do mkdir -p "${NAME%.tar}"; tar -xvf "${NAME}" -C "${NAME%.tar}"; rm -f "${NAME}"; done
```

  - After extracting the training set, check if the number of directories is 1,000 (class 1 is n01728572 and class 1000 is n15075141).

- Extract bounding boxes

```
$ mkdir bounding_boxes $ mv ILSVRC2012_bbox_train_v2.tar.gz bounding_boxes $ mv ILSVRC2012_bbox_val_v3.tgz bounding_boxes $ cd bounding_boxes $ tar xzf ILSVRC2012_bbox_val_v3.tgz $ mkdir train $ mv ILSVRC2012_bbox_train_v2.tar.gz train $ cd train $ tar xzf ILSVRC2012_bbox_train_v2.tar.gz
```

```
-rw-rw-r--    1 ec2-user ec2-user     19537608 10월  1 00:01 ILSVRC2012_bbox_train_v2.tar.gz
-rw-rw-r--    1 ec2-user ec2-user      2221290  9월 18  2012 ILSVRC2012_bbox_val_v3.tgz
-rw-rw-r--    1 ec2-user ec2-user 147897477120  6월 14  2012 ILSVRC2012_img_train.tar
-rw-rw-r--    1 ec2-user ec2-user   6744924160  6월 14  2012 ILSVRC2012_img_val.tar
drwxrwxr-x 1002 ec2-user ec2-user        32768  9월 17 01:36 bounding_boxes
-rw-r--r--    1 root     root        29709928  9월 17 06:02 imagenet_2012_bounding_boxes.csv
drwxrwxr-x 1002 ec2-user ec2-user        32768  9월 17 02:29 train
drwxrwxr-x 1002 ec2-user ec2-user      2691072  9월 17 05:54 validation
(base) [ec2-user@ip-172-31-35-5 data]$
```

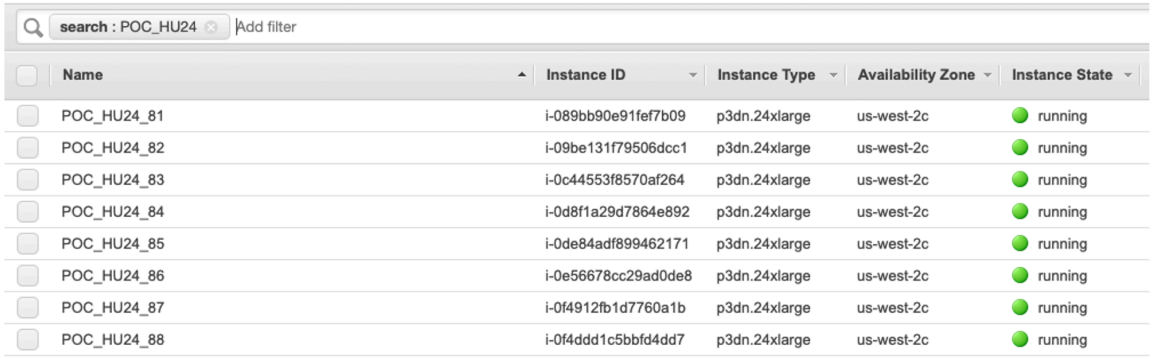
Data Transformation

RecordIO format

- Use `im2rec.py` the same way Simon did. (https://medium.com/@julsimon/imagenet-part-1-going-on-an-adventure-c0a62976dc72). It takes 1.5 days on the `t2.large` instance. I think he did some typos (ImageNet baseline usually uses 224x224 size image, but he uses 480x480).

TFRecord format

- Create an EC2 instance for Training (Deep Learning AMI (Ubuntu 16.04) or Deep Learning AMI (Amazon Linux)). `p3.16xlarge` or `p3dn.24xlarge` is recommended if you need to do distributed GPU training using Uber's Horovod or Tensorflow's DistributedStrategy). Please also note that the default root volume size is 75GB, but I recommend you to increase 100GB since training logs and model checkpoints are stored in the root volume if you do not modify training configuration. If you not want to increase the volume size, then you can delete some conda environments such as Theano, Chainer, Caffe, and Caffe2 after logging in to the EC2 instance.
  - https://aws.amazon.com/ko/getting-started/tutorials/get-started-dlami/
- If you want to train on distributed GPUs, then you need to create multiple GPU instances with the same setting. For example, the below figure shows 8 `p3dn.24xlarge` instances.

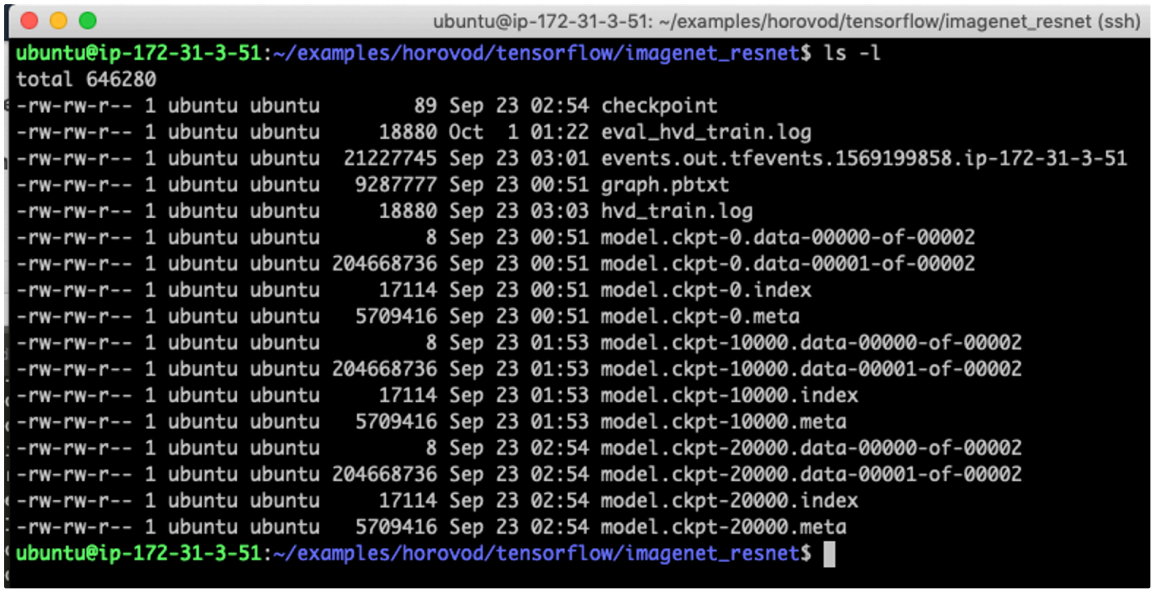


search : POC_HU24 | Add filter

Name	Instance ID	Instance Type	Availability Zone	Instance State
POC_HU24_81	i-089bb90e91fef7b09	p3dn.24xlarge	us-west-2c	running
POC_HU24_82	i-09be131f79506dcc1	p3dn.24xlarge	us-west-2c	running
POC_HU24_83	i-0c44553f8570af264	p3dn.24xlarge	us-west-2c	running
POC_HU24_84	i-0d8f1a29d7864e892	p3dn.24xlarge	us-west-2c	running
POC_HU24_85	i-0de84adf899462171	p3dn.24xlarge	us-west-2c	running
POC_HU24_86	i-0e56678cc29ad0de8	p3dn.24xlarge	us-west-2c	running
POC_HU24_87	i-0f4912fb1d7760a1b	p3dn.24xlarge	us-west-2c	running
POC_HU24_88	i-0f4ddd1c5bbfd4dd7	p3dn.24xlarge	us-west-2c	running

- Please refer to the website for the remaining steps; https://docs.aws.amazon.com/dlami/latest/devguide/tutorial-horovod-tensorflow.html. Note that all code and all feature sets(TFRecord and RecordIO) must be on the same path on each server.

- After training, please check the training log and evaluation log by checking `imagenet_resnet` folder:



- vd_train_log (32 GPUS; 4 p3dn.24xlarge instances)

```
- Step Epoch Speed Loss FinLoss LR - 0 0.0 952.2 6.923 8.262
0.00100 - 1 0.0 2686.6 6.928 8.267 0.00305 - 50 0.3 22243.7
6.586 7.919 0.10353 - .. - 14000 89.5 21021.1 0.750 1.152
0.00012 - 14050 89.8 21818.7 0.583 0.985 0.00002 - Finished in
5289.161954164505
```

- eval_hvd_train.log (32 GPUS; 4 p3dn.24xlarge instances)