This topic describes how to download, preprocess, and upload the ImageNet dataset to use with Cloud TPU. Machine learning models that use the ImageNet dataset include:

- ResNet
- AmoebaNet
- EfficientNet
- MNASNet

ImageNet is an image database. The images in the database are organized into a hierarchy, with each node of the hierarchy depicted by hundreds and thousands of images.

The size of the ImageNet database means it can take a considerable amount of time to train a model. An alternative is to use a demonstration version of the dataset, referred to as *fake_imagenet*. This demonstration version allows you to test the model, while reducing the storage and time requirements typically associated with using the full ImageNet database.

You need about 300GB of space available on your local machine or VM to use the full ImageNet dataset.

If you use **`ctpu up`** to set up your VM, it will allocate 250GB by default.

You can increase the size of the VM disk using one of the following methods:

- Specify the **`--disk-size-gb`** flag on the **`ctpu up`** command line with the size, in GB, that you want allocated.
- Follow the Compute Engine guide to add a disk (/compute/docs/disks/add-persistent-disk) to your VM.

- Set **When deleting instance** to **Delete disk** to ensure that the disk is removed when you remove the VM.
- Make a note of the path to your new disk. For example: `/mnt/disks/mnt-dir`.

For the following commands, a prefix of (vm) means you should run the command on the Compute Engine VM instar
mand does not have the (vm) prefix, run it on your local workstation.

1. Sign up for an ImageNet account (http://image-net.org/signup). Remember the username and password you used to create the account.
2. Set up a `DATA_DIR` environment variable pointing to a path on your Cloud Storage bucket:

3. Download the `imagenet_to_gcs.py` script from GitHub:

4. Set a `SCRATCH_DIR` variable to contain the script's working files. The variable must specify a location on your local machine or on your Compute Engine VM. For example, on your local machine:

   Or if you're processing the data on the VM:

5. Run the `imagenet_to_gcs.py` script to download, format, and upload the ImageNet data to the bucket. Replace `[USERNAME]` and `[PASSWORD]` with the username and password you used to create your ImageNet account.

Optionally if the raw data, in JPEG format, has already been downloaded, you can provide a direct `raw_data_directory` path. If a raw data directory for training or validation data is provided, it should be in the format:

- Training images (https://github.com/awslabs/deeplearning-benchmark/blob/master/tensorflow/inception/inception/data/build_imagenet_data.py) : train/n03062245/n03062245_4620.JPEG
- Validation images (https://github.com/tensorflow/models/blob/master/research/inception/inception/data/preprocess_imagenet_validation_data.py) : validation/ILSVRC2012_val_00000001.JPEG
- Validation labels (http://data.dmlc.ml/mxnet/models/imagenet/synset.txt): synset_labels.txt

The training subdirectory names (for example, n03062245) are "WordNet IDs" (wnid). The ImageNet API (http://www.image-net.org/download-API) shows the mapping of WordNet IDs to their associated validation labels in the `synset_labels.txt` file. A synset in this context is a visually-similar group of images.

Note: Downloading and preprocessing the data can take 10 or more hours, depending on your network and computer speed. Do not interrupt the script.

When the script finishes processing, a message like the following appears:

The script produces a series of directories (for both training and validation) of the form:

and

After the data has been uploaded to your Cloud bucket, run your model and set `--data_dir=${DATA_DIR}`.

Cloud TPU provides a demonstration version of the ImageNet dataset, referred to as *fake_imagenet*. This dataset contains randomly-selected images. You can use this dataset when you want to test how a model works, but don't need the full ImageNet dataset.

The *fake_imagenet* dataset is available in the following Cloud Storage bucket:

gs://cloud-tpu-test-datasets/fake_imagenet

This bucket is read only. When you use the *fake_imagenet* dataset, remember to create a different Cloud Storage buc
our training results and other data.