Before you can train a model, you must prepare the training data for TPU use.

This topic describes how to prepare the COCO (http://cocodataset.org) dataset for models on Cloud TPU.

COCO is a large-scale object detection, segmentation, and captioning dataset. In this step, you convert this dataset into a set of TFRecords (`*.tfrecord`) that the training application can use.

To prepare the COCO dataset, start a VM and run the COCO setup script. You do not need the Cloud TPU set up until after you have prepared the dataset. Since Cloud TPU charges begin when the TPU is set up, best practice is to set up the Compute Engine VM, prepare the dataset, and then set up the Cloud TPU.

Use the TPU setup procedure (/tpu/docs/creating-deleting-tpus#setup_TPU_only) to set up the Cloud TPU after the dataset is prepared.

Machine learning models that use the COCO dataset include:

- Mask-RCNN
- Retinanet

The COCO dataset will be stored on your Cloud Storage. If you have not previously set the storage bucket variable, do that now:

Run the `download_and_preprocess_coco.sh` script to convert the COCO dataset into a set of TFRecords (`*.tfrecord`) that the training application expects.

This installs the required libraries and then runs the preprocessing script. It outputs a number of `*.tfrecord` files in your local data directory. The COCO download and conversion script takes approximately 1 hour to complete.

After you convert the data into TFRecords, copy them from local storage to your Cloud Storage bucket using the `gsutil` command. You must also copy the annotation files. These files help validate the model's performance.