Alpha

This feature is in a pre-release state and might change or have limited support. For more information, see the product launch stages (/products/#product-launch-stages).

For information about access to this release, see the access request page (http://www.google.com).

This tutorial specifically focuses on the FairSeq version of Transformer, and the WMT 18 translation task, translating English to German.

Warning: This model uses a third-party dataset. Google provides no representation, warranty, or other guarantees about the validity, or any other aspects of this dataset.

- Prepare the dataset.
- Run the training job.
- Verify the output results.

This tutorial uses billable components of Google Cloud, including:

- Compute Engine
- Cloud TPU

Use the pricing calculator (/products/calculator/) to generate a cost estimate based on your projected usage. New Google Cloud users might be eligible for a free trial (/free/).

Before starting this tutorial, check that your Google Cloud project is correctly set up.

1. Sign in (https://accounts.google.com/Login) to your Google Account.

   If you don't already have one, sign up for a new account (https://accounts.google.com/SignUp).

2. In the Cloud Console, on the project selector page, select or create a Cloud project.

★ **Note**: If you don't plan to keep the resources that you create in this procedure, create a project instead of selecting an existing project. After you finish these steps, you can delete the project, removing all resources associated with the project.

Go to the project selector page (https://console.cloud.google.com/projectselector2/home/dashboard)

3. Make sure that billing is enabled for your Google Cloud project. Learn how to confirm billing is enabled for your project (/billing/docs/how-to/modify-project).

   This walkthrough uses billable components of Google Cloud. Check the Cloud TPU pricing page (/tpu/docs/pricing) to estimate your costs. Be sure to clean up (#clean_up) resources you create when you've finished with them to avoid unnecessary charges.

1. Open a Cloud Shell window.

   Open Cloud Shell (https://console.cloud.google.com/?cloudshell=true)

2. Create a variable for your project's name.

3. Configure `gcloud` command-line tool to use the project where you want to create Cloud TPU.

4. Launch the Compute Engine resource required for this tutorial.

5. Connect to the new Compute Engine instance.

From this point on, a prefix of `(vm)$` means you should run the command on the Compute Engine VM instance.

1. From the Compute Engine virtual machine, launch a Cloud TPU resource using the following command:

   For more information about IP address ranges and Cloud TPU, see Set up TPU internal IP addresses (/tpu/docs/internal-ip-blocks).

2. Identify the IP address for the Cloud TPU resource.

The IP address is located under the `NETWORK_ENDPOINTS` column. You will need this IP address when you create and configure the PyTorch environment.

1. Create a directory, `pytorch-tutorial-data` to store the model data.

2. Navigate to the `pytorch-tutorial-data` directory.

3. Download the model data.

4. Extract the data.

1. Start a `conda` environment.

2. Configure environmental variables for the Cloud TPU resource.

★ **Note:** The **`TPU_IP_ADDRESS`** variable must equal the IP address of the Cloud TPU you identified in when you launched the Cloud TPU resource (#launch-tpu).

To train the model, run the following script:

Note: Due to the CPU time it takes to prepare the dataset and the initial TPU graph compilations, it can take approximately 20 minutes to train 100 steps and generate the first log output.

Note: Changing the value for the `input_shapes` hyperparameter may lead to improved performance. for example:

These changes can cause significantly slower initial compiles but faster epoch times after stabilization occurs.

To use these input shapes, you must enable bfloat16 use. To do so, run the following command:

After the training job completes, you can find your model checkpoints in the following directory:

Perform a cleanup to avoid incurring unnecessary charges to your account after using the resources you created:

1. Disconnect from the Compute Engine instance, if you have not already done so:

Your prompt should now be `user@projectname`, showing you are in the Cloud Shell.

2. In your Cloud Shell, use the `gcloud` command-line tool to delete the Compute Engine instance.

3. Use `gcloud` command-line tool to delete the Cloud TPU resource.

Try the PyTorch colabs:

- Training MNIST on TPUs (https://colab.sandbox.google.com/github/pytorch/xla/blob/master/contrib/colab/mnist-training-xrt-1-15.ipynb)
- Training ResNet18 on TPUs with Cifar10 dataset (https://colab.sandbox.google.com/github/pytorch/xla/blob/master/contrib/colab/resnet18-training-xrt-1-15.ipynb)
- Inference with Pretrained ResNet50 Model (https://colab.sandbox.google.com/github/pytorch/xla/blob/master/contrib/colab/resnet50-inference-xrt-1-15.ipynb)