Deployment

So far you have been using data that was present as a part of the training framework. In reality, data is usually present externally and stored in a database like an S3 bucket. On this page, we will see how we can use a dataset stored in S3 for training our model using Hyperparameter Optimization and then deploy that model to a Sagemaker endpoint and query it.

Using External Datasets and Deploying Models

Step 1: Uploading Data to S3 There are many ways to upload data to an S3 bucket. For instance, you can use your Sagemaker session, the `aws` CLI or the `boto3` client. Here is how you can upload data to using your Sagemaker session:

```
inputs = sagemaker_session.upload_data(path="data", bucket=bucket, key_prefix=
print("input spec (in this case, just an S3 path): {}".format(inputs))
```

Step 2: Passing data path to your training script When submitting your job for training, you can pass a dictionary of the training data and its path.

```
tuner.fit({"training": inputs})
```

The data present in that path will be included in your training instance. Your training script can get that data by taking an argument as follows:

```
parser.add_argument("--data-dir", type=str, default=os.environ["SM_CHANNEL_TRA
```

Step 3: Deploying your model By default, Sagemaker will deploy the model with the best metric to the instance that you specify:

```
predictor = tuner.deploy(initial_instance_count=1, instance_type="ml.t2.mediu
```

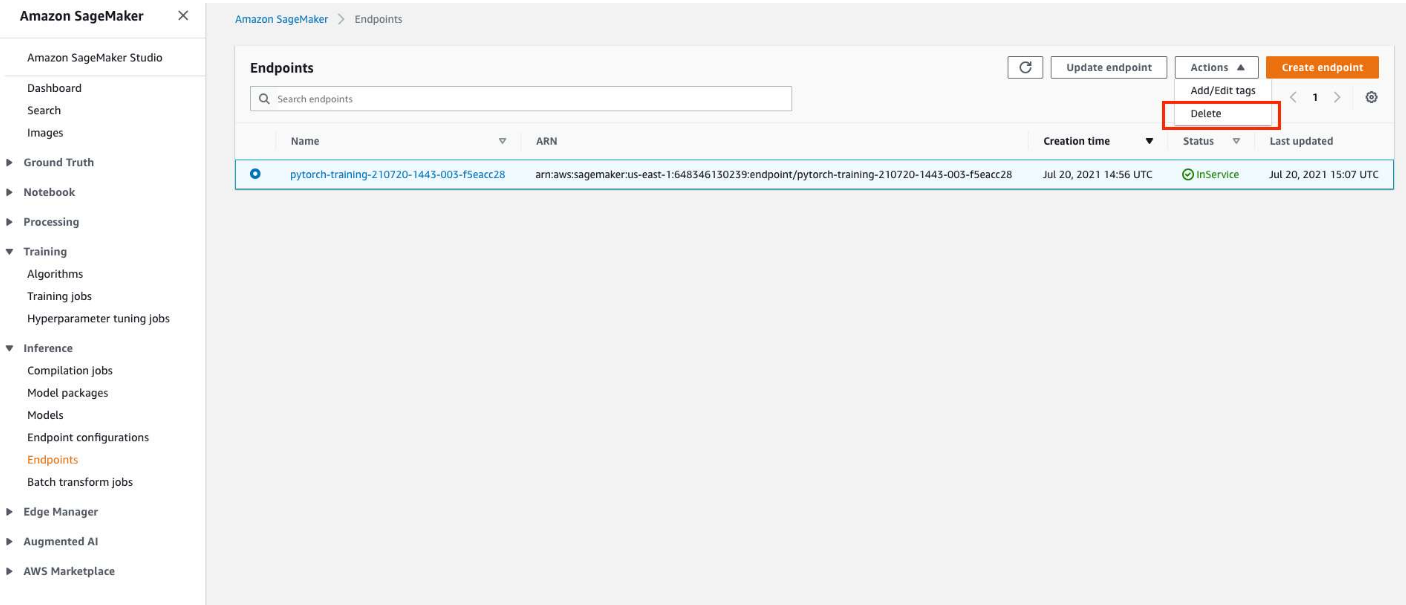
Step 4: Querying your model You can use the `predict()` method of your `predictor` object to query the endpoint with some data

```
response = predictor.predict()
```

Step 5: Deleting your Endpoint Make sure you delete your endpoint before you move on to the next page or close your Sagemaker session:

```
tuner.delete_endpoint()
```

You can also use the UI to delete your endpoint by navigating to Inference, then Endpoints, choose your endpoint, and click Actions followed by Delete:



	Name	ARN	Creation time	Status	Last updated
●	pytorch-training-210720-1443-003-f5eacc28	arn:aws:sagemaker:us-east-1:648346130239:endpoint/pytorch-training-210720-1443-003-f5eacc28	Jul 20, 2021 14:56 UTC	✓ InService	Jul 20, 2021 15:07 UTC

Make sure that you delete your active endpoints by selecting your endpoint and clicking delete

Additional Resources

The following learning resources will help you better understand SageMaker's script mode.

- Before deploying a model, make sure you check the size and pricing of the different **Real-Time Inference** instances here

- You can also perform inference on a lot of data at once using **Batch Transform**. You can read more about it here
- You can also deploy multiple models to the same instance. These is known as **Multi-Model Endpoints**. You can read more about them here
  - **Note** Endpoints will cost money if they are not deleted.
- You can find the API reference for the different inference classes here