10c

November 1, 2024

[1]:
```
!pip install tensorflow tf2onnx onnxruntime numpy
```

```
Requirement already satisfied: tensorflow in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (2.13.0)
Requirement already satisfied: tf2onnx in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (1.16.1)
Collecting onnxruntime
  Downloading onnxruntime-1.19.2-cp38-cp38-win_amd64.whl (11.1 MB)
Requirement already satisfied: numpy in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (1.22.3)
Requirement already satisfied: tensorflow-intel==2.13.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow)
(2.13.0)
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (0.31.0)
Requirement already satisfied: libclang>=13.0.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (16.0.6)
Requirement already satisfied: google-pasta>=0.1.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (0.2.0)
Requirement already satisfied: six>=1.12.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (1.16.0)
Requirement already satisfied: tensorflow-estimator<2.14,>=2.13.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (2.13.0)
Requirement already satisfied: opt-einsum>=2.3.2 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (3.3.0)
Collecting typing-extensions<4.6.0,>=3.6.6
  Using cached typing_extensions-4.5.0-py3-none-any.whl (27 kB)
Requirement already satisfied: gast<=0.4.0,>=0.2.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (0.4.0)
Requirement already satisfied: flatbuffers>=23.1.21 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
```

```
intel==2.13.0->tensorflow) (23.5.26)
Requirement already satisfied: absl-py>=1.0.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (1.0.0)
Requirement already satisfied: h5py>=2.9.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (3.7.0)
Requirement already satisfied: keras<2.14,>=2.13.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (2.13.1)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (1.59.2)
Requirement already satisfied: termcolor>=1.1.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (2.3.0)
Requirement already satisfied: packaging in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (23.1)
Requirement already satisfied: wrapt>=1.11.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (1.14.1)
Requirement already satisfied: setuptools in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (61.2.0)
Requirement already satisfied: astunparse>=1.6.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (1.6.3)
Requirement already satisfied: tensorboard<2.14,>=2.13 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (2.13.0)
Requirement already satisfied:
protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,<5.0.0dev,>=3.20.3
in c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tensorflow-
intel==2.13.0->tensorflow) (3.20.3)
Requirement already satisfied: onnx>=1.4.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tf2onnx) (1.17.0)
Requirement already satisfied: requests in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from tf2onnx) (2.31.0)
Collecting coloredlogs
  Downloading coloredlogs-15.0.1-py2.py3-none-any.whl (46 kB)
Requirement already satisfied: sympy in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from onnxruntime)
(1.12)
Requirement already satisfied: wheel<1.0,>=0.23.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
astunparse>=1.6.0->tensorflow-intel==2.13.0->tensorflow) (0.37.1)
Requirement already satisfied: google-auth-oauthlib<1.1,>=0.5 in
```

```
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow) (1.0.0)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow) (0.7.2)
Requirement already satisfied: werkzeug>=1.0.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow) (3.0.4)
Requirement already satisfied: google-auth<3,>=1.6.3 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow) (2.23.4)
Requirement already satisfied: markdown>=2.6.8 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow) (3.3.7)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(0.2.8)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(5.1.0)
Requirement already satisfied: rsa<5,>=3.1.4 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(4.8)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from google-auth-
oauthlib<1.1,>=0.5->tensorboard<2.14,>=2.13->tensorflow-
intel==2.13.0->tensorflow) (1.3.1)
Requirement already satisfied: importlib-metadata>=4.4 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
markdown>=2.6.8->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(6.0.0)
Requirement already satisfied: zipp>=0.5 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from importlib-
metadata>=4.4->markdown>=2.6.8->tensorboard<2.14,>=2.13->tensorflow-
intel==2.13.0->tensorflow) (3.8.0)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
pyasn1-modules>=0.2.1->google-
auth<3,>=1.6.3->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(0.4.8)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from requests->tf2onnx)
(3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from requests->tf2onnx)
```

```
(1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from requests->tf2onnx)
(2024.8.30)
Requirement already satisfied: charset-normalizer<4,>=2 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from requests->tf2onnx)
(2.0.12)
Requirement already satisfied: oauthlib>=3.0.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from requests-
oauthlib>=0.7.0->google-auth-
oauthlib<1.1,>=0.5->tensorboard<2.14,>=2.13->tensorflow-
intel==2.13.0->tensorflow) (3.2.0)
Requirement already satisfied: MarkupSafe>=2.1.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
werkzeug>=1.0.1->tensorboard<2.14,>=2.13->tensorflow-intel==2.13.0->tensorflow)
(2.1.5)
Collecting humanfriendly>=9.1
  Downloading humanfriendly-10.0-py2.py3-none-any.whl (86 kB)
Collecting pyreadline3
  Downloading pyreadline3-3.5.4-py3-none-any.whl (83 kB)
Requirement already satisfied: mpmath>=0.19 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
sympy->onnxruntime) (1.3.0)
Installing collected packages: pyreadline3, typing-extensions, humanfriendly,
coloredlogs, onnxruntime
  Attempting uninstall: typing-extensions
    Found existing installation: typing-extensions 4.12.2
    Uninstalling typing-extensions-4.12.2:
      Successfully uninstalled typing-extensions-4.12.2
Successfully installed coloredlogs-15.0.1 humanfriendly-10.0 onnxruntime-1.19.2
pyreadline3-3.5.4 typing-extensions-4.5.0

WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
    WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
    WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
```

```
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
ERROR: pip's dependency resolver does not currently take into account all the
packages that are installed. This behaviour is the source of the following
dependency conflicts.
torchvision 0.12.0 requires torch==1.11.0, but you have torch 2.0.1+cu117 which
is incompatible.
emoji 2.13.0 requires typing-extensions>=4.7.0, but you have typing-extensions
4.5.0 which is incompatible.
altair 5.4.1 requires typing-extensions>=4.10.0; python_version < "3.13", but
you have typing-extensions 4.5.0 which is incompatible.
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
```

[7]:

```
import tensorflow as tf
import tf2onnx
import numpy as np
import onnx
import onnxruntime as ort
```

```
import time
import matplotlib.pyplot as plt

# Step 1: Load the MNIST dataset
mnist = tf.keras.datasets.mnist
(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0  # Normalize the images

# Reshape data to add a channel dimension
x_train = x_train[..., np.newaxis].astype("float32")
x_test = x_test[..., np.newaxis].astype("float32")

# Step 2: Define a simple CNN model
model = tf.keras.models.Sequential([
    tf.keras.layers.Conv2D(32, kernel_size=(3, 3), activation='relu',␣
 ↪input_shape=(28, 28, 1)),
    tf.keras.layers.MaxPooling2D(pool_size=(2, 2)),
    tf.keras.layers.Conv2D(64, kernel_size=(3, 3), activation='relu'),
    tf.keras.layers.MaxPooling2D(pool_size=(2, 2)),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(10, activation='softmax')
])

# Compile and train the model
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',␣
 ↪metrics=['accuracy'])
model.fit(x_train, y_train, epochs=1, validation_data=(x_test, y_test))  #␣
 ↪Train for 1 epoch for demo purposes

# Step 3: Convert the model to ONNX format
onnx_model_path = "mnist_cnn_model.onnx"
spec = (tf.TensorSpec((None, 28, 28, 1), tf.float32, name="input"),)
model_proto, _ = tf2onnx.convert.from_keras(model, input_signature=spec,␣
 ↪output_path=onnx_model_path)
print(f"Model exported to {onnx_model_path}")

# Step 4: Load and validate the ONNX model
onnx_model = onnx.load(onnx_model_path)
onnx.checker.check_model(onnx_model)
print("ONNX model is valid")

# Step 5: Set up ONNX Runtime session
ort_session = ort.InferenceSession(onnx_model_path)

tf.config.run_functions_eagerly(True)
```

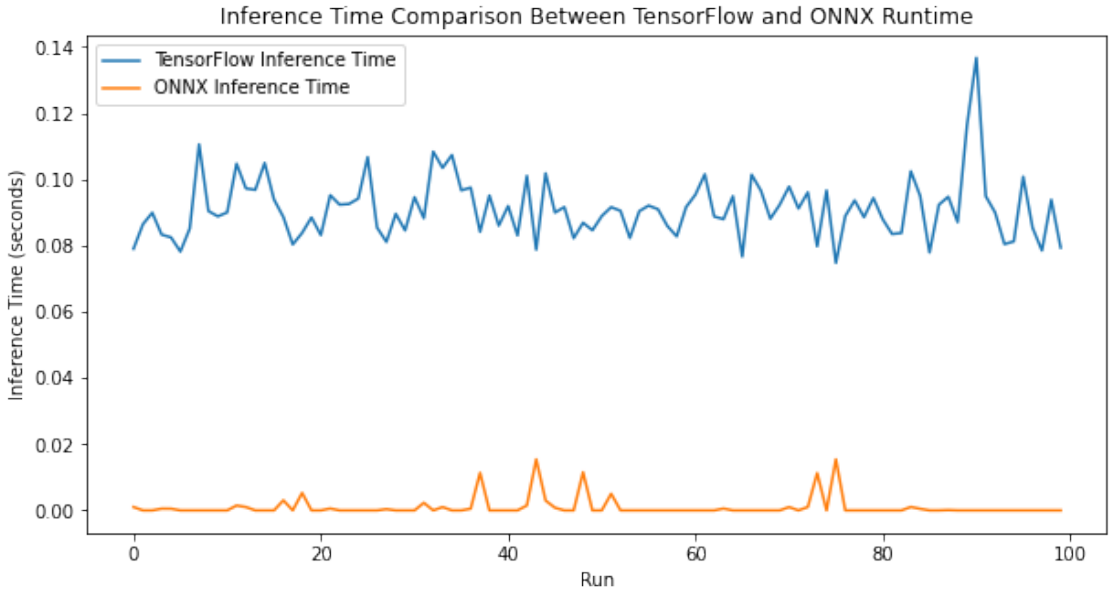
```
1875/1875 [==============================] - 221s 118ms/step - loss: 0.1373 -
accuracy: 0.9573 - val_loss: 0.0360 - val_accuracy: 0.9881
Model exported to mnist_cnn_model.onnx
ONNX model is valid
```

[10]:
```python
# Step 8: Measure and plot inference time for TensorFlow and ONNX models
def measure_inference_time(tf_model, ort_session, x_test, num_runs=100):
    tf_times = []
    onnx_times = []

    for _ in range(num_runs):
        # Select a single random sample image and ensure correct shape for the␣
 ↪model
        test_image = np.expand_dims(x_test[np.random.randint(len(x_test))],␣
 ↪axis=0)

        # TensorFlow inference time
        start_time = time.time()
        tf_output = tf_model.predict(test_image)
        tf_times.append(time.time() - start_time)

        # ONNX inference time
        start_time = time.time()
        onnx_output = ort_session.run(None, {'input': test_image.astype(np.
 ↪float32)})
        onnx_times.append(time.time() - start_time)

    return tf_times, onnx_times

# Run inference timing
tf_times, onnx_times = measure_inference_time(tf_model, ort_session, x_test)

# Step 9: Plot the inference time comparison
plt.figure(figsize=(10, 5))
plt.plot(tf_times, label="TensorFlow Inference Time")
plt.plot(onnx_times, label="ONNX Inference Time")
plt.xlabel("Run")
plt.ylabel("Inference Time (seconds)")
plt.legend()
plt.title("Inference Time Comparison Between TensorFlow and ONNX Runtime")
plt.show()
```

```
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 31ms/step
```

```
1/1 [==============================] - 0s 28ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 36ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 37ms/step
1/1 [==============================] - 0s 45ms/step
1/1 [==============================] - 0s 36ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 36ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 16ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 47ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 36ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 18ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 18ms/step
1/1 [==============================] - 0s 21ms/step
1/1 [==============================] - 0s 37ms/step
1/1 [==============================] - 0s 41ms/step
1/1 [==============================] - 0s 48ms/step
1/1 [==============================] - 0s 22ms/step
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 25ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 25ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 33ms/step
```

```
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 28ms/step
1/1 [==============================] - 0s 28ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 28ms/step
1/1 [==============================] - 0s 22ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 37ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 29ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 42ms/step
1/1 [==============================] - 0s 47ms/step
1/1 [==============================] - 0s 32ms/step
1/1 [==============================] - 0s 33ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 35ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 34ms/step
1/1 [==============================] - 0s 22ms/step
1/1 [==============================] - 0s 31ms/step
```



[11]:
```
pip install onnx2pytorch
```

```
Collecting onnx2pytorch
  Downloading onnx2pytorch-0.5.0-py3-none-any.whl (45 kB)
Requirement already satisfied: torchvision>=0.9.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from onnx2pytorch)
(0.12.0)
Requirement already satisfied: torch>=1.4.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from onnx2pytorch)
(2.0.1+cu117)
Requirement already satisfied: onnx>=1.6.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from onnx2pytorch)
(1.17.0)
Requirement already satisfied: protobuf>=3.20.2 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
onnx>=1.6.0->onnx2pytorch) (3.20.3)
Requirement already satisfied: numpy>=1.20 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
onnx>=1.6.0->onnx2pytorch) (1.22.3)
Requirement already satisfied: filelock in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torch>=1.4.0->onnx2pytorch) (3.7.0)
Requirement already satisfied: jinja2 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torch>=1.4.0->onnx2pytorch) (3.1.4)
Requirement already satisfied: sympy in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
```

```
torch>=1.4.0->onnx2pytorch) (1.12)
Requirement already satisfied: networkx in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torch>=1.4.0->onnx2pytorch) (3.0)
Requirement already satisfied: typing-extensions in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torch>=1.4.0->onnx2pytorch) (4.5.0)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torchvision>=0.9.0->onnx2pytorch) (10.2.0)
Collecting torch>=1.4.0
  Using cached torch-1.11.0-cp38-cp38-win_amd64.whl (158.0 MB)
Requirement already satisfied: requests in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
torchvision>=0.9.0->onnx2pytorch) (2.31.0)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
jinja2->torch>=1.4.0->onnx2pytorch) (2.1.5)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
requests->torchvision>=0.9.0->onnx2pytorch) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
requests->torchvision>=0.9.0->onnx2pytorch) (2024.8.30)
Requirement already satisfied: urllib3<3,>=1.21.1 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
requests->torchvision>=0.9.0->onnx2pytorch) (1.26.18)
Requirement already satisfied: charset-normalizer<4,>=2 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
requests->torchvision>=0.9.0->onnx2pytorch) (2.0.12)
Requirement already satisfied: mpmath>=0.19 in
c:\users\karthi\anaconda3\envs\object\lib\site-packages (from
sympy->torch>=1.4.0->onnx2pytorch) (1.3.0)
Installing collected packages: torch, onnx2pytorch
  Attempting uninstall: torch
    Found existing installation: torch 2.0.1+cu117
    Uninstalling torch-2.0.1+cu117:
      Successfully uninstalled torch-2.0.1+cu117
Successfully installed onnx2pytorch-0.5.0 torch-1.11.0
Note: you may need to restart the kernel to use updated packages.

WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
```

```
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
    WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
    WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -rotobuf
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
WARNING: Ignoring invalid distribution -illow
(c:\users\karthi\anaconda3\envs\object\lib\site-packages)
```

[12]:

```python
# Import necessary libraries
import onnx
import torch
import numpy as np
from onnx2pytorch import ConvertModel

# Step 1: Load the ONNX model
onnx_model_path = "mnist_cnn_model.onnx"  # Update this path if your model has␣
 ↪a different name
onnx_model = onnx.load(onnx_model_path)

# Step 2: Convert the ONNX model to a PyTorch model
pytorch_model = ConvertModel(onnx_model)

# Step 3: Prepare sample input (similar to the original input shape)
# For the MNIST dataset, the input should be in the shape (batch_size, 1, 28,␣
 ↪28)
```

```
sample_input = np.random.rand(1, 1, 28, 28).astype(np.float32)  # Replace with␣
 ↪actual test data as needed
sample_input_tensor = torch.tensor(sample_input)

# Step 4: Run inference on the PyTorch model
with torch.no_grad():  # No need to calculate gradients during inference
    output = pytorch_model(sample_input_tensor)

# Step 5: Process and display the output
print("PyTorch model output:", output)
```

```
PyTorch model output: tensor([[0.0699, 0.0730, 0.1236, 0.1025, 0.0771, 0.1461,
0.0522, 0.0757, 0.2124,
         0.0674]])
```

```
C:\Users\karthi\anaconda3\envs\object\lib\site-
packages\onnx2pytorch\convert\layer.py:30: UserWarning: The given NumPy array is
not writable, and PyTorch does not support non-writable tensors. This means
writing to this tensor will result in undefined behavior. You may want to copy
the array to protect its data or make it writable before converting it to a
tensor. This type of warning will be suppressed for the rest of this program.
(Triggered internally at  C:\actions-runner\_work\pytorch\pytorch\builder\window
s\pytorch\torch\csrc\utils\tensor_numpy.cpp:178.)
  layer.weight.data = torch.from_numpy(numpy_helper.to_array(weight))
C:\Users\karthi\anaconda3\envs\object\lib\site-
packages\torch\nn\functional.py:749: UserWarning: Note that order of the
arguments: ceil_mode and return_indices will changeto match the args list in
nn.MaxPool2d in a future release.
  warnings.warn("Note that order of the arguments: ceil_mode and return_indices
will change"
```

```
[ ]:
```

10cb

November 1, 2024

[6]:
```
import tensorflow as tf
import numpy as np

# Create a simple CNN model
def create_model():
    model = tf.keras.Sequential([
        tf.keras.layers.Conv2D(16, (3, 3), activation='relu', input_shape=(28,␣
 ↪28, 1)),
        tf.keras.layers.MaxPooling2D((2, 2)),
        tf.keras.layers.Conv2D(32, (3, 3), activation='relu'),
        tf.keras.layers.MaxPooling2D((2, 2)),
        tf.keras.layers.Flatten(),
        tf.keras.layers.Dense(10, activation='softmax')
    ])
    return model

# Load and preprocess the data
(x_train, y_train), (x_test, y_test) = tf.keras.datasets.mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0  # Normalize the data
x_train = x_train[..., np.newaxis]  # Add channel dimension
x_test = x_test[..., np.newaxis]

# Compile and train the model briefly
model = create_model()
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',␣
 ↪metrics=['accuracy'])
model.fit(x_train, y_train, epochs=2, validation_data=(x_test, y_test))

# Convert the trained model to TensorFlow Lite format with quantization
converter = tf.lite.TFLiteConverter.from_keras_model(model)
converter.optimizations = [tf.lite.Optimize.DEFAULT]  # Enable default␣
 ↪quantization

# Convert the model
quantized_model = converter.convert()

# Save the quantized model
```

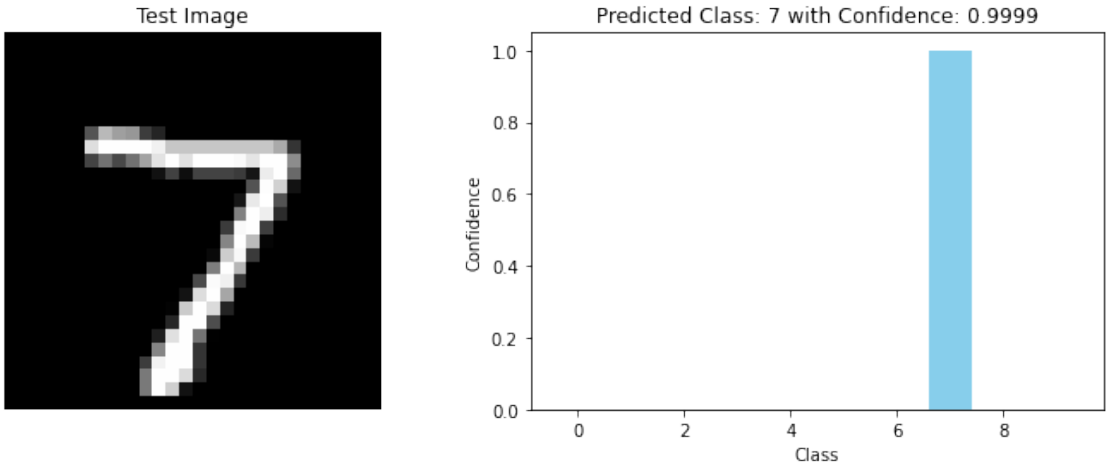
```
with open("quantized_model.tflite", "wb") as f:
    f.write(quantized_model)

print("Quantized model has been saved as 'quantized_model.tflite'")

# Optional: Load and test the quantized model for inference
interpreter = tf.lite.Interpreter(model_path="quantized_model.tflite")
interpreter.allocate_tensors()
```

```
Epoch 1/2
1875/1875 [==============================] - 31s 16ms/step - loss: 0.1954 -
accuracy: 0.9431 - val_loss: 0.0661 - val_accuracy: 0.9790
Epoch 2/2
1875/1875 [==============================] - 30s 16ms/step - loss: 0.0681 -
accuracy: 0.9785 - val_loss: 0.0631 - val_accuracy: 0.9787
INFO:tensorflow:Assets written to:
C:\Users\karthi\AppData\Local\Temp\tmp_yrzl_gd\assets

INFO:tensorflow:Assets written to:
C:\Users\karthi\AppData\Local\Temp\tmp_yrzl_gd\assets

Quantized model has been saved as 'quantized_model.tflite'
```

[7]:
```
import numpy as np
import matplotlib.pyplot as plt

# Ensure the input sample is in FLOAT32 format
sample_input = x_test[0:1].astype(np.float32)  # Take one sample and convert to␣
 ↪FLOAT32
interpreter.set_tensor(input_details[0]['index'], sample_input)
interpreter.invoke()
output_data = interpreter.get_tensor(output_details[0]['index'])

# Find the maximum value in the model output and its index (predicted class)
max_value = np.max(output_data)
predicted_class = np.argmax(output_data)

print("Quantized model output:", output_data)
print("Predicted class:", predicted_class)
print("Maximum value (confidence):", max_value)

# Plot the test image and the output data
plt.figure(figsize=(10, 4))

# Plot the test image
plt.subplot(1, 2, 1)
plt.imshow(x_test[0].squeeze(), cmap='gray')
plt.title("Test Image")
```

```
plt.axis('off')

# Plot the output confidence values
plt.subplot(1, 2, 2)
plt.bar(range(10), output_data[0], color='skyblue')
plt.xlabel('Class')
plt.ylabel('Confidence')
plt.title(f'Predicted Class: {predicted_class} with Confidence: {max_value:.
  ↪4f}')

plt.tight_layout()
plt.show()
```

```
Quantized model output: [[6.4943869e-08 4.0027018e-08 3.4217155e-05
3.7877217e-05 3.7895242e-11
  1.5684165e-08 3.8829763e-13 9.9989438e-01 1.3171524e-07 3.3238786e-05]]
Predicted class: 7
Maximum value (confidence): 0.9998944
```



[ ]:

10cc

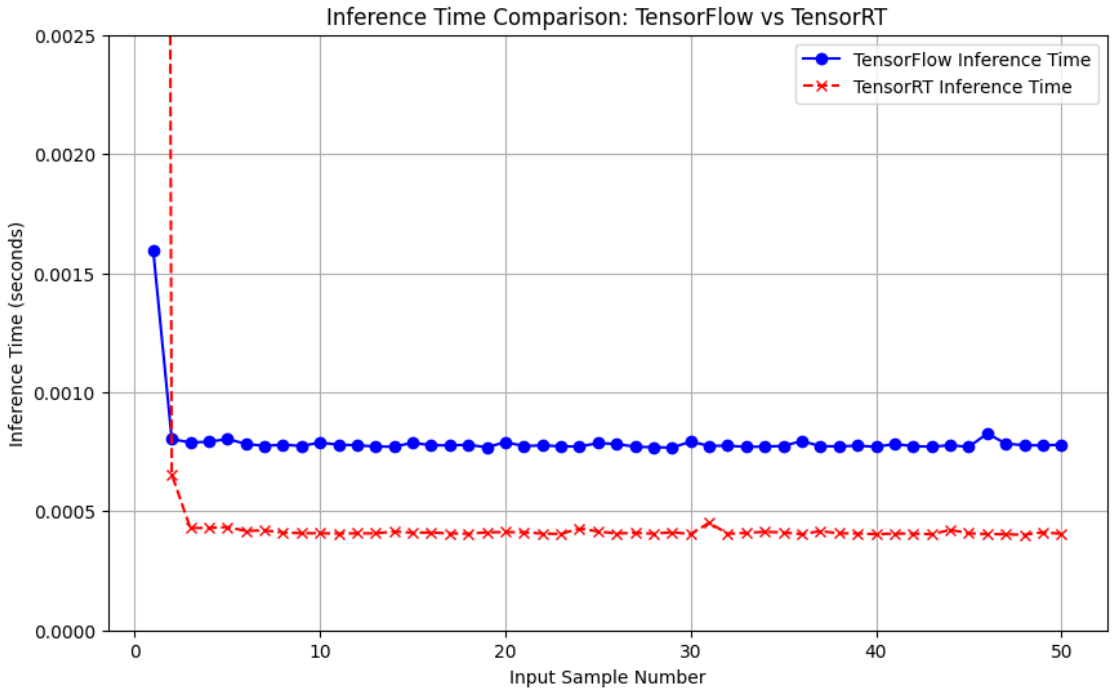
November 1, 2024

[4]:
```
import tensorflow as tf
import numpy as np
import time
import matplotlib.pyplot as plt
from tensorflow.keras.datasets import mnist
from tensorflow.python.compiler.tensorrt import trt_convert as trt

def create_model():
    model = tf.keras.Sequential([
        tf.keras.layers.Flatten(input_shape=(28, 28)),
        tf.keras.layers.Dense(128, activation='relu'),
        tf.keras.layers.Dense(10, activation='softmax')
    ])

    model.compile(optimizer='adam',
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model

def measure_inference_times(model, input_data, num_samples=50):
    inference_times = []

    # Measure inference time for each sample
    for i in range(num_samples):
        start_time = time.time()
        # Use the model directly for inference
        model(input_data[i % len(input_data)][np.newaxis, ...])  # Call the␣
 ↪model directly
        end_time = time.time()

        # Calculate inference time for the current sample
        inference_times.append(end_time - start_time)

    return inference_times

def convert_to_tensorrt(model):
    # Save the TensorFlow model
```

```
    tf.saved_model.save(model, 'saved_model')

    # Use TensorRT to optimize the model
    converter = trt.TrtGraphConverterV2(input_saved_model_dir='saved_model')
    converter.convert()
    converter.save('tensorrt_model')

    # Load the TensorRT model
    return tf.saved_model.load('tensorrt_model')

# Load the MNIST dataset
(x_train, y_train), (x_test, y_test) = mnist.load_data()

# Normalize the data to [0, 1]
x_test = x_test.astype(np.float32) / 255.0

# Ensure test_data has at least 50 samples
if x_test.shape[0] < 50:
    raise ValueError("Test data must contain at least 50 samples.")

# Create and train the model
model = create_model()
model.fit(x_train, y_train, epochs=5, batch_size=32, verbose=2)

# Measure inference times for the first 50 inputs using TensorFlow
tf_inference_times = measure_inference_times(model, x_test[:50])

# Convert the model to TensorRT
tensorrt_model = convert_to_tensorrt(model)

# Measure inference times for the first 50 inputs using TensorRT
tensorrt_inference_times = measure_inference_times(tensorrt_model, x_test[:50])

# Plot the inference times for comparison
plt.figure(figsize=(10, 6))
plt.plot(range(1, 51), tf_inference_times, marker='o', linestyle='-',␣
 ↪color='b', label='TensorFlow Inference Time')
plt.plot(range(1, 51), tensorrt_inference_times, marker='x', linestyle='--',␣
 ↪color='r', label='TensorRT Inference Time')
plt.xlabel("Input Sample Number")
plt.ylabel("Inference Time (seconds)")
plt.title("Inference Time Comparison: TensorFlow vs TensorRT")
plt.ylim(0,0.0025)
plt.legend()
plt.grid(True)
plt.show()
```

```
Epoch 1/5
1875/1875 - 2s - loss: 2.8073 - accuracy: 0.8515 - 2s/epoch - 1ms/step
Epoch 2/5
1875/1875 - 2s - loss: 0.3995 - accuracy: 0.9046 - 2s/epoch - 959us/step
Epoch 3/5
1875/1875 - 2s - loss: 0.3048 - accuracy: 0.9232 - 2s/epoch - 956us/step
Epoch 4/5
1875/1875 - 2s - loss: 0.2644 - accuracy: 0.9340 - 2s/epoch - 889us/step
Epoch 5/5
1875/1875 - 2s - loss: 0.2313 - accuracy: 0.9432 - 2s/epoch - 914us/step
INFO:tensorflow:Assets written to: saved_model/assets
INFO:tensorflow:Linked TensorRT version: (8, 4, 2)
INFO:tensorflow:Loaded TensorRT version: (8, 4, 2)
INFO:tensorflow:Clearing prior device assignments in loaded saved model
INFO:tensorflow:Automatic mixed precision has been deactivated.
INFO:tensorflow:Could not find TRTEngineOp_002_000 in TF-TRT cache. This can
happen if build() is not called, which means TensorRT engines will be built and
cached at runtime.

2024-10-29 10:03:05.770875: I tensorflow/core/grappler/devices.cc:66] Number of
eligible GPUs (core count >= 8, compute capability >= 0.0): 1
2024-10-29 10:03:05.771169: I
tensorflow/core/grappler/clusters/single_machine.cc:358] Starting new session
2024-10-29 10:03:05.777741: I
tensorflow/core/common_runtime/gpu/gpu_device.cc:1532] Created device
/job:localhost/replica:0/task:0/device:GPU:0 with 17007 MB memory:  -> device:
0, name: NVIDIA A100-SXM4-40GB MIG 3g.20gb, pci bus id: 0000:90:00.0, compute
capability: 8.0
2024-10-29 10:03:05.855855: I tensorflow/core/grappler/devices.cc:66] Number of
eligible GPUs (core count >= 8, compute capability >= 0.0): 1
2024-10-29 10:03:05.856116: I
tensorflow/core/grappler/clusters/single_machine.cc:358] Starting new session
2024-10-29 10:03:05.862402: I
tensorflow/core/common_runtime/gpu/gpu_device.cc:1532] Created device
/job:localhost/replica:0/task:0/device:GPU:0 with 17007 MB memory:  -> device:
0, name: NVIDIA A100-SXM4-40GB MIG 3g.20gb, pci bus id: 0000:90:00.0, compute
capability: 8.0
2024-10-29 10:03:05.884669: W
tensorflow/compiler/tf2tensorrt/convert/trt_optimization_pass.cc:198]
Calibration with FP32 or FP16 is not implemented. Falling back to
use_calibration = False.Note that the default value of use_calibration is True.
2024-10-29 10:03:05.885713: W
tensorflow/compiler/tf2tensorrt/segment/segment.cc:952]

################################################################################
TensorRT unsupported/non-converted OP Report:
        - NoOp -> 2x
        - Identity -> 1x
```

```
        - Placeholder -> 1x
--------------------------------------------------------------------------------
        - Total nonconverted OPs: 4
        - Total nonconverted OP Types: 3
For more information see https://docs.nvidia.com/deeplearning/frameworks/tf-trt-
user-guide/index.html#supported-ops.
################################################################################

2024-10-29 10:03:05.885776: W
tensorflow/compiler/tf2tensorrt/segment/segment.cc:1280] The environment
variable TF_TRT_MAX_ALLOWED_ENGINES=20 has no effect since there are only 1 TRT
Engines with  at least minimum_segment_size=3 nodes.
2024-10-29 10:03:05.885787: I
tensorflow/compiler/tf2tensorrt/convert/convert_graph.cc:799] Number of TensorRT
candidate segments: 1
2024-10-29 10:03:05.886368: I
tensorflow/compiler/tf2tensorrt/convert/convert_graph.cc:916] Replaced segment 0
consisting of 12 nodes by TRTEngineOp_002_000.

INFO:tensorflow:Assets written to: tensorrt_model/assets
```



[ ]: