Module03

June 18, 2024

[1]:
```
import tensorflow as tf
from tensorflow.keras.datasets import mnist

tf.compat.v1.enable_v2_behavior()

(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0
x_train = x_train.reshape((-1, 28, 28, 1))
x_test = x_test.reshape((-1, 28, 28, 1))
```

```
2024-06-18 23:05:46.904554: I tensorflow/core/util/port.cc:113] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-06-18 23:05:47.017146: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 AVX_VNNI FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
2024-06-18 23:05:49.112813: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
find TensorRT
```

[2]:
```
model = tf.keras.Sequential(
    [    tf.keras.layers.Conv2D(32, (3, 3),
        activation='relu', input_shape=(28, 28, 1)),
        tf.keras.layers.MaxPooling2D((2, 2)),
        tf.keras.layers.Flatten(),
        tf.keras.layers.Dense(128, activation='relu'),
        tf.keras.layers.Dense(10, activation='softmax')
    ])
```

```
/usr/local/lib/python3.10/dist-
packages/keras/src/layers/convolutional/base_conv.py:107: UserWarning: Do not
pass an `input_shape`/`input_dim` argument to a layer. When using Sequential
models, prefer using an `Input(shape)` object as the first layer in the model
instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
```

```
2024-06-18 23:05:51.623364: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.716479: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.717001: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.721798: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.722348: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.722620: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.827317: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.827797: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
```

```
2024-06-18 23:05:51.828160: I
external/local_xla/xla/stream_executor/cuda/cuda_executor.cc:998] successful
NUMA node read from SysFS had negative value (-1), but there must be at least
one NUMA node, so returning NUMA node zero. See more at
https://github.com/torvalds/linux/blob/v6.0/Documentation/ABI/testing/sysfs-bus-
pci#L344-L355
2024-06-18 23:05:51.828480: I
tensorflow/core/common_runtime/gpu/gpu_device.cc:1928] Created device
/job:localhost/replica:0/task:0/device:GPU:0 with 1216 MB memory:  -> device: 0,
name: NVIDIA RTX A1000 Laptop GPU, pci bus id: 0000:01:00.0, compute capability:
8.6
```

[3]:
```
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])
```

[4]:
```
model.fit(x_train, y_train, epochs=5, validation_data=(x_test, y_test))
```

```
Epoch 1/5

WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
I0000 00:00:1718751977.949332   27394 service.cc:145] XLA service 0x7206e8009f70
initialized for platform CUDA (this does not guarantee that XLA will be used).
Devices:
I0000 00:00:1718751977.949415   27394 service.cc:153]   StreamExecutor device
(0): NVIDIA RTX A1000 Laptop GPU, Compute Capability 8.6
2024-06-18 23:06:18.010017: I
tensorflow/compiler/mlir/tensorflow/utils/dump_mlir_util.cc:268] disabling MLIR
crash reproducer, set env var `MLIR_CRASH_REPRODUCER_DIRECTORY` to enable.
2024-06-18 23:06:18.259925: I
external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:465] Loaded cuDNN
version 8906

  41/1875            6s 4ms/step - accuracy:
0.5195 - loss: 1.5012

I0000 00:00:1718751981.685848   27394 device_compiler.h:188] Compiled cluster
using XLA!  This line is logged at most once for the lifetime of the process.

1875/1875            15s 5ms/step -
accuracy: 0.9032 - loss: 0.3182 - val_accuracy: 0.9757 - val_loss: 0.0748
Epoch 2/5
1875/1875            7s 4ms/step -
accuracy: 0.9845 - loss: 0.0540 - val_accuracy: 0.9851 - val_loss: 0.0479
Epoch 3/5
1875/1875            7s 4ms/step -
accuracy: 0.9898 - loss: 0.0323 - val_accuracy: 0.9834 - val_loss: 0.0533
Epoch 4/5
1875/1875            7s 4ms/step -
```

```
accuracy: 0.9936 - loss: 0.0204 - val_accuracy: 0.9866 - val_loss: 0.0407
Epoch 5/5
1875/1875             7s 4ms/step -
accuracy: 0.9958 - loss: 0.0137 - val_accuracy: 0.9865 - val_loss: 0.0421
```

[4]:
```
<keras.src.callbacks.history.History at 0x72083eca6650>
```

[12]:
```
# Save model to disk
# This was needed to work with the TFLite conversion process
model.export('mnist_model')
```

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

INFO:tensorflow:Assets written to: mnist_model/assets

Saved artifact at 'mnist_model'. The following endpoints are available:

* Endpoint 'serve'
  args_0 (POSITIONAL_ONLY): TensorSpec(shape=(None, 28, 28, 1),
dtype=tf.float32, name='keras_tensor')
Output Type:
  TensorSpec(shape=(None, 10), dtype=tf.float32, name=None)
Captures:
  125379738248368: TensorSpec(shape=(), dtype=tf.resource, name=None)
  125379738252416: TensorSpec(shape=(), dtype=tf.resource, name=None)
  125379738569712: TensorSpec(shape=(), dtype=tf.resource, name=None)
  125379738568656: TensorSpec(shape=(), dtype=tf.resource, name=None)
  125379738571824: TensorSpec(shape=(), dtype=tf.resource, name=None)
  125379738568304: TensorSpec(shape=(), dtype=tf.resource, name=None)
```

Conversion needed to be updated to account for the saved model for use with TFLite, the model needed to be saved in a different format to be imported

Information for this was found in the documentation https://www.tensorflow.org/tutorials/keras/save_and_load

This post also helped identify the problem and resolution https://stackoverflow.com/questions/62358745/oserror-savedmodel-file-does-not-exist-at-c-users-munib-new-folder-saved-mod

[20]:
```
converter = tf.lite.TFLiteConverter.from_saved_model("mnist_model")
tflite_model = converter.convert()
open("model.tflite", "wb").write(tflite_model)
```

```
W0000 00:00:1718753168.071540   26911 tf_tfl_flatbuffer_helpers.cc:390] Ignored
output_format.
W0000 00:00:1718753168.071736   26911 tf_tfl_flatbuffer_helpers.cc:393] Ignored
drop_control_dependency.
2024-06-18 23:26:08.074486: I tensorflow/cc/saved_model/reader.cc:83] Reading
SavedModel from: mnist_model
2024-06-18 23:26:08.076129: I tensorflow/cc/saved_model/reader.cc:51] Reading
meta graph with tags { serve }
```

```
2024-06-18 23:26:08.076153: I tensorflow/cc/saved_model/reader.cc:146] Reading
SavedModel debug info (if present) from: mnist_model
2024-06-18 23:26:08.095138: I tensorflow/cc/saved_model/loader.cc:234] Restoring
SavedModel bundle.
2024-06-18 23:26:08.184630: I tensorflow/cc/saved_model/loader.cc:218] Running
initialization op on SavedModel bundle at path: mnist_model
2024-06-18 23:26:08.205136: I tensorflow/cc/saved_model/loader.cc:317]
SavedModel load for tags { serve }; Status: success: OK. Took 130872
microseconds.
```

[20]: 2778540

[14]:
```
with open('model.tflite', 'wb') as f:
    f.write(tflite_model)
```

[18]:
```
!edge-impulse-uploader --api-key␣
↪ei_57339804411126b0ac00c95dcf593f5989bd296457674b6130ab878938581f02 model.
↪tflite
```

```
Edge Impulse uploader v1.26.0
Endpoints:
    API:         https://studio.edgeimpulse.com
    Ingestion:   https://ingestion.edgeimpulse.com

Upload configuration:
    Label:       Not set, will be inferred from file name
    Category:    training
Cannot handle this file, only .wav, .cbor, .json, .jpg, .jpeg, .png, .csv, .txt,
.mp4, .avi supported: model.tflite
```

As we can see above, the command provided does not support the tflite model, also this is confirmed in the online documentation https://docs.edgeimpulse.com/docs/tools/edge-impulse-cli/cli-uploader. Uploading a model can be done through the Edge Impulse web GUI.

[ ]: