t81_558_class_11_04_hf_train

June 23, 2025

1 T81-558: Applications of Deep Neural Networks

Module 11: Natural Language Processing with Hugging Face * Instructor: Jeff Heaton, McKelvey School of Engineering, Washington University in St. Louis * For more information visit the class website.

2 Module 11 Material

- Part 11.1: Introduction to Hugging Face [Video] [Notebook]
- Part 11.2: Hugging Face Tokenizers [Video] [Notebook]
- Part 11.3: Hugging Face Datasets [Video] [Notebook]
- **Part 11.4: Training Hugging Face Models** [Video] [Notebook]
- Part 11.5: What are Embedding Layers in Keras [Video] [Notebook]

3 Google CoLab Instructions

The following code ensures that Google CoLab is running the correct version of TensorFlow.

```
[ ]: try:
         %tensorflow_version 2.x
         COLAB = True
         print("Note: using Google CoLab")
     except:
         print("Note: not using Google CoLab")
         COLAB = False
```

```
Note: using Google CoLab
```

4 Part 11.4: Training Hugging Face Models

Up to this point, we've used data and models from the Hugging Face hub unmodified. In this section, we will transfer and train a Hugging Face model. We will use Hugging Face data sets, tokenizers, and pretrained models to achieve this training.

We begin by installing Hugging Face if needed. It is also essential to install Hugging Face datasets.

```
[ ]: # HIDE OUTPUT
     !pip install transformers
```

```
!pip install transformers[sentencepiece]
!pip install datasets
```

```
Collecting transformers
  Downloading transformers-4.17.0-py3-none-any.whl (3.8 MB)
     |                 | 3.8 MB 15.1 MB/s
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-
packages (from transformers) (2.23.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.7/dist-packages (from transformers) (2019.12.20)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-
packages (from transformers) (3.6.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-
packages (from transformers) (1.21.5)
Collecting tokenizers!=0.11.3,>=0.11.1
  Downloading
tokenizers-0.11.6-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (6.5
MB)
     |                 | 6.5 MB 56.8 MB/s
Collecting pyyaml>=5.1
  Downloading PyYAML-6.0-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_6
4.whl (596 kB)
     |                 | 596 kB 72.2 MB/s
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.7/dist-packages (from transformers) (21.3)
Collecting sacremoses
  Downloading sacremoses-0.0.49-py3-none-any.whl (895 kB)
     |                 | 895 kB 65.0 MB/s
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.7/dist-packages (from transformers) (4.63.0)
Requirement already satisfied: importlib-metadata in
/usr/local/lib/python3.7/dist-packages (from transformers) (4.11.3)
Collecting huggingface-hub<1.0,>=0.1.0
  Downloading huggingface_hub-0.4.0-py3-none-any.whl (67 kB)
     |                 | 67 kB 6.9 MB/s
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.7/dist-packages (from huggingface-
hub<1.0,>=0.1.0->transformers) (3.10.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
/usr/local/lib/python3.7/dist-packages (from packaging>=20.0->transformers)
(3.0.7)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-
packages (from importlib-metadata->transformers) (3.7.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.7/dist-packages (from requests->transformers) (2021.10.8)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.7/dist-packages (from requests->transformers) (3.0.4)
```

```
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.7/dist-packages (from requests->transformers) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-
packages (from requests->transformers) (2.10)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages
(from sacremoses->transformers) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages
(from sacremoses->transformers) (1.1.0)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages
(from sacremoses->transformers) (1.15.0)
Installing collected packages: pyyaml, tokenizers, sacremoses, huggingface-hub,
transformers
  Attempting uninstall: pyyaml
    Found existing installation: PyYAML 3.13
    Uninstalling PyYAML-3.13:
      Successfully uninstalled PyYAML-3.13
Successfully installed huggingface-hub-0.4.0 pyyaml-6.0 sacremoses-0.0.49
tokenizers-0.11.6 transformers-4.17.0
Requirement already satisfied: transformers[sentencepiece] in
/usr/local/lib/python3.7/dist-packages (4.17.0)
Requirement already satisfied: sacremoses in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (0.0.49)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (3.6.0)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (4.63.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (21.3)
Requirement already satisfied: importlib-metadata in
/usr/local/lib/python3.7/dist-packages (from transformers[sentencepiece])
(4.11.3)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (2.23.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.7/dist-packages (from transformers[sentencepiece])
(2019.12.20)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (6.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (1.21.5)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in
/usr/local/lib/python3.7/dist-packages (from transformers[sentencepiece])
(0.4.0)
Requirement already satisfied: tokenizers!=0.11.3,>=0.11.1 in
/usr/local/lib/python3.7/dist-packages (from transformers[sentencepiece])
(0.11.6)
Collecting sentencepiece!=0.1.92,>=0.1.91
  Downloading
```

```
sentencepiece-0.1.96-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(1.2 MB)
     |                 | 1.2 MB 14.4 MB/s
Requirement already satisfied: protobuf in /usr/local/lib/python3.7/dist-
packages (from transformers[sentencepiece]) (3.17.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.7/dist-packages (from huggingface-
hub<1.0,>=0.1.0->transformers[sentencepiece]) (3.10.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
/usr/local/lib/python3.7/dist-packages (from
packaging>=20.0->transformers[sentencepiece]) (3.0.7)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-
packages (from importlib-metadata->transformers[sentencepiece]) (3.7.0)
Requirement already satisfied: six>=1.9 in /usr/local/lib/python3.7/dist-
packages (from protobuf->transformers[sentencepiece]) (1.15.0)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-
packages (from requests->transformers[sentencepiece]) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.7/dist-packages (from
requests->transformers[sentencepiece]) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.7/dist-packages (from
requests->transformers[sentencepiece]) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.7/dist-packages (from
requests->transformers[sentencepiece]) (2021.10.8)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages
(from sacremoses->transformers[sentencepiece]) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages
(from sacremoses->transformers[sentencepiece]) (1.1.0)
Installing collected packages: sentencepiece
Successfully installed sentencepiece-0.1.96
Collecting datasets
  Downloading datasets-2.0.0-py3-none-any.whl (325 kB)
     |                 | 325 kB 14.5 MB/s
Collecting xxhash
  Downloading
xxhash-3.0.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (212 kB)
     |                 | 212 kB 70.9 MB/s
Requirement already satisfied: multiprocess in
/usr/local/lib/python3.7/dist-packages (from datasets) (0.70.12.2)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.1.0 in
/usr/local/lib/python3.7/dist-packages (from datasets) (0.4.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-
packages (from datasets) (1.21.5)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.7/dist-packages (from datasets) (2.23.0)
Requirement already satisfied: pyarrow>=5.0.0 in /usr/local/lib/python3.7/dist-
```

```
packages (from datasets) (6.0.1)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.7/dist-
packages (from datasets) (4.63.0)
Collecting responses<0.19
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Requirement already satisfied: dill in /usr/local/lib/python3.7/dist-packages
(from datasets) (0.3.4)
Requirement already satisfied: importlib-metadata in
/usr/local/lib/python3.7/dist-packages (from datasets) (4.11.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-
packages (from datasets) (21.3)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages
(from datasets) (1.3.5)
Collecting aiohttp
  Downloading aiohttp-3.8.1-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_6
4.whl (1.1 MB)
     |                 | 1.1 MB 62.9 MB/s
Collecting fsspec[http]>=2021.05.0
  Downloading fsspec-2022.2.0-py3-none-any.whl (134 kB)
     |                 | 134 kB 74.3 MB/s
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-
packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (3.6.0)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.7/dist-packages
(from huggingface-hub<1.0.0,>=0.1.0->datasets) (6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.7/dist-packages (from huggingface-
hub<1.0.0,>=0.1.0->datasets) (3.10.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
/usr/local/lib/python3.7/dist-packages (from packaging->datasets) (3.0.7)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets)
(1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-
packages (from requests>=2.19.0->datasets) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets)
(2021.10.8)
Collecting urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1
  Downloading urllib3-1.25.11-py2.py3-none-any.whl (127 kB)
     |                 | 127 kB 76.1 MB/s
Collecting asynctest==0.13.0
  Downloading asynctest-0.13.0-py3-none-any.whl (26 kB)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-
packages (from aiohttp->datasets) (21.4.0)
Collecting multidict<7.0,>=4.5
```

```
  Downloading
multidict-6.0.2-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (94
kB)
     |                 | 94 kB 4.6 MB/s
Collecting yarl<2.0,>=1.0
  Downloading yarl-1.7.2-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_6
4.whl (271 kB)
     |                 | 271 kB 65.7 MB/s
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.3.0-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_6
4.whl (144 kB)
     |                 | 144 kB 78.0 MB/s
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Collecting aiosignal>=1.1.2
  Downloading aiosignal-1.2.0-py3-none-any.whl (8.2 kB)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in
/usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (2.0.12)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-
packages (from importlib-metadata->datasets) (3.7.0)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-
packages (from pandas->datasets) (2018.9)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-
packages (from python-dateutil>=2.7.3->pandas->datasets) (1.15.0)
Installing collected packages: multidict, frozenlist, yarl, urllib3, asynctest,
async-timeout, aiosignal, fsspec, aiohttp, xxhash, responses, datasets
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.24.3
    Uninstalling urllib3-1.24.3:
      Successfully uninstalled urllib3-1.24.3
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.

datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 which is

incompatible.
Successfully installed aiohttp-3.8.1 aiosignal-1.2.0 async-timeout-4.0.2
asynctest-0.13.0 datasets-2.0.0 frozenlist-1.3.0 fsspec-2022.2.0 multidict-6.0.2
responses-0.18.0 urllib3-1.25.11 xxhash-3.0.0 yarl-1.7.2
```

We begin by loading the emotion data set from the Hugging Face hub. Emotion is a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. The following code loads the emotion data set from the Hugging Face hub.

```
[ ]: # HIDE OUTPUT
     from datasets import load_dataset

     emotions = load_dataset("emotion")
```

```
Downloading builder script:   0%|          | 0.00/1.66k [00:00<?, ?B/s]

Downloading metadata:   0%|          | 0.00/1.61k [00:00<?, ?B/s]

Using custom data configuration default

Downloading and preparing dataset emotion/default (download: 1.97 MiB,
generated: 2.07 MiB, post-processed: Unknown size, total: 4.05 MiB) to /root/.ca
che/huggingface/datasets/emotion/default/0.0.0/348f63ca8e27b3713b6c04d723efe6d82
4a56fb3d1449794716c0f0296072705…

Downloading data:   0%|          | 0.00/1.66M [00:00<?, ?B/s]

Downloading data:   0%|          | 0.00/204k [00:00<?, ?B/s]

Downloading data:   0%|          | 0.00/207k [00:00<?, ?B/s]

Generating train split:   0%|          | 0/16000 [00:00<?, ? examples/s]

Generating validation split:   0%|          | 0/2000 [00:00<?, ? examples/s]

Generating test split:   0%|          | 0/2000 [00:00<?, ? examples/s]

Dataset emotion downloaded and prepared to /root/.cache/huggingface/datasets/emo
tion/default/0.0.0/348f63ca8e27b3713b6c04d723efe6d824a56fb3d1449794716c0f0296072
705. Subsequent calls will reuse this data.

  0%|          | 0/3 [00:00<?, ?it/s]
```

You can see a single observation from the training data set here. This observation includes both the text sample and the assigned emotion label. The label is a numeric index representing the assigned emotion.

```
[ ]: emotions['train'][2]
```

```
[ ]: {'label': 3, 'text': 'im grabbing a minute to post i feel greedy wrong'}
```

We can display the labels in order of their index labels.

```
[ ]: emotions['train'].features
```

```
[ ]: {'label': ClassLabel(num_classes=6, names=['sadness', 'joy', 'love', 'anger',
     'fear', 'surprise'], id=None),
      'text': Value(dtype='string', id=None)}
```

Next, we utilize Hugging Face tokenizers and data sets together. The following code tokenizes the entire emotion data set. You can see below that the code has transformed the training set into subword tokens that are now ready to be used in conjunction with a transformer for either inference or training.

```
[ ]: # HIDE OUTPUT
     from transformers import AutoTokenizer


     def tokenize(rows):
         return tokenizer(rows['text'], padding="max_length", truncation=True)


     model_ckpt = "distilbert-base-uncased"
     tokenizer = AutoTokenizer.from_pretrained(model_ckpt)

     emotions.set_format(type=None)

     tokenized_datasets = emotions.map(tokenize, batched=True)
```

```
Downloading:   0%|          | 0.00/28.0 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/483 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/226k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/455k [00:00<?, ?B/s]

  0%|          | 0/16 [00:00<?, ?ba/s]

  0%|          | 0/2 [00:00<?, ?ba/s]

  0%|          | 0/2 [00:00<?, ?ba/s]
```

We will utilize the Hugging Face **DefaultDataCollator** to transform the emotion data set into TensorFlow type data that we can use to finetune a neural network.

```
[ ]: from transformers import DefaultDataCollator

     data_collator = DefaultDataCollator(return_tensors="tf")
```

Now we generate a shuffled training and evaluation data set.

```
[ ]: small_train_dataset = tokenized_datasets["train"].shuffle(seed=42)
     small_eval_dataset = tokenized_datasets["test"].shuffle(seed=42)
```

We can now generate the TensorFlow data sets. We specify which columns should map to the input features and labels. We do not need to shuffle because we previously shuffled the data.

```
[ ]: tf_train_dataset = small_train_dataset.to_tf_dataset(
         columns=["attention_mask", "input_ids", "token_type_ids"],
         label_cols=["labels"],
         shuffle=True,
         collate_fn=data_collator,
         batch_size=8,
     )
```

```
tf_validation_dataset = small_eval_dataset.to_tf_dataset(
    columns=["attention_mask", "input_ids", "token_type_ids"],
    label_cols=["labels"],
    shuffle=False,
    collate_fn=data_collator,
    batch_size=8,
)
```

We will now load the distilbert model for classification. We will adjust the pretrained weights to predict the emotions of text lines.

[ ]:

```
# HIDE OUTPUT
import tensorflow as tf
from transformers import TFAutoModelForSequenceClassification

model = TFAutoModelForSequenceClassification.from_pretrained(\
    "distilbert-base-uncased", num_labels=6)
```

```
Downloading:   0%|          | 0.00/347M [00:00<?, ?B/s]
```

```
Some layers from the model checkpoint at distilbert-base-uncased were not used
when initializing TFDistilBertForSequenceClassification: ['vocab_layer_norm',
'vocab_transform', 'vocab_projector', 'activation_13']
- This IS expected if you are initializing TFDistilBertForSequenceClassification
from the checkpoint of a model trained on another task or with another
architecture (e.g. initializing a BertForSequenceClassification model from a
BertForPreTraining model).
- This IS NOT expected if you are initializing
TFDistilBertForSequenceClassification from the checkpoint of a model that you
expect to be exactly identical (initializing a BertForSequenceClassification
model from a BertForSequenceClassification model).
Some layers of TFDistilBertForSequenceClassification were not initialized from
the model checkpoint at distilbert-base-uncased and are newly initialized:
['pre_classifier', 'classifier', 'dropout_19']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

We now train the neural network. Because the network is already pretrained, we use a small learning rate.

[ ]:

```
model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=5e-5),
    loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
    metrics=tf.metrics.SparseCategoricalAccuracy(),
)

model.fit(tf_train_dataset, validation_data=tf_validation_dataset,
          epochs=5)
```

```
Epoch 1/5
```

```
2000/2000 [==============================] - 360s 174ms/step - loss: 0.3720 -
sparse_categorical_accuracy: 0.8669 - val_loss: 0.1728 -
val_sparse_categorical_accuracy: 0.9180
Epoch 2/5
2000/2000 [==============================] - 347s 174ms/step - loss: 0.1488 -
sparse_categorical_accuracy: 0.9338 - val_loss: 0.1496 -
val_sparse_categorical_accuracy: 0.9295
Epoch 3/5
2000/2000 [==============================] - 347s 173ms/step - loss: 0.1253 -
sparse_categorical_accuracy: 0.9420 - val_loss: 0.1617 -
val_sparse_categorical_accuracy: 0.9245
Epoch 4/5
2000/2000 [==============================] - 346s 173ms/step - loss: 0.1092 -
sparse_categorical_accuracy: 0.9486 - val_loss: 0.1654 -
val_sparse_categorical_accuracy: 0.9295
Epoch 5/5
2000/2000 [==============================] - 347s 173ms/step - loss: 0.0960 -
sparse_categorical_accuracy: 0.9585 - val_loss: 0.1830 -
val_sparse_categorical_accuracy: 0.9220
```

```
[ ]: <keras.callbacks.History at 0x7f42e84a49d0>
```