SageMaker Project

December 31, 2019

1 Creating a Sentiment Analysis Web App

1.1 Using PyTorch and SageMaker

Deep Learning Nanodegree Program | Deployment

Now that we have a basic understanding of how SageMaker works we will try to use it to construct a complete project from end to end. Our goal will be to have a simple web page which a user can use to enter a movie review. The web page will then send the review off to our deployed model which will predict the sentiment of the entered review.

1.2 Instructions

Some template code has already been provided for you, and you will need to implement additional functionality to successfully complete this notebook. You will not need to modify the included code beyond what is requested. Sections that begin with '**TODO**' in the header indicate that you need to complete or implement some portion within them. Instructions will be provided for each section and the specifics of the implementation are marked in the code block with a `#` `TODO: ...` comment. Please be sure to read the instructions carefully!

In addition to implementing code, there will be questions for you to answer which relate to the task and your implementation. Each section where you will answer a question is preceded by a '**Question:**' header. Carefully read each question and provide your answer below the '**Answer:**' header by editing the Markdown cell.

> **Note**: Code and Markdown cells can be executed using the **Shift+Enter** keyboard shortcut. In addition, a cell can be edited by typically clicking it (double-click for Markdown cells) or by pressing **Enter** while it is highlighted.

1.3 General Outline

Recall the general outline for SageMaker projects using a notebook instance.

1. Download or otherwise retrieve the data.
2. Process / Prepare the data.
3. Upload the processed data to S3.
4. Train a chosen model.

5. Test the trained model (typically using a batch transform job).
6. Deploy the trained model.
7. Use the deployed model.

For this project, you will be following the steps in the general outline with some modifications.

First, you will not be testing the model in its own step. You will still be testing the model, however, you will do it by deploying your model and then using the deployed model by sending the test data to it. One of the reasons for doing this is so that you can make sure that your deployed model is working correctly before moving forward.

In addition, you will deploy and use your trained model a second time. In the second iteration you will customize the way that your trained model is deployed by including some of your own code. In addition, your newly deployed model will be used in the sentiment analysis web app.

1.4 Step 1: Downloading the data

As in the XGBoost in SageMaker notebook, we will be using the IMDb dataset

Maas, Andrew L., et al. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011.

```
In [1]: %mkdir ../data
        !wget -O ../data/aclImdb_v1.tar.gz http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_
        !tar -zxf ../data/aclImdb_v1.tar.gz -C ../data

mkdir: cannot create directory ../data: File exists
--2019-12-03 08:55:45--  http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz
Resolving ai.stanford.edu (ai.stanford.edu)... 171.64.68.10
Connecting to ai.stanford.edu (ai.stanford.edu)|171.64.68.10|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 84125825 (80M) [application/x-gzip]
Saving to: ../data/aclImdb_v1.tar.gz

../data/aclImdb_v1. 100%[===================>]  80.23M  22.4MB/s    in 5.2s

2019-12-03 08:55:51 (15.5 MB/s) - ../data/aclImdb_v1.tar.gz saved [84125825/84125825]
```

1.5 Step 2: Preparing and Processing the data

Also, as in the XGBoost notebook, we will be doing some initial data processing. The first few steps are the same as in the XGBoost example. To begin with, we will read in each of the reviews and combine them into a single input structure. Then, we will split the dataset into a training set and a testing set.

```
In [2]: import os
        import glob
```

```python
def read_imdb_data(data_dir='../data/aclImdb'):
    data = {}
    labels = {}

    for data_type in ['train', 'test']:
        data[data_type] = {}
        labels[data_type] = {}

        for sentiment in ['pos', 'neg']:
            data[data_type][sentiment] = []
            labels[data_type][sentiment] = []

            path = os.path.join(data_dir, data_type, sentiment, '*.txt')
            files = glob.glob(path)

            for f in files:
                with open(f) as review:
                    data[data_type][sentiment].append(review.read())
                    # Here we represent a positive review by '1' and a negative review b
                    labels[data_type][sentiment].append(1 if sentiment == 'pos' else 0)

            assert len(data[data_type][sentiment]) == len(labels[data_type][sentiment]),
                    "{}/{} data size does not match labels size".format(data_type, senti

    return data, labels
```

```
In [3]: data, labels = read_imdb_data()
        print("IMDB reviews: train = {} pos / {} neg, test = {} pos / {} neg".format(
                len(data['train']['pos']), len(data['train']['neg']),
                len(data['test']['pos']), len(data['test']['neg'])))

IMDB reviews: train = 12500 pos / 12500 neg, test = 12500 pos / 12500 neg
```

Now that we've read the raw training and testing data from the downloaded dataset, we will combine the positive and negative reviews and shuffle the resulting records.

```python
In [4]: from sklearn.utils import shuffle

        def prepare_imdb_data(data, labels):
            """Prepare training and test sets from IMDb movie reviews."""

            #Combine positive and negative reviews and labels
            data_train = data['train']['pos'] + data['train']['neg']
            data_test = data['test']['pos'] + data['test']['neg']
            labels_train = labels['train']['pos'] + labels['train']['neg']
            labels_test = labels['test']['pos'] + labels['test']['neg']

            #Shuffle reviews and corresponding labels within training and test sets
```

3

```
            data_train, labels_train = shuffle(data_train, labels_train)
            data_test, labels_test = shuffle(data_test, labels_test)

            # Return a unified training data, test data, training labels, test labets
            return data_train, data_test, labels_train, labels_test
```

```
In [5]: train_X, test_X, train_y, test_y = prepare_imdb_data(data, labels)
        print("IMDb reviews (combined): train = {}, test = {}".format(len(train_X), len(test_X))
```

```
IMDb reviews (combined): train = 25000, test = 25000
```

Now that we have our training and testing sets unified and prepared, we should do a quick check and see an example of the data our model will be trained on. This is generally a good idea as it allows you to see how each of the further processing steps affects the reviews and it also ensures that the data has been loaded correctly.

```
In [6]: print(train_X[100])
        print(train_y[100])
```

```
Frailty--8/10--It's non-sensical title and "Bill Paxton Directs" headline aside, this is a prett
1
```

The first step in processing the reviews is to make sure that any html tags that appear should be removed. In addition we wish to tokenize our input, that way words such as *entertained* and *entertaining* are considered the same with regard to sentiment analysis.

```
In [7]: import nltk
        from nltk.corpus import stopwords
        from nltk.stem.porter import *

        import re
        from bs4 import BeautifulSoup

        def review_to_words(review):
            nltk.download("stopwords", quiet=True)
            stemmer = PorterStemmer()

            text = BeautifulSoup(review, "html.parser").get_text() # Remove HTML tags
            text = re.sub(r"[^a-zA-Z0-9]", " ", text.lower()) # Convert to lower case
            words = text.split() # Split string into words
            words = [w for w in words if w not in stopwords.words("english")] # Remove stopwords
            words = [PorterStemmer().stem(w) for w in words] # stem

            return words
```

The `review_to_words` method defined above uses `BeautifulSoup` to remove any html tags that appear and uses the `nltk` package to tokenize the reviews. As a check to ensure we know how everything is working, try applying `review_to_words` to one of the reviews in the training set.

```
In [8]: # TODO: Apply review_to_words to a review (train_X[100] or any other review)
        review_to_words(train_X[0])

Out[8]: ['skip',
         'mission',
         'galactica',
         'watch',
         'origin',
         'live',
         'legend',
         'episod',
         'instead',
         'network',
         'took',
         'part',
         '1',
         '2',
         'live',
         'legend',
         'jam',
         'one',
         'plot',
         'aw',
         'fire',
         'space',
         'episod',
         'although',
         'galactica',
         'suffer',
         'network',
         'control',
         'write',
         'lack',
         'time',
         'prepar',
         'proper',
         'product',
         'live',
         'legend',
         'best',
         '1978',
         'tv',
         'seri',
         'fire',
         'space',
         'one',
         'worst',
         'episod',
```

```
    'histor',
    'note',
    'watch',
    'galactica',
    'origin',
    'star',
    'trek',
    'reviv',
    'trek',
    'seri',
    'see',
    'differ',
    'qualiti',
    'network',
    'produc',
    'sci',
    'fi',
    'syndic',
    'sci',
    'fi']
```

Question: Above we mentioned that `review_to_words` method removes html formatting and allows us to tokenize the words found in a review, for example, converting *entertained* and *entertaining* into *entertain* so that they are treated as though they are the same word. What else, if anything, does this method do to the input?

Answer: It conveniently puts all the stemmed words into an array. This is helpful to easily convert integer representations of most common words using CountVectorizer. Also it puts all the extracted words into lowercase to make sure that the formatting does not affect the influence of a word representation especially when uppercase and lowercase versions of a particular word is deemed by the model separate which in this case is not.

The method below applies the `review_to_words` method to each of the reviews in the training and testing datasets. In addition it caches the results. This is because performing this processing step can take a long time. This way if you are unable to complete the notebook in the current session, you can come back without needing to process the data a second time.

```
In [9]: import pickle

        cache_dir = os.path.join("../cache", "sentiment_analysis")  # where to store cache files
        os.makedirs(cache_dir, exist_ok=True)  # ensure cache directory exists

        def preprocess_data(data_train, data_test, labels_train, labels_test,
                            cache_dir=cache_dir, cache_file="preprocessed_data.pkl"):
            """Convert each review to words; read from cache if available."""

            # If cache_file is not None, try to read from it first
            cache_data = None
            if cache_file is not None:
                try:
```

```python
            with open(os.path.join(cache_dir, cache_file), "rb") as f:
                cache_data = pickle.load(f)
            print("Read preprocessed data from cache file:", cache_file)
        except:
            pass  # unable to read from cache, but that's okay

    # If cache is missing, then do the heavy lifting
    if cache_data is None:
        # Preprocess training and test data to obtain words for each review
        #words_train = list(map(review_to_words, data_train))
        #words_test = list(map(review_to_words, data_test))
        words_train = [review_to_words(review) for review in data_train]
        words_test = [review_to_words(review) for review in data_test]

        # Write to cache file for future runs
        if cache_file is not None:
            cache_data = dict(words_train=words_train, words_test=words_test,
                              labels_train=labels_train, labels_test=labels_test)
            with open(os.path.join(cache_dir, cache_file), "wb") as f:
                pickle.dump(cache_data, f)
            print("Wrote preprocessed data to cache file:", cache_file)
    else:
        # Unpack data loaded from cache file
        words_train, words_test, labels_train, labels_test = (cache_data['words_train'],
                cache_data['words_test'], cache_data['labels_train'], cache_data['labels

    return words_train, words_test, labels_train, labels_test
```

```python
In [10]: # Preprocess data
         train_X, test_X, train_y, test_y = preprocess_data(train_X, test_X, train_y, test_y)
```

```
Read preprocessed data from cache file: preprocessed_data.pkl
```

1.6 Transform the data

In the XGBoost notebook we transformed the data from its word representation to a bag-of-words feature representation. For the model we are going to construct in this notebook we will construct a feature representation which is very similar. To start, we will represent each word as an integer. Of course, some of the words that appear in the reviews occur very infrequently and so likely don't contain much information for the purposes of sentiment analysis. The way we will deal with this problem is that we will fix the size of our working vocabulary and we will only include the words that appear most frequently. We will then combine all of the infrequent words into a single category and, in our case, we will label it as 1.

Since we will be using a recurrent neural network, it will be convenient if the length of each review is the same. To do this, we will fix a size for our reviews and then pad short reviews with the category 'no word' (which we will label 0) and truncate long reviews.

7

1.6.1 (TODO) Create a word dictionary

To begin with, we need to construct a way to map words that appear in the reviews to integers. Here we fix the size of our vocabulary (including the 'no word' and 'infrequent' categories) to be 5000 but you may wish to change this to see how it affects the model.

> **TODO:** Complete the implementation for the `build_dict()` method below. Note that even though the vocab_size is set to 5000, we only want to construct a mapping for the most frequently appearing 4998 words. This is because we want to reserve the special labels 0 for 'no word' and 1 for 'infrequent word'.

```python
In [58]: import numpy as np
         import collections

         def build_dict(data, vocab_size = 5000):
             """Construct and return a dictionary mapping each of the most frequently appearing

             # TODO: Determine how often each word appears in `data`. Note that `data` is a list
             #        sentence is a list of words.
             c = collections.Counter()

             for sentence in data:
                 c.update(sentence)

             word_count = dict(c.most_common()) # A dict storing the words that appear in the re

             # TODO: Sort the words found in `data` so that sorted_words[0] is the most frequent
             #        sorted_words[-1] is the least frequently appearing word.

             sorted_words = [word for word in word_count.keys()]

             word_dict = {} # This is what we are building, a dictionary that translates words i
             for idx, word in enumerate(sorted_words[:vocab_size - 2]): # The -2 is so that we s
                 word_dict[word] = idx + 2                                # 'infrequent' labels

             return word_dict
```

```python
In [60]: word_dict = build_dict(train_X)
```

> **Question:** What are the five most frequently appearing (tokenized) words in the training set? Does it makes sense that these words appear frequently in the training set?
> **Answer:**

```python
In [66]: # TODO: Use this space to determine the five most frequently appearing words in the tra
         print(list(word_dict.keys())[:5])
```

```
['movi', 'film', 'one', 'like', 'time']
```

For me it makes sense to have **movie** and **film** keywords to be the most frequent keywords. **One**, **like** and **time** in the other hand came on top as well since most reviews talk about comparison to other movies or judging if the movie is best or worst at that time. So words like that are used.

1.6.2 Save `word_dict`

Later on when we construct an endpoint which processes a submitted review we will need to make use of the `word_dict` which we have created. As such, we will save it to a file now for future use.

```
In [67]: data_dir = '../data/pytorch' # The folder we will use for storing data
         if not os.path.exists(data_dir): # Make sure that the folder exists
             os.makedirs(data_dir)
```

```
In [68]: with open(os.path.join(data_dir, 'word_dict.pkl'), "wb") as f:
             pickle.dump(word_dict, f)
```

1.6.3 Transform the reviews

Now that we have our word dictionary which allows us to transform the words appearing in the reviews into integers, it is time to make use of it and convert our reviews to their integer sequence representation, making sure to pad or truncate to a fixed length, which in our case is 500.

```
In [69]: def convert_and_pad(word_dict, sentence, pad=500):
             NOWORD = 0 # We will use 0 to represent the 'no word' category
             INFREQ = 1 # and we use 1 to represent the infrequent words, i.e., words not appear

             working_sentence = [NOWORD] * pad

             for word_index, word in enumerate(sentence[:pad]):
                 if word in word_dict:
                     working_sentence[word_index] = word_dict[word]
                 else:
                     working_sentence[word_index] = INFREQ

             return working_sentence, min(len(sentence), pad)

         def convert_and_pad_data(word_dict, data, pad=500):
             result = []
             lengths = []

             for sentence in data:
                 converted, leng = convert_and_pad(word_dict, sentence, pad)
                 result.append(converted)
                 lengths.append(leng)

             return np.array(result), np.array(lengths)
```

9

```
In [70]: train_X, train_X_len = convert_and_pad_data(word_dict, train_X)
         test_X, test_X_len = convert_and_pad_data(word_dict, test_X)
```

As a quick check to make sure that things are working as intended, check to see what one of the reviews in the training set looks like after having been processeed. Does this look reasonable? What is the length of a review in the training set?

```
In [75]: # Use this cell to examine one of the processed reviews to make sure everything is work
         print(train_X[0])
         print("Number of words for the first observation in the training set: " + str(train_X_l
```

```
[ 541    1  627 3560   76  489 1974    1    1 4314 4314   89 3507 2266
  150    1  575 2075 4403 3560 3871    1    1 1323 2068 3816    5    1
  152    8  399 1865  150  561   52    1  216    9 3311    1 1278   22
    2   82    2  312   52 3560  336  373    4    9  613   57    1  398
  711 3638   26    1    1   37  820    1 1496    1  297 2678    1    4
    9   33    1  132 1780 1261 1336    1 1261   13  722   89 1581 4543
    1   49 3560   76   71 1870   42 1444    1 4436  130 2210 3198 1198
  305 2236 1991 2644  951    1  613 1023 3376   88   72  985   31  540
  423  714   35  939 2196 1336   89  118  286  461 1258  176   26 1214
   30  116 1701 3131 1480   89   29  616  119   25 2108  537    1  848
 3560    1    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0]
Number of words for the first observation in the training set: 142
```

Question: In the cells above we use the `preprocess_data` and `convert_and_pad_data` methods to process both the training and testing set. Why or why not might this be a problem?

Answer:

It may not be a problem since neural networks work optimally on vector inputs. It may be a problem since most of the values inside a vector are zeros assuming that the review length is short. Remembering what Andrew Trask discussed about using neural networks on zero values, it may deal a significant amount of time computing gradients on zero inputs which as we all know returns zero as result.

1.7   Step 3: Upload the data to S3

As in the XGBoost notebook, we will need to upload the training dataset to S3 in order for our training code to access it. For now we will save it locally and we will upload to S3 later on.

1.7.1   Save the processed training dataset locally

It is important to note the format of the data that we are saving as we will need to know it when we write the training code. In our case, each row of the dataset has the form `label, length, review[500]` where `review[500]` is a sequence of 500 integers representing the words in the review.

```
In [76]: import pandas as pd

         pd.concat([pd.DataFrame(train_y), pd.DataFrame(train_X_len), pd.DataFrame(train_X)], ax
                 .to_csv(os.path.join(data_dir, 'train.csv'), header=False, index=False)
```

1.7.2   Uploading the training data

Next, we need to upload the training data to the SageMaker default S3 bucket so that we can provide access to it while training our model.

```
In [77]: import sagemaker

         sagemaker_session = sagemaker.Session()

         bucket = sagemaker_session.default_bucket()
         prefix = 'sagemaker/sentiment_rnn'

         role = sagemaker.get_execution_role()

In [78]: input_data = sagemaker_session.upload_data(path=data_dir, bucket=bucket, key_prefix=pre
```

NOTE: The cell above uploads the entire contents of our data directory. This includes the `word_dict.pkl` file. This is fortunate as we will need this later on when we create an endpoint that accepts an arbitrary review. For now, we will just take note of the fact that it resides in the data directory (and so also in the S3 training bucket) and that we will need to make sure it gets saved in the model directory.

1.8 Step 4: Build and Train the PyTorch Model

In the XGBoost notebook we discussed what a model is in the SageMaker framework. In particular, a model comprises three objects

- Model Artifacts,
- Training Code, and
- Inference Code,

each of which interact with one another. In the XGBoost example we used training and inference code that was provided by Amazon. Here we will still be using containers provided by Amazon with the added benefit of being able to include our own custom code.

We will start by implementing our own neural network in PyTorch along with a training script. For the purposes of this project we have provided the necessary model object in the `model.py` file, inside of the `train` folder. You can see the provided implementation by running the cell below.

```
In [79]: !pygmentize train/model.py

import torch.nn as nn

class LSTMClassifier(nn.Module):
    """
    This is the simple RNN model we will be using to perform Sentiment Analysis.
    """

    def __init__(self, embedding_dim, hidden_dim, vocab_size):
        """
        Initialize the model by settingg up the various layers.
        """
        super(LSTMClassifier, self).__init__()

        self.embedding = nn.Embedding(vocab_size, embedding_dim, padding_idx=0)
        self.lstm = nn.LSTM(embedding_dim, hidden_dim)
        self.dense = nn.Linear(in_features=hidden_dim, out_features=1)
        self.sig = nn.Sigmoid()

        self.word_dict = None

    def forward(self, x):
        """
        Perform a forward pass of our model on some input.
        """
        x = x.t()
        lengths = x[0,:]
        reviews = x[1:,:]
        embeds = self.embedding(reviews)
        lstm_out, _ = self.lstm(embeds)
        out = self.dense(lstm_out)
        out = out[lengths - 1, range(len(lengths))]
```

```
        return self.sig(out.squeeze())
```

The important takeaway from the implementation provided is that there are three parameters that we may wish to tweak to improve the performance of our model. These are the embedding dimension, the hidden dimension and the size of the vocabulary. We will likely want to make these parameters configurable in the training script so that if we wish to modify them we do not need to modify the script itself. We will see how to do this later on. To start we will write some of the training code in the notebook so that we can more easily diagnose any issues that arise.

First we will load a small portion of the training data set to use as a sample. It would be very time consuming to try and train the model completely in the notebook as we do not have access to a gpu and the compute instance that we are using is not particularly powerful. However, we can work on a small bit of the data to get a feel for how our training script is behaving.

```
In [80]: import torch
         import torch.utils.data

         # Read in only the first 250 rows
         train_sample = pd.read_csv(os.path.join(data_dir, 'train.csv'), header=None, names=None

         # Turn the input pandas dataframe into tensors
         train_sample_y = torch.from_numpy(train_sample[[0]].values).float().squeeze()
         train_sample_X = torch.from_numpy(train_sample.drop([0], axis=1).values).long()

         # Build the dataset
         train_sample_ds = torch.utils.data.TensorDataset(train_sample_X, train_sample_y)
         # Build the dataloader
         train_sample_dl = torch.utils.data.DataLoader(train_sample_ds, batch_size=50)
```

1.8.1 (TODO) Writing the training method

Next we need to write the training code itself. This should be very similar to training methods that you have written before to train PyTorch models. We will leave any difficult aspects such as model saving / loading and parameter loading until a little later.

```
In [83]: def train(model, train_loader, epochs, optimizer, loss_fn, device):
             for epoch in range(1, epochs + 1):
                 model.train()
                 total_loss = 0
                 for batch in train_loader:
                     batch_X, batch_y = batch

                     batch_X = batch_X.to(device)
                     batch_y = batch_y.to(device)

                     # TODO: Complete this train method to train the model provided.
                     optimizer.zero_grad()
                     output = model(batch_X)
                     loss = loss_fn(output, batch_y)
```

13

```
                loss.backward()
                optimizer.step()

                total_loss += loss.data.item()
            print("Epoch: {}, BCELoss: {}".format(epoch, total_loss / len(train_loader)))
```

Supposing we have the training method above, we will test that it is working by writing a bit of code in the notebook that executes our training method on the small sample training set that we loaded earlier. The reason for doing this in the notebook is so that we have an opportunity to fix any errors that arise early when they are easier to diagnose.

```
In [84]: import torch.optim as optim
         from train.model import LSTMClassifier

         device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
         model = LSTMClassifier(32, 100, 5000).to(device)
         optimizer = optim.Adam(model.parameters())
         loss_fn = torch.nn.BCELoss()

         train(model, train_sample_dl, 5, optimizer, loss_fn, device)

Epoch: 1, BCELoss: 0.6955353617668152
Epoch: 2, BCELoss: 0.6862450242042542
Epoch: 3, BCELoss: 0.6784801125526428
Epoch: 4, BCELoss: 0.6700268268585206
Epoch: 5, BCELoss: 0.659886360168457
```

In order to construct a PyTorch model using SageMaker we must provide SageMaker with a training script. We may optionally include a directory which will be copied to the container and from which our training code will be run. When the training container is executed it will check the uploaded directory (if there is one) for a `requirements.txt` file and install any required Python libraries, after which the training script will be run.

1.8.2   (TODO) Training the model

When a PyTorch model is constructed in SageMaker, an entry point must be specified. This is the Python file which will be executed when the model is trained. Inside of the `train` directory is a file called `train.py` which has been provided and which contains most of the necessary code to train our model. The only thing that is missing is the implementation of the `train()` method which you wrote earlier in this notebook.

TODO: Copy the `train()` method written above and paste it into the `train/train.py` file where required.

The way that SageMaker passes hyperparameters to the training script is by way of arguments. These arguments can then be parsed and used in the training script. To see how this is done take a look at the provided `train/train.py` file.

```
In [85]: from sagemaker.pytorch import PyTorch
```

```python
estimator = PyTorch(entry_point="train.py",
                    source_dir="train",
                    role=role,
                    framework_version='0.4.0',
                    train_instance_count=1,
                    train_instance_type='ml.p2.xlarge',
                    hyperparameters={
                        'epochs': 10,
                        'hidden_dim': 200,
                    })
```

```
In [86]: estimator.fit({'training': input_data})
```

```
2019-12-03 10:09:05 Starting - Starting the training job...
2019-12-03 10:09:07 Starting - Launching requested ML instances...
2019-12-03 10:10:10 Starting - Preparing the instances for training...
2019-12-03 10:11:32 Downloading - Downloading input data...
2019-12-03 10:12:04 Training - Downloading the training image...bash: cannot set terminal proces
bash: no job control in this shell
2019-12-03 10:12:38,517 sagemaker-containers INFO     Imported framework sagemaker_pytorch_conta
2019-12-03 10:12:38,542 sagemaker_pytorch_container.training INFO     Block until all host DNS l
2019-12-03 10:12:38,545 sagemaker_pytorch_container.training INFO     Invoking user training scr
2019-12-03 10:12:38,776 sagemaker-containers INFO     Module train does not provide a setup.py.
Generating setup.py
2019-12-03 10:12:38,776 sagemaker-containers INFO     Generating setup.cfg
2019-12-03 10:12:38,777 sagemaker-containers INFO     Generating MANIFEST.in
2019-12-03 10:12:38,777 sagemaker-containers INFO     Installing module with the following comma
/usr/bin/python -m pip install -U . -r requirements.txt
Processing /opt/ml/code
Collecting pandas (from -r requirements.txt (line 1))  Downloading https://files.pythonhosted.or
Collecting numpy (from -r requirements.txt (line 2))  Downloading https://files.pythonhosted.org
Collecting nltk (from -r requirements.txt (line 3))  Downloading https://files.pythonhosted.org/
Collecting beautifulsoup4 (from -r requirements.txt (line 4))  Downloading https://files.pythonh
Collecting html5lib (from -r requirements.txt (line 5))  Downloading https://files.pythonhosted.
Collecting pytz>=2011k (from pandas->-r requirements.txt (line 1))  Downloading https://files.py
Requirement already satisfied, skipping upgrade: python-dateutil>=2.5.0 in /usr/local/lib/python
Requirement already satisfied, skipping upgrade: six in /usr/local/lib/python3.5/dist-packages (
Collecting soupsieve>=1.2 (from beautifulsoup4->-r requirements.txt (line 4))  Downloading https
Collecting webencodings (from html5lib->-r requirements.txt (line 5))  Downloading https://files
Building wheels for collected packages: nltk, train  Running setup.py bdist_wheel for nltk: star
  Running setup.py bdist_wheel for nltk: finished with status 'done'  Stored in directory: /root
Successfully built nltk train
Installing collected packages: pytz, numpy, pandas, nltk, soupsieve, beautifulsoup4, webencoding
Successfully installed beautifulsoup4-4.8.1 html5lib-1.0.1 nltk-3.4.5 numpy-1.17.4 pandas-0.24.2
You are using pip version 18.1, however version 19.3.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
2019-12-03 10:12:50,550 sagemaker-containers INFO     Invoking user script
Training Env:
```

```
{    "framework_module": "sagemaker_pytorch_container.training:main",    "additional_framework_p
}
Environment variables:
SM_FRAMEWORK_MODULE=sagemaker_pytorch_container.training:main
SM_HP_EPOCHS=10
SM_INPUT_DATA_CONFIG={"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicat
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_GPUS=1
SM_NETWORK_INTERFACE_NAME=eth0
SM_CHANNELS=["training"]
SM_USER_ARGS=["--epochs","10","--hidden_dim","200"]
SM_LOG_LEVEL=20
SM_CURRENT_HOST=algo-1
SM_HP_HIDDEN_DIM=200
SM_NUM_CPUS=4
SM_MODULE_NAME=train
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"training":"/opt/ml/
SM_USER_ENTRY_POINT=train.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_host":"algo-1","hosts":["algo-1"],"network_interface_name":"eth0"}
SM_INPUT_DIR=/opt/ml/input
SM_HOSTS=["algo-1"]
PYTHONPATH=/usr/local/bin:/usr/lib/python35.zip:/usr/lib/python3.5:/usr/lib/python3.5/plat-x86_6
SM_MODULE_DIR=s3://sagemaker-us-east-1-212294556799/sagemaker-pytorch-2019-12-03-10-09-04-848/so
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_MODEL_DIR=/opt/ml/model
SM_HPS={"epochs":10,"hidden_dim":200}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNEL_TRAINING=/opt/ml/input/data/training
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
Invoking script with the following command:
/usr/bin/python -m train --epochs 10 --hidden_dim 200
Using device cuda.
Get train data loader.

2019-12-03 10:12:38 Training - Training image download completed. Training in progress.Model loa
Epoch: 1, BCELoss: 0.6734105640528153
Epoch: 2, BCELoss: 0.5873278513246653
Epoch: 3, BCELoss: 0.5109929630951006
Epoch: 4, BCELoss: 0.44405045618816297
Epoch: 5, BCELoss: 0.37813643411714204
Epoch: 6, BCELoss: 0.34157360056225133
Epoch: 7, BCELoss: 0.35614595182087955
Epoch: 8, BCELoss: 0.3200960384339702
Epoch: 9, BCELoss: 0.2869105664442997
Epoch: 10, BCELoss: 0.2874879040280167
2019-12-03 10:15:48,075 sagemaker-containers INFO     Reporting training SUCCESS
```

```
2019-12-03 10:15:56 Uploading - Uploading generated training model
2019-12-03 10:15:56 Completed - Training job completed
Training seconds: 264
Billable seconds: 264
```

1.9 Step 5: Testing the model

As mentioned at the top of this notebook, we will be testing this model by first deploying it and then sending the testing data to the deployed endpoint. We will do this so that we can make sure that the deployed model is working correctly.

1.10 Step 6: Deploy the model for testing

Now that we have trained our model, we would like to test it to see how it performs. Currently our model takes input of the form `review_length, review[500]` where `review[500]` is a sequence of 500 integers which describe the words present in the review, encoded using `word_dict`. Fortunately for us, SageMaker provides built-in inference code for models with simple inputs such as this.

There is one thing that we need to provide, however, and that is a function which loads the saved model. This function must be called `model_fn()` and takes as its only parameter a path to the directory where the model artifacts are stored. This function must also be present in the python file which we specified as the entry point. In our case the model loading function has been provided and so no changes need to be made.

NOTE: When the built-in inference code is run it must import the `model_fn()` method from the `train.py` file. This is why the training code is wrapped in a main guard (ie, `if __name__ == '__main__':`)

Since we don't need to change anything in the code that was uploaded during training, we can simply deploy the current model as-is.

NOTE: When deploying a model you are asking SageMaker to launch an compute instance that will wait for data to be sent to it. As a result, this compute instance will continue to run until *you* shut it down. This is important to know since the cost of a deployed endpoint depends on how long it has been running for.

In other words **If you are no longer using a deployed endpoint, shut it down!**

TODO: Deploy the trained model.

```
In [87]: # TODO: Deploy the trained model
         predictor = estimator.deploy(initial_instance_count=1, instance_type='ml.p2.xlarge')
```

1.11 Step 7 - Use the model for testing

Once deployed, we can read in the test data and send it off to our deployed model to get some results. Once we collect all of the results we can determine how accurate our model is.

```
In [88]: test_X = pd.concat([pd.DataFrame(test_X_len), pd.DataFrame(test_X)], axis=1)
```

```
In [89]: # We split the data into chunks and send each chunk seperately, accumulating the result

         def predict(data, rows=512):
             split_array = np.array_split(data, int(data.shape[0] / float(rows) + 1))
             predictions = np.array([])
             for array in split_array:
                 predictions = np.append(predictions, predictor.predict(array))

             return predictions

In [95]: predictions_unrounded = predict(test_X.values)
         predictions = [round(num) for num in predictions_unrounded]

In [96]: from sklearn.metrics import accuracy_score
         accuracy_score(test_y, predictions)

Out[96]: 0.85256

In [97]: print(predictions_unrounded)

[0.25083616 0.01193954 0.00822546 ... 0.95973063 0.91134888 0.97261649]
```

Question: How does this model compare to the XGBoost model you created earlier? Why might these two models perform differently on this dataset? Which do *you* think is better for sentiment analysis?

Answer: XGBoost provided a good result when dealing with sentiment analysis but fails when we have temporal dependency. That is, the mood of the review depends on the general tone of the review itself. So if for example, the review goes negative on its first part but becomes positive on its later parts or anything with similar fluctuations, XGBoost fails. LSTM has the power to retain this long-term dependency, hence called Long-Term Short-Term Memory or LSTM.

Also, since we are dealing with a fairly large dataset assuming the dataset quality is acceptable, XGBoost may fail to generalize from these. Luckily, with LSTM, it performs well on large dataset, thus a much viable option to this use case.

But LSTM is not good when we do not have access to GPU-enabled environments. And since training data is large, it is slow compared to XGBoost. Also, LSTM depends on the data itself. As much as it is good at finding intricate patterns of data, it fails when the data has significant variance. In scenarios like this, XGBoost is a preferred option where it achieves acceptable result at a fraction of computational cost.

Bottomline, I prefer LSTM in this project. It achieves high accuracy in just 10 epochs. Observing BCELoss pattern, it can definitely improve on more epochs but I prefer to train it to just 10.

1.11.1 (TODO) More testing

We now have a trained model which has been deployed and which we can send processed reviews to and which returns the predicted sentiment. However, ultimately we would like to be able to send our model an unprocessed review. That is, we would like to send the review itself as a string. For example, suppose we wish to send the following review to our model.

18

```
In [98]: test_review = 'The simplest pleasures in life are the best, and this film is one of the
```

The question we now need to answer is, how do we send this review to our model?

Recall in the first section of this notebook we did a bunch of data processing to the IMDb dataset. In particular, we did two specific things to the provided reviews. - Removed any html tags and stemmed the input - Encoded the review as a sequence of integers using `word_dict`

In order process the review we will need to repeat these two steps.

TODO: Using the `review_to_words` and `convert_and_pad` methods from section one, convert `test_review` into a numpy array `test_data` suitable to send to our model. Remember that our model expects input of the form `review_length, review[500]`.

```
In [139]: # TODO: Convert test_review into a form usable by the model and save the results in te
         converted, leng = convert_and_pad(word_dict, review_to_words(test_review), 500)
         test_data = converted
         test_data.insert(0, leng)
         test_data = np.array([test_data])
```

```
In [143]: print(test_data.shape)
```

```
(1, 501)
```

Now that we have processed the review, we can send the resulting array to our model to predict the sentiment of the review.

```
In [144]: predictor.predict(test_data)
```

```
Out[144]: array(0.81779, dtype=float32)
```

Since the return value of our model is close to 1, we can be certain that the review we submitted is positive.

1.11.2 Delete the endpoint

Of course, just like in the XGBoost notebook, once we've deployed an endpoint it continues to run until we tell it to shut down. Since we are done using our endpoint for now, we can delete it.

```
In [155]: estimator.delete_endpoint()
```

1.12 Step 6 (again) - Deploy the model for the web app

Now that we know that our model is working, it's time to create some custom inference code so that we can send the model a review which has not been processed and have it determine the sentiment of the review.

As we saw above, by default the estimator which we created, when deployed, will use the entry script and directory which we provided when creating the model. However, since we now wish to accept a string as input and our model expects a processed review, we need to write some custom inference code.

We will store the code that we write in the `serve` directory. Provided in this directory is the `model.py` file that we used to construct our model, a `utils.py` file which contains the

`review_to_words` and `convert_and_pad` pre-processing functions which we used during the initial data processing, and `predict.py`, the file which will contain our custom inference code. Note also that `requirements.txt` is present which will tell SageMaker what Python libraries are required by our custom inference code.

When deploying a PyTorch model in SageMaker, you are expected to provide four functions which the SageMaker inference container will use. - `model_fn`: This function is the same function that we used in the training script and it tells SageMaker how to load our model. - `input_fn`: This function receives the raw serialized input that has been sent to the model's endpoint and its job is to de-serialize and make the input available for the inference code. - `output_fn`: This function takes the output of the inference code and its job is to serialize this output and return it to the caller of the model's endpoint. - `predict_fn`: The heart of the inference script, this is where the actual prediction is done and is the function which you will need to complete.

For the simple website that we are constructing during this project, the `input_fn` and `output_fn` methods are relatively straightforward. We only require being able to accept a string as input and we expect to return a single value as output. You might imagine though that in a more complex application the input or output may be image data or some other binary data which would require some effort to serialize.

1.12.1 (TODO) Writing inference code

Before writing our custom inference code, we will begin by taking a look at the code which has been provided.

```
In [164]: !pygmentize serve/predict.py

import argparse
import json
import os
import pickle
import sys
import sagemaker_containers
import pandas as pd
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
import torch.utils.data

from model import LSTMClassifier

from utils import review_to_words, convert_and_pad

def model_fn(model_dir):
    """Load the PyTorch model from the `model_dir` directory."""
    print("Loading model.")

    # First, load the parameters used to create the model.
    model_info = {}
```

```python
        model_info_path = os.path.join(model_dir, 'model_info.pth')
        with open(model_info_path, 'rb') as f:
            model_info = torch.load(f)

        print("model_info: {}".format(model_info))

        # Determine the device and construct the model.
        device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
        model = LSTMClassifier(model_info['embedding_dim'], model_info['hidden_dim'], model_info['vo

        # Load the store model parameters.
        model_path = os.path.join(model_dir, 'model.pth')
        with open(model_path, 'rb') as f:
            model.load_state_dict(torch.load(f))

        # Load the saved word_dict.
        word_dict_path = os.path.join(model_dir, 'word_dict.pkl')
        with open(word_dict_path, 'rb') as f:
            model.word_dict = pickle.load(f)

        model.to(device).eval()

        print("Done loading model.")
        return model

def input_fn(serialized_input_data, content_type):
    print('Deserializing the input data.')
    if content_type == 'text/plain':
        data = serialized_input_data.decode('utf-8')
        return data
    raise Exception('Requested unsupported ContentType in content_type: ' + content_type)

def output_fn(prediction_output, accept):
    print('Serializing the generated output.')
    return str(prediction_output)

def predict_fn(input_data, model):
    print('Inferring sentiment of input data.')

    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

    if model.word_dict is None:
        raise Exception('Model has not been loaded properly, no word_dict.')

    # TODO: Process input_data so that it is ready to be sent to our model.
    #       You should produce two variables:
    #         data_X   - A sequence of length 500 which represents the converted review
    #         data_len - The length of the review
```

```
        converted, leng = convert_and_pad(model.word_dict, review_to_words(input_data), 500)

        data_X = converted
        data_len = leng

        # Using data_X and data_len we construct an appropriate input tensor. Remember
        # that our model expects input data of the form 'len, review[500]'.
        data_pack = np.hstack((data_len, data_X))
        data_pack = data_pack.reshape(1, -1)

        data = torch.from_numpy(data_pack)
        data = data.to(device)

        # Make sure to put the model into evaluation mode
        model.eval()

        # TODO: Compute the result of applying the model to the input data. The variable `result` sh
        #       be a numpy array which contains a single integer which is either 1 or 0

        with torch.no_grad():
            output = model.forward(data)

        result = int(torch.round(output.squeeze()))

        return result
```

As mentioned earlier, the `model_fn` method is the same as the one provided in the training code and the `input_fn` and `output_fn` methods are very simple and your task will be to complete the `predict_fn` method. Make sure that you save the completed file as `predict.py` in the `serve` directory.

TODO: Complete the `predict_fn()` method in the `serve/predict.py` file.

1.12.2   Deploying the model

Now that the custom inference code has been written, we will create and deploy our model. To begin with, we need to construct a new PyTorchModel object which points to the model artifacts created during training and also points to the inference code that we wish to use. Then we can call the deploy method to launch the deployment container.

NOTE: The default behaviour for a deployed PyTorch model is to assume that any input passed to the predictor is a `numpy` array. In our case we want to send a string so we need to construct a simple wrapper around the `RealTimePredictor` class to accomodate simple strings. In a more complicated situation you may want to provide a serialization object, for example if you wanted to sent image data.

```
In [170]: from sagemaker.predictor import RealTimePredictor
          from sagemaker.pytorch import PyTorchModel
```

22

```
class StringPredictor(RealTimePredictor):
    def __init__(self, endpoint_name, sagemaker_session):
        super(StringPredictor, self).__init__(endpoint_name, sagemaker_session, conten

model = PyTorchModel(model_data=estimator.model_data,
                     role = role,
                     framework_version='0.4.0',
                     entry_point='predict.py',
                     source_dir='serve',
                     predictor_cls=StringPredictor)
predictor = model.deploy(initial_instance_count=1, instance_type='ml.p2.xlarge')
```

1.12.3 Testing the model

Now that we have deployed our model with the custom inference code, we should test to see if everything is working. Here we test our model by loading the first 250 positive and negative reviews and send them to the endpoint, then collect the results. The reason for only sending some of the data is that the amount of time it takes for our model to process the input and then perform inference is quite long and so testing the entire data set would be prohibitive.

```
In [171]: import glob

def test_reviews(data_dir='../data/aclImdb', stop=250):

    results = []
    ground = []

    # We make sure to test both positive and negative reviews
    for sentiment in ['pos', 'neg']:

        path = os.path.join(data_dir, 'test', sentiment, '*.txt')
        files = glob.glob(path)

        files_read = 0

        print('Starting ', sentiment, ' files')

        # Iterate through the files and send them to the predictor
        for f in files:
            with open(f) as review:
                # First, we store the ground truth (was the review positive or negativ
                if sentiment == 'pos':
                    ground.append(1)
                else:
                    ground.append(0)
```

```
                    # Read in the review and convert to 'utf-8' for transmission via HTTP
                    review_input = review.read().encode('utf-8')
                    # Send the review to the predictor and store the results
                    results.append(int(predictor.predict(review_input)))

                    # Sending reviews to our endpoint one at a time takes a while so we
                    # only send a small number of reviews
                    files_read += 1
                    if files_read == stop:
                        break

            return ground, results
```

```
In [172]: ground, results = test_reviews()

Starting  pos  files
Starting  neg  files
```

```
In [173]: from sklearn.metrics import accuracy_score
          accuracy_score(ground, results)
```

```
Out[173]: 0.864
```

As an additional test, we can try sending the `test_review` that we looked at earlier.

```
In [174]: predictor.predict(test_review)
```

```
Out[174]: b'1'
```

Now that we know our endpoint is working as expected, we can set up the web page that will interact with it. If you don't have time to finish the project now, make sure to skip down to the end of this notebook and shut down your endpoint. You can deploy it again when you come back.

1.13   Step 7 (again): Use the model for the web app

> **TODO:** This entire section and the next contain tasks for you to complete, mostly using the AWS console.

So far we have been accessing our model endpoint by constructing a predictor object which uses the endpoint and then just using the predictor object to perform inference. What if we wanted to create a web app which accessed our model? The way things are set up currently makes that not possible since in order to access a SageMaker endpoint the app would first have to authenticate with AWS using an IAM role which included access to SageMaker endpoints. However, there is an easier way! We just need to use some additional AWS services.

The diagram above gives an overview of how the various services will work together. On the far right is the model which we trained above and which is deployed using SageMaker. On the far left is our web app that collects a user's movie review, sends it off and expects a positive or negative sentiment in return.

In the middle is where some of the magic happens. We will construct a Lambda function, which you can think of as a straightforward Python function that can be executed whenever a specified event occurs. We will give this function permission to send and recieve data from a SageMaker endpoint.

Lastly, the method we will use to execute the Lambda function is a new endpoint that we will create using API Gateway. This endpoint will be a url that listens for data to be sent to it. Once it gets some data it will pass that data on to the Lambda function and then return whatever the Lambda function returns. Essentially it will act as an interface that lets our web app communicate with the Lambda function.

1.13.1   Setting up a Lambda function

The first thing we are going to do is set up a Lambda function. This Lambda function will be executed whenever our public API has data sent to it. When it is executed it will receive the data, perform any sort of processing that is required, send the data (the review) to the SageMaker endpoint we've created and then return the result.

Part A: Create an IAM Role for the Lambda function   Since we want the Lambda function to call a SageMaker endpoint, we need to make sure that it has permission to do so. To do this, we will construct a role that we can later give the Lambda function.

Using the AWS Console, navigate to the **IAM** page and click on **Roles**. Then, click on **Create role**. Make sure that the **AWS service** is the type of trusted entity selected and choose **Lambda** as the service that will use this role, then click **Next: Permissions**.

In the search box type `sagemaker` and select the check box next to the **AmazonSageMakerFullAccess** policy. Then, click on **Next: Review**.

Lastly, give this role a name. Make sure you use a name that you will remember later on, for example `LambdaSageMakerRole`. Then, click on **Create role**.

Part B: Create a Lambda function   Now it is time to actually create the Lambda function.

Using the AWS Console, navigate to the AWS Lambda page and click on **Create a function**. When you get to the next page, make sure that **Author from scratch** is selected. Now, name your Lambda function, using a name that you will remember later on, for example `sentiment_analysis_func`. Make sure that the **Python 3.6** runtime is selected and then choose the role that you created in the previous part. Then, click on **Create Function**.

On the next page you will see some information about the Lambda function you've just created. If you scroll down you should see an editor in which you can write the code that will be executed when your Lambda function is triggered. In our example, we will use the code below.

```python
# We need to use the low-level library to interact with SageMaker since the SageMaker API
# is not available natively through Lambda.
import boto3

def lambda_handler(event, context):

    # The SageMaker runtime is what allows us to invoke the endpoint that we've created.
    runtime = boto3.Session().client('sagemaker-runtime')

    # Now we use the SageMaker runtime to invoke our endpoint, sending the review we were given
```

```
    response = runtime.invoke_endpoint(EndpointName = '**ENDPOINT NAME HERE**',   # The name of
                                       ContentType = 'text/plain',                # The data fo
                                       Body = event['body'])                      # The actual

    # The response is an HTTP response whose body contains the result of our inference
    result = response['Body'].read().decode('utf-8')

    return {
        'statusCode' : 200,
        'headers' : { 'Content-Type' : 'text/plain', 'Access-Control-Allow-Origin' : '*' },
        'body' : result
    }
```

Once you have copy and pasted the code above into the Lambda code editor, replace the **ENDPOINT NAME HERE** portion with the name of the endpoint that we deployed earlier. You can determine the name of the endpoint using the code cell below.

`In [175]:` `predictor.endpoint`

`Out[175]:` `'sagemaker-pytorch-2019-12-03-13-38-05-406'`

Once you have added the endpoint name to the Lambda function, click on **Save**. Your Lambda function is now up and running. Next we need to create a way for our web app to execute the Lambda function.

1.13.2   Setting up API Gateway

Now that our Lambda function is set up, it is time to create a new API using API Gateway that will trigger the Lambda function we have just created.

Using AWS Console, navigate to **Amazon API Gateway** and then click on **Get started**.

On the next page, make sure that **New API** is selected and give the new api a name, for example, `sentiment_analysis_api`. Then, click on **Create API**.

Now we have created an API, however it doesn't currently do anything. What we want it to do is to trigger the Lambda function that we created earlier.

Select the **Actions** dropdown menu and click **Create Method**. A new blank method will be created, select its dropdown menu and select **POST**, then click on the check mark beside it.

For the integration point, make sure that **Lambda Function** is selected and click on the **Use Lambda Proxy integration**. This option makes sure that the data that is sent to the API is then sent directly to the Lambda function with no processing. It also means that the return value must be a proper response object as it will also not be processed by API Gateway.

Type the name of the Lambda function you created earlier into the **Lambda Function** text entry box and then click on **Save**. Click on **OK** in the pop-up box that then appears, giving permission to API Gateway to invoke the Lambda function you created.

The last step in creating the API Gateway is to select the **Actions** dropdown and click on **Deploy API**. You will need to create a new Deployment stage and name it anything you like, for example `prod`.

You have now successfully set up a public API to access your SageMaker model. Make sure to copy or write down the URL provided to invoke your newly created public API as this will be needed in the next step. This URL can be found at the top of the page, highlighted in blue next to the text **Invoke URL**.