Introduction - TensorFlow and application in NLP

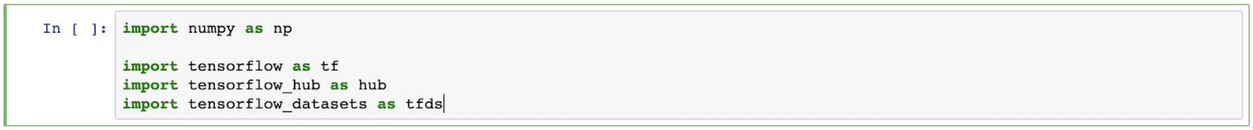
In this document, we are going to be discussing TensorFlow and its applications in Natural Language Processing techniques. Before we dive deeper into the application of how the toolkit is used, it is in our best interest to have a higher level understanding of TensorFlow and what it is. For starters, TensorFlow is an open source library that is primarily used in machine learning and has a focus on training deep neural networks. It was developed by the Google Brain team for all kinds of internal machine learning use. Later, it has gained much popularity in recent years and is being used for all kinds of NLP use as well. One specific one that we are going to be walking through and learning more about is Text Classification using Movie Reviews. We will also learn about other applications of TensorFlow and how it is being used in industry context.

Body - IMDb dataset Tutorial and Application

Using TensorFlow for Text Clustering of movie Reviews

In this class, we deal and talk about various different NLP techniques in regard to Text Retrieval and Text Mining. In the latter part of the course, we have been diving deeper into Text Mining techniques and we learned about Text Clustering. Text Clustering is a way to text mine where we try to group similar objects together in order to discover their “natural structure.” These objects of similarity can be documents, terms, passages or even websites. Examples of text clustering include clustering documents in a whole collection, clustering of websites or even term clustering defined by theme or topic. It can be useful for getting a sense of the overall content of a collection and get an understanding of the sentiment behind the text.

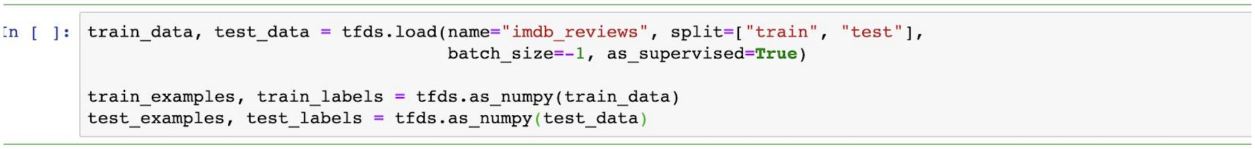
TensorFlow is one tool kit that can be used in text clustering applications and proves to be helpful in really doing semantic analysis on a piece of text. In this example, we are going to talk about how TensorFlow is used in order to perform text clustering techniques on movie reviews from the IMDb database in order to classify them as being positive or negative. This would give us an overall better understanding of how people are viewing/rating different movies and what aspect of those movies do the people like. Let’s take this case study and see some sample code on how we would use TensorFlow in order to work on text clustering. The first step is to import in tensorflow and split the test and training data up in the built in IMDb database. (Note. not all bits of code are shown)



```
In [ ]: import numpy as np

        import tensorflow as tf
        import tensorflow_hub as hub
        import tensorflow_datasets as tfds
```

Figure 1. Importing in TensorFlow

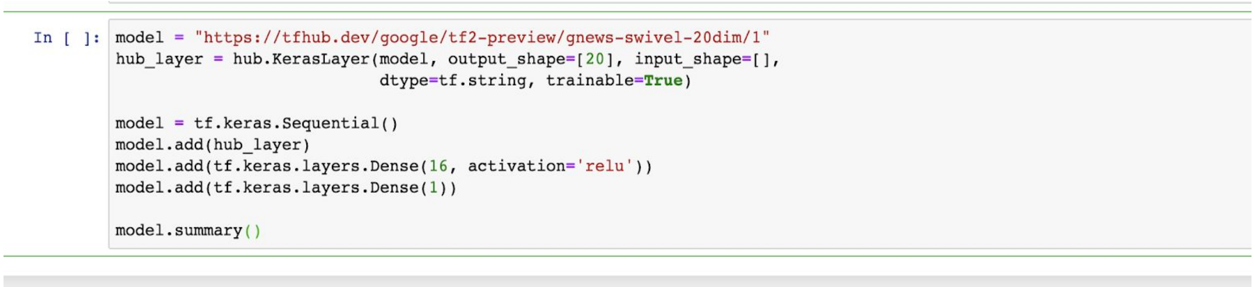


```
n [ ]: train_data, test_data = tfds.load(name="imdb_reviews", split=["train", "test"],
                                         batch_size=-1, as_supervised=True)

       train_examples, train_labels = tfds.as_numpy(train_data)
       test_examples, test_labels = tfds.as_numpy(test_data)
```

Figure 2. Training and Testing the data

After splitting our data into both testing and training, we would want to build a model for a neural network that works towards clustering our dataset. The neural network is made through stacking layers and that is and the labels are going to predict either a 1 or 0 for every review sentence. By converting sentences to embedding vectors, we can pre-train as a part of the first layer. For this example we can create a Keras layer that uses a built in TensorFlow model to embed the sentences. A layer in tensorflow can be thought of as just an object that takes in inputs and outputs a tensor through some computation. In order our sentences in the reviews to vectors, we are using a specific layer called the Keras layer and will proceed to build a full model.



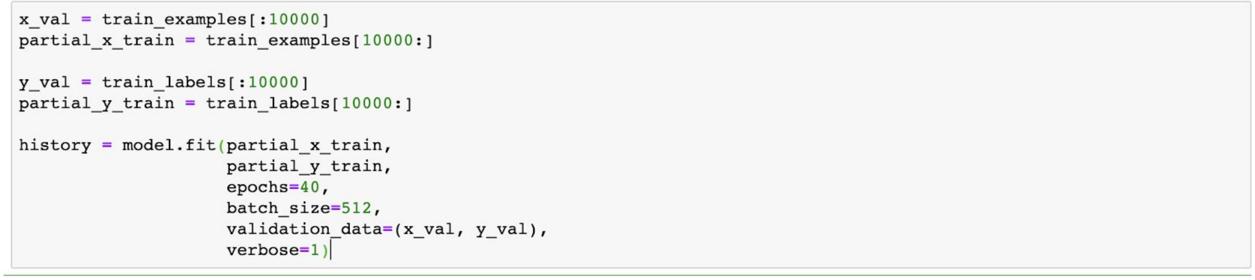
```
In [ ]: model = "https://tfhub.dev/google/tf2-preview/gnews-swivel-20dim/1"
        hub_layer = hub.KerasLayer(model, output_shape=[20], input_shape=[],
                                   dtype=tf.string, trainable=True)

        model = tf.keras.Sequential()
        model.add(hub_layer)
        model.add(tf.keras.layers.Dense(16, activation='relu'))
        model.add(tf.keras.layers.Dense(1))

        model.summary()
```

Figure 3. Building the full model and using the Keras layer

In the above figure, we have added layers to our model. The first layer is the Keras layer to map a sentence into its embedding vector. The second layer is the Dense layer where the output vector goes through with 16 hidden units. The hidden units can be thought of as the amount of freedom that the network is allowed when learning an internal representation.



```
x_val = train_examples[:10000]
partial_x_train = train_examples[10000:]

y_val = train_labels[:10000]
partial_y_train = train_labels[10000:]

history = model.fit(partial_x_train,
                    partial_y_train,
                    epochs=40,
                    batch_size=512,
                    validation_data=(x_val, y_val),
                    verbose=1)
```

Figure 4. Training the model and fitting it

Finally, we are going to create a validation set and train the model. After training the model, we will have to evaluate it to check for the accuracy against the training set. Overall, this process lets us see which reviews are positive and negative and builds an effective model in order to do text clustering. This is one way in which NLP techniques can be used for ways to classify text data through the TensorFlow library. This is just one of the applications of TensorFlow. We can see how TensorFlow is being used in various cases within the industry.

Applications in Industry

There are many applications of TensorFlow currently being used in the industry. One interesting application that relates closely to text clustering such as the technique as we saw above when we classified IMDb reviews as being positive and negative is picture classification as done by Airbnb. One case study that was done was how Airbnb is using TensorFlow and machine learning to help categorize its listing photos. Using this picture classification model that they have built, Airbnb is able to tag pictures of homes and identify the room that a specific picture belongs to, tag it and return that back to the user. Similar to our example, Airbnb used ResNet50, Keras and TF serving to build their layer.

Another example of an application of Text classification used in the industry was done by Twitter. They had used it in order to rank tweets that were showing up on the home timeline. For the ranking, the candidate Tweet was scored by a relevance model in order to see how useful it would be for each user. Relevance was then judged by seeing how the user would engage with the tweet and using that feedback, was used to improve the existing relevance feedback model. This is very similar to the recommender push system formations that we had talked about in the class.

Conclusion

In this technology review, we had discussed TensorFlow and how it is a software library used for various machine learning purposes. In this specific review, we took a certain scenario of creating text clusters for movie reviews using IMDb database. After creating a model and evaluating the accuracy of our model for predicting reviews to be positive or negative, we took a look at the applications of TensorFlow used in the industry by companies like Airbnb and Twitter in the field of NLP and clustering. It is safe to say that TensorFlow is a useful machine learning tool kit that can be used for a wide array of NLP applications.