201900 MUHAMMAD HASSAN MUKHTAR

BSAI Fall 2020

Data Selected

This is a dataset of 25,000 movie reviews from IMDB, labeled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a list of word indexes (integers). For convenience, words are indexed by overall frequency in the dataset, so that for instance the integer "3" encodes the 3rd most frequent word in the data. This allows for quick filtering operations such as: "only consider the top 10,000 most common words, but eliminate the top 20 most common words".

The files provided include "**imdb.npz**" housing both training and testing data, and "**imdb_word_index.json**" acting as a dictionary for reverting the labeling process. In this labeling approach, each word is assigned a distinct numerical identifier. Additionally, the input is of variable length, with each sentence restricted to 80 words for consistency.

RNN Implementation

The TensorFlow library is employed to construct a neural network model for sentiment analysis. The model architecture is built sequentially, beginning with an Embedding layer followed by an **LSTM** layer and concluding with a Dense layer.

Firstly, an Embedding layer is added to the model with an input dimension of 20,000 and an output dimension of 128. This layer serves to convert the input word indices into dense vectors of fixed size, facilitating the representation of textual data in a continuous vector space.

Following the Embedding layer, an LSTM (Long Short-Term Memory) layer is incorporated into the model. The LSTM layer comprises 128 units, with dropout applied to both input and recurrent connections at a rate of 0.2. Dropout is utilized as a regularization technique to prevent overfitting by randomly disabling a fraction of the units during training.

Finally, a Dense layer with a single neuron and a sigmoid activation function is added to the model. This layer produces the final output of the network, representing the predicted sentiment probability (positive or negative). The sigmoid activation function squashes the output values into the range [0, 1], allowing them to be interpreted as probabilities.

After constructing the model architecture, it is compiled using binary cross-entropy as the loss function, the Adam optimizer for gradient descent, and accuracy as the evaluation metric. This configuration sets the model up for training and evaluation on sentiment analysis tasks, where the goal is to classify textual data into positive or negative sentiment categories.

Model Architecture and Working

Embedding Layer:

- Converts integer indices representing words into dense vectors of fixed size.
- Learn dense representations of words, capturing semantic relationships.
- Helps the model understand the contextual meaning of words in the input sequence.

model.add(Embedding(20000, 128))

LSTM Layer:

- Processes sequential data, capturing temporal dependencies between words.
- Capable of learning long-term dependencies in sequential data.
- Dropout and recurrent dropout parameters prevent overfitting.
- Effectively captures contextual information and sequential patterns.

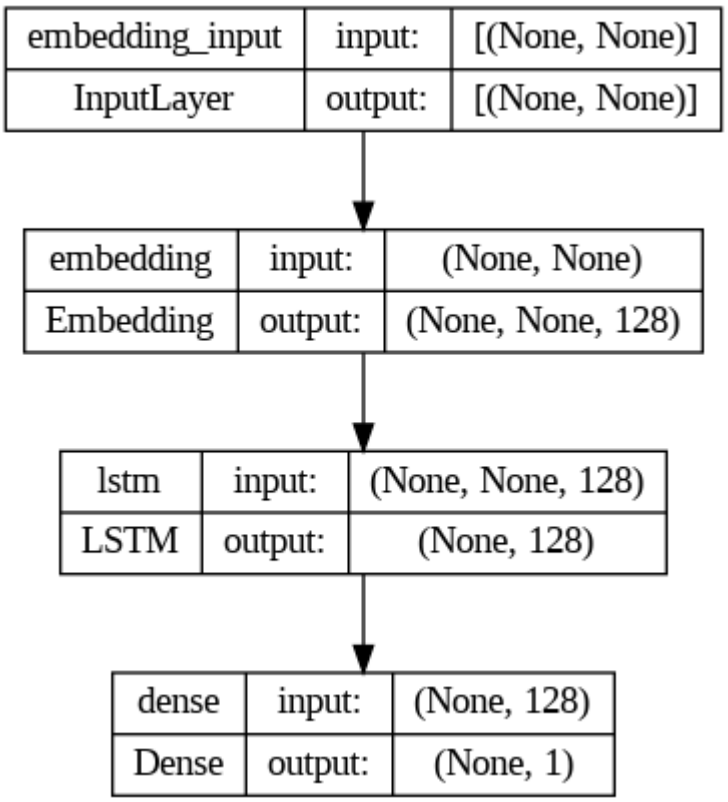
model.add(LSTM(128, dropout=0.2, recurrent_dropout=0.2))

Dense Layer:

- Standard fully connected layer that performs classification/regression.
- Applies a linear transformation followed by an activation function.
- Learns complex decision boundaries based on features learned by preceding layers.

- Outputs a probability score indicating likelihood of input sequence belonging to a class.

model.add(Dense(1, activation='sigmoid'))




```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 128)         2560000

 lstm (LSTM)                 (None, 128)               131584

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 2691713 (10.27 MB)
Trainable params: 2691713 (10.27 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Training and Evaluation

The training and testing datasets are imported, with a constraint to include only the top 20,000 most frequent words to manage the computational load. The dataset comprises 25,000 training reviews and 5,000 testing reviews.

The model is then trained using the training data (x_train and y_train) with the following specifications: a batch size of 32, training over 15 epochs, verbose output level set to 2, and validation data provided as (x_test, y_test).

Subsequently, the model's performance is evaluated on the testing data using the evaluate() function. The evaluation is conducted with a batch size of 32, and the results (score and accuracy) are printed out.

During training, the model achieved a loss of 1.0490 and an accuracy of approximately 80.96% after each epoch. This process took about 18 seconds per epoch, with an average step time of 23 milliseconds.

After evaluation, the test score is reported as 1.0490, and the test accuracy is approximately 80.96%. These metrics provide insight into the model's performance on unseen data, indicating its effectiveness in classifying sentiment in text reviews.

Experiment and Analysis

The model shows varying success in sentiment prediction, adept at identifying negative sentiments but struggling with positive ones, indicating a need for further training or adjustments to better understand nuanced language and context.

```
Example 1
Original Review: ? ? ? ? ? ? ? ? ? ? ? ? ? please give this one a miss br br kristy swanson and
Ground Truth Label: 0
Predicted Label: 0
1/1 [==============================] - 0s 36ms/step


Example 2
Original Review: wonderfully written script br br i praise robert altman this is one of his many
Ground Truth Label: 1
Predicted Label: 0
1/1 [==============================] - 0s 35ms/step


Example 3
Original Review: events ? may or may not have had ? turmoil in mind when he made ? but whatever
Ground Truth Label: 1
Predicted Label: 0
1/1 [==============================] - 0s 36ms/step
```