CA4023

Assignment Two

Semester Two, 2022

PART ONE (10 marks)

A baseline sentiment analysis system which uses logistic regression is shown in this notebook. This system obtains ~77% accuracy when trained on 80% of the training section of the IMDb movie review dataset, and tested on the remaining 20%.

Your goal is to attempt to improve the accuracy of the baseline model by trying out TWO ideas. Your new model should still use Logistic Regression.

Your experiments should be documented in a Jupyter notebook.

Marking Criteria

Marks will be awarded for

1. Two ideas for improving the results (4 marks)
2. Implementation of ideas (2 marks)
3. Description of experiments (3 marks)
4. Clear, readable code (1 mark)

PART TWO (10 marks)

BERT is a neural network language model architecture introduced by Google in 2018 (Devlin et al. 2018). When training a BERT model, the network is trained not to predict the next token in a sequence but to predict a masked token as in a cloze test, e.g.

Predict the MASK in this sequence

It was the best of times, it was the MASK of times

A secondary training objective is to predict whether two sequences follow on from each other, e.g.

Input Sequence 1 *My father's family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip. So, I called myself Pip, and came to be called Pip.*

Input Sequence 2 *I give Pirrip as my father's family name, on the authority of his tombstone and my sister — Mrs. Joe Gargery, who married the blacksmith. As I never saw my father or my mother, and never saw any likeness of either of them (for their days were long before the days of* *photographs), my first fancies regarding what they were like, were unreasonably derived from their tombstones.*

Output: TRUE

Sequence 1 *My father's family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip. So, I called myself Pip, and came to be called Pip.*

Sequence 2 *It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.*

Output: FALSE

BERT represents tokens as a subword units which can either be full words or subwords, e.g. *comparison, compar*, *##ison*.

BERT has become hugely popular because you can take a pretrained model, add a layer on top and fine-tune BERT for any NLP task (question answering, sentiment analysis, named entity recognition, and so on). During this fine-tuning process, the labelled data for the task is orders of magnitude lower than the amount of data that BERT is pre-trained on.

Several versions of BERT are available for English. There are versions available for many other languages too, plus a multilingual model trained on all languages for which Wikipedia is available.

Here is a Jupyter notebook which shows how BERT can be fine-tuned on a small subset of the same movie review dataset that was used in Part Two. It takes roughly 20 mins to train on Google Colab with one GPU. To run it using a GPU, select Edit -> Notebook Settings -> GPU.

What you have to do

The aim of this assignment is for you to gain experience in using BERT.

Choose one of the following:

1. **Analysis**
   Compare the output of the fine-tuned BERT model and your logistic regression model from Part Two. Try to answer the following questions

   - What are the strengths and weaknesses of each model?
   - What do both models have trouble with? Can you invent examples that "break" both models?

- What do both models do well?

2. **Review Text Selection**
   A disadvantage of BERT is there is a limit to the number of tokens that can be in the input. The number of tokens in the notebook model is restricted to the default value of 512. This means that the review text is truncated after 512 tokens.
   Attempt to improve the BERT model by using an intelligent approach to selecting the part of the review to include. But note that BERT expects fluent text rather than a bag of words.

3. **Fine-tune BERT on a different task/dataset.**
   Find a labelled dataset and fine-tune BERT on it. Here are some datasets that are available through HuggingFace

   https://github.com/huggingface/datasets/tree/master/datasets

   Caveat: don't choose a dataset which has already been used as an example in an online tutorial, e.g.
   https://huggingface.co/transformers/custom_datasets.html

4. **Something else BERT-related**
   Just run the idea by me first!

Marking Criteria

1. Task goal achieved (5 marks)
2. Description of experiments (5 marks)

ASSIGNMENT SUBMISSION

1. Via gitlab
2. A Jupyter notebook for each part.

ASSIGNMENT DEADLINE

Sat 2nd April, midnight