[COM4513-6513] Lab 5: Neural Language Modelling

Instructor: Nikos Aletras

Teaching Assistants: George Chrysostomou, Hardy and Zeerak Waseem

In this lab assignment, you will implement a neural language model.

N-Gram Neural Language Modelling

First, you need to download the code for N-Gram Language Modelling from here https://sheffieldnlp.github.io/com4513-6513/labs/word_embeddings_tutorial.py. Second, read the documentation for it here http://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html

- Run `word_embeddings_tutorial.py` on a Linux/Unix-based machine, i.e. Ubuntu or MacOS with PyTorch installed (see instruction in Lab 0). **Note 1:** You do not have to implement the CBOW model presented at the end of the file. [0 marks]
- Describe in your report the neural network language model using mathematical equations, fully detailing the dimensionsality of each parameter (**Hint:** it is a kind of multilayer perceptron). You could use a table to summarise the dimensionality of each layer of the network. [2 marks]
- Modify the code given in `word_embeddings_tutorial.py` to model the following toy training set (**Tip:** You need to create a list of sentences, where each sentence is represented as a list of tokens. You need to include start/end of sentence tokens):
  - The mathematician ran .
  - The mathematician ran to the store .
  - The physicist ran to the store .
  - The philosopher thought about it .
  - The mathematician solved the open problem .
- **Run a Sanity check:** make sure your model can learn how to predict correctly your training data. Take the sentence
  - The mathematician ran to the store .

  and check that for every trigram (i.e. context and prediction) you get the right answer. Does it work? You need to play with the hyper-parameters, such as learning rate, epoch number etc. You will observe some variance in the results, so find and report hyper-parameters that get the correct results in 5 consecutive runs. Among others, your model should be predicting for the context "START The" the word "mathematician". Why is this happening instead of predicting "physicist"? [3 marks]
- **Test:** Given a sentence with a gap
  - The _____ solved the open problem.

  which is more likely to fill it in: "physicist" or "philosopher"?

  Get the model to predict this correctly by changing the hyper-parameters (and report them). Discuss whether this would be possible with the bigram ML model from lab 2. Ensure that the model is predicting correctly for the right reason, i.e. that the embeddings for "physicist" and "mathematician" are closer together than the embeddings for "philosopher" and "mathematician". Use cosine similarity for this (it is implemented in PyTorch, check the API documentation). [3 marks]

Submission Details

Your code should be executable by running:

```
python3 lab5.py
```

Do not submit jupyter notebooks. You need to present both the results for the sanity check as well as the actual test. Your report `lab5.pdf`(2 pages max length) should be submitted as a PDF document.

This lab will be marked out of 8, 2 points for the model description part, 3 points for the sanity check and 3 for the test. Model description has only a report component (all of 2 points), the other two have both code and report (1.5 point each). It is worth 8% of your final grade in the module.

The deadline for this assignment is Monday 29/4, 11:00 and it needs to be submitted via MOLE. Standard departmental penalties for lateness will be applied.