Assignments 4 and 5

Submission date: End of day October 21, 2024

Assignment 4 (40 marks)

Build a skipgram model to generate the word embedding

Tasks

1. **Use the COVID-19 corpus**
   Extract all the abstracts from the COVID-19 text files and use them as the corpus. Ensure that you create a vocabulary of around 10,000 words. (5 marks)

2. **Construct One-Hot Vectors (OHVs)**
   You may construct each one on the fly (just in time) during training, or you can pre-create and store them in memory (volatile or non-volatile). If you associate numerical indexes with the vocabulary, you can generate each OHV on demand during training. (5 marks)

3. **Describe Your Architecture**
   Provide a brief description of the architecture used. Use appropriate markdown or formatting to emphasize key design decisions. Optionally, you can include a figure illustrating your neural network architecture, with explanations, instead of text. (5 marks)

4. **Use Stochastic Gradient Descent (SGD)**
   Apply the SGD algorithm for learning in your model. (5 marks)

5. **Plot Epochs vs. Training Error**
   Plot the relationship between the number of epochs and the training error. (10 marks)

6. **Use Negative Sampling**
   Replace naive softmax with negative sampling (use five negative samples per training instance)(5 marks)

7. **Test Analogies**
   Provide at least one example to test word analogies. The chosen words should not include the following terms: ”man,” ”woman,” ”boy,” ”girl,” ”country,” ”city,” ”malware,” ”virus,” and their synonyms. Ensure that two of the selected words used in the analogy are related to COVID-19. (5 marks)

Assignment 5 (25 marks)

Tasks

There are two matrices in this word2vec model

- W_{in} - connecting the input and hidden layer
- W_{out} - connecting the hidden and the output layer

Usually W_{in} is used for word embedding and W_{out} is ignored. In this assignment, you will execute the following:

1. Find similar words for a word of your choice using W_{in} (5 marks)
2. Find the similar words for the same word chosen in (1) using W_{out} (5 marks)
3. Find the similar words for the same word chosen in (1) after combining W_{in} and W_{out} - either concatenate them to have a longer vector or average them out (5 marks)
4. Compare the results of (1), (2) and (3) and write a brief description of the outcome (5 marks)
5. A slide related to the complexities of skipgram and CBOW model was shown in the class. Check if they are correct. If not, what are the correct entries? (5 marks)

Note

All fine prints from the earlier assignments apply here with respect to submission guidelines.