MLHVL Individual Assignment Part II

Please submit as a pdf via Blackboard before 23:59 2/11/2021, and please maintain the order of questions.

1. Before learning distributional word vectors from a text corpus, the texts in the corpus are usually preprocessed. In this preprocessing, a standard practice is to filter out the words that are (i) extremely common or (ii) extremely uncommon (i.e. the words with too few or too many global occurrences). Say why for each case, in a few sentences each. 2 points

2. In the original Skip-gram model, if v_j is the vector for the target word w_j and c_k is the vector for the context word w_k , the softmax is used to convert their dot product into probabilities.

   $$p(w_k|w_j) = \frac{e^{c_k \cdot v_j}}{\sum_{c_i \in V} e^{c_i \cdot v_j}}$$

   Here, the normalisation term in the denominator is expensive to compute (as for every word it has to be computed over the entire vocabulary V). The solution to this problem in Word2Vec is "negative sampling".

   Describe in a few sentences (4-6 sentences, and you don't need to use equations) how negative sampling works in the Skip-gram model. Use examples to illustrate your point, and make sure to say exactly what the negative samples are. 2 points

3. **Mitchell et al. 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns** : The authors compare the performance of their model in different settings, specifically, using different sets of "semantic features" — the words that are used for representing the target nouns. Which set of semantic features achieved the best results in their experiments? Why, do you think, that set worked better than any of the others? 2 points

   Hint: Do you think it's related to the authors' selection of the target words (the words they were trying to represent)?

4. **Corkery et al. 2019. Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection** : The authors train several instances of the model with different random initializations. How do these different model instances perform on the set

of real verbs (regular and irregular)? How do they perform on the set of nonce verbs? How do authors propose to treat the different instances of the model to account for their varying behavior? 3 points

5. **Vinyals et al. 2015. Show and Tell: A Neural Image Caption Generator** : In the Show and Tell paper, Figure 5 shows examples of the generated captions for the images in the test set. Describe at least four types of errors that you can observe there, illustrating your answer with examples. 2 points

6. Recall from Lab2 that in order for a neural network to work with the data, the data needs to be represented as a set of vectors, a tensor. When the input sequences in the data are of varying length (e.g. when the sentences in your training dataset are of varying length), we apply padding to create a well-formed tensor. Padding is appending zero tensors to all input sequences that are shorter than the maximum in-batch length to make them all equally long. What is the advantage to applying padding on the batch rather than the entire dataset? 1 point