Project 2 Report

CBOW Word embeddings with PyTorch

In this project, we are asked to create our own word-embedding using the Continuous-Bag-of-Word(CBOW) model with two corpuses(hotel reviews and and a scifi story). We have build and trained a CBOW2 model based on the hotel reviews dataset and the scifi text.

Part 1

1. Preprocessing

Hotel Reviews Dataset

After loading the data into colab, we clean the data. For each review, we substitute all numbers with spaces, make all words lowercase and delete all punctuation marks. We also remove stopwords and words which appear infrequently (because they usually contain spelling errors). After cleaning we get a list of reviews. Next, we delete all reviews that have less than 5 words, because to build a CBOW2 model, we need to have 4 content words (2 to the left and 2 to the right) for each of target word. We also eliminate words that occurs only 1 time, since they are likely to be misspelled.

Scifi Dataset

For the scifi dataset, the processing is very much the same. Firstly, all numbers and all punctional marks. Then we get the word list and the unique vocabularies list to form the word2list dictionary to map the words to numbers. We then eliminate words that occurs less than 100 times. In this relatively large corpus, we consider words occuring less than 100 times as uncommon words. They very much influence the training outcome.

2. Learning the word embeddings

Wtih the word list, we then built the vocabulary list with all unique words, with which we can later build the word2index and index2word dictionaries. These two dictionary will help us to map words two their index number for training the embedding. Next, for the hotel reviews dataset, for every review in the reviews list, we create context-target tuples. The tuple contains the target word itself and a list composed of its former 2 words and later 2 words. These tuples are later transformed into a tensor using the word2index lookup-table. Then we defined a CBOW models with two linear layers, one of which is followed by a ReLU activation function and the other of which is followed by a

LogSoftmax function. A number of 128 neurons is used in the hidden layer and the embeddings have dimension 50. The CBOW model is then trained for 30 epochs for the reviews dataset and 2 epochs for the scifi dataset.

We didn't do the optional CBOW5 part, because the processes to build a CBOW5 model is largely the same except building a longer content for each of the target. As for the performance result, as we discussed in the lecture, the CBOW5 model usually gets a more sementic result (since it uses a larger word context) and while the CBOW2 result is more syntactic.

Part 2

In Part 2, we are asked to find 3 nouns, 3 verbs and 3 adjectives from with differend word frequencies. So we create three group of words (very frequent words, medium frequent words, low frequent) based on their absolute number of occurence in the corpus. We then select one noun, one verb and one adjective from these three groups

-The 9 chosen words from the hotel review dataset and their neighbours are as follows:

frequency	selected words	5 closest neighbours
very frequent	hotel	'tuilieres', 'overload', 'bam', 'asun', 'visiting'
very frequent	great	'fantastic', 'excellent', 'superb', 'best', 'palmetto'
very frequent	clean	'beds', 'kalverstraat', 'bathrooms', 'victorian', 'spacious'
medium frequent	issue	'amzing', 'hawai', 'talk', 'greasy', 'like'
medium frequent	adequate	'small', 'tiananmen', 'single', 'expecting', 'wood'
medium frequent	smoking	'suitehotel', 'scenarios', 'legacy', 'caretakers', 'pillow'
low frequent	italy	'hotel', 'embargo', 'declare', 'crazy', 'hairstylist'
low frequent	filled	'stupid', 'hanson', 'fra', 'ankle', 'caren'
low frequent	comment	'hotel', 'bedding', 'intensive', 'walls', 'definitely'

We can see that most of the neighbouring words given by the model make sense, while some do not. For very frequent adjectives, such as “great”, the result(‘fantastic’, ‘excellent’, ‘superb’, ‘best’, ‘palmetto’) very good. The first 4 are synonyms and the last one(‘palmetto’) is like to be in a good hotel. While the result for the other less frequent words, such as “common” or other ambiguous word such as “clean”, “filled” (they can both be a verb or an adjective) are less satisfactory.

-The 9 chosen words from the scifi text and their neighbours are as follows:

frequency	selected words.	5 closest neighbours
very frequent	time	‘said’, ‘first’, ‘thought’, ‘made’, ‘like’
very frequent	think	‘know’, ‘tell’, ‘enough’, ‘want’, ‘sure’
very frequent	right	‘said’, ‘man’, ‘made’, ‘course’, ‘little’
medium frequent	blood	‘single’, ‘small’, ‘back’, ‘little’, ‘light’
medium frequent	smile	‘power’, ‘ship’, ‘back’, ‘right’, ‘said’
medium frequent	tiny	‘small’, ‘dark’, ‘turned’, ‘came’, ‘bright’
low frequent	party	‘course’, ‘way’, ‘point’, ‘man’, ‘men’
low frequent	worry	‘seems’, ‘however’, ‘right’, ‘girl’, ‘must’
low frequent	warm	‘watched’, ‘snapped’, ‘went’, ‘saw’, ‘looking’

As for the scifi dataset, the result is less satisfactory compared to the hotels reviews dataset. The difference in the result could be caused by the difference between novel and reviews. In reviews, the sentence and phrases are more repetitive and the the use cases are more explicit and fix.

We then chose two words [‘good’, ‘job’] to get the 5 cloest neighbours from both of the datasets and the result is as follows:

dataset	selected word	5 closest neighbours
hotel reviews	good	‘amandari’, ‘blooming’, ‘best’, ‘great’, ‘decent’
hotel reviews	job	‘nameless’, ‘experience’, ‘precinct’, ‘bouncy’, ‘whopping’
scifi	good	‘time’, ‘said’, ‘knew’, ‘like’, ‘first’
scifi	job	‘even’, ‘work’, ‘point’, ‘time’, ‘though’

Their five neigbouring words for both CBOW models are quite different, since in different use cases (hotel reviews and novel) words are used differently and the neighbouring words used with them are very different. For example in a hotel review, when a customer comments that a hotel is good, he or she may mean that the hotel is decent, so when we train a hotel model. The model put these vertors(‘good’,‘best’, ‘great’, ‘decent’) closer. However, “decent” is not a common word in the scifi dataset, saying ‘good’ in the scifi text may means the timing is good. So, the corpus of text you are using to train your CBOW model has very strong influence on the final word embeddings you are learning.