Assignment 3

Zhenhao Zhang

Question 1: Increase the skip window. Looking at the training error and the closest words, does the model seem to get better or worse? Explain why. (3 marks)
Answer: The initial value of skip window is 1. When it is increased to 10, the average loss seems get a little bit increase. The model seems to get worse. However, this parameter is the number of context words. The more skip window, the more paired data that is manyfold fed to the model. Then the model needs more calculation to get less average loss. But the practicability of the model will be better when it used to application scenarios. A paper says that selecting 2-5 as the skip window works well depending on the size of dataset.

Question 2: Research and explain NCE loss. (3 marks)
Answer: Noise-Contrastive Estimation, as known as NCE, is a method for fitting unnormalized model, adapted to neural language modelling. The purpose of using NCE is to eliminate the normalization costs during training. It enables fast training without the complexity of working with tree-structured models. The main advantage of NCE is that it makes the training time effectively independent of the vocabulary size.

Its principle is that the model uses the global unigram distribution of the training data as the noise distribution, which is called negative sampling. Its training time linear in the number of noise samples and independent of the vocabulary size. As the number of noise samples increases, the estimate approaches of likelihood gradient of the normalized model offer the best trade-off between computational and statistical efficiency.

Question 3: Why replace rare words with UNK rather than keeping them? (2 marks)
Answer: 'UNK' is used for the out-of-vocabulary issue. The 'UNK' tags can simply be used to tell the model that there is stuff, which is not semantically important to the output. The very low frequency words tend not to get very good word-vectors from their few training examples, and further given typical word-distributions (where there are many such 'long tail' few-occurrence words). There may be a lot of such words, which if retained tend to make the word-vectors for other words worse. With too few and insufficiently varied examples, they're essentially 'noise'. However, when lots of UNKs appear in the model, it definitely would have a negative effect on the performance.

Question 4: If you run the model more than once the t-SNE plot looks different each time. Why? (2 marks)
Answer: Because plot_only is 500, it shows the 1st - 500th words on the final_embeddings vector. And the sequence of words in final_embeddings vector is random.

Question 5: What happens to accuracy if you set vocabulary_size to 500? Explain why. (3 marks)
Answer: The vocabulary size will be the number of unique tokens seen in the training corpus. If set vocabulary_size to 500, the average loss will decrease faster than before. The main reason is the target words significantly decrease and the dimensions of the hidden layers also decrease a lot.

Question 6: You may see antonyms like “less” and “more” next to each other in the t-SNE. How does that make sense rather than them being at opposite ends of the plot? (2 marks)
Answer: This model is used to cluster words that are often in the same context together. The result of word2vec means semantically correlated, which is somewhat regarded as "relevance". The antonyms like "less" and "more" often appear near one another, for literary contrast or other emphasis so they are contextually quite relevant to one another. On the other hand, the points that are close in the high dimensions are easily shown close to each other in 2D space. However, the points that are far from each other in the high dimensions are possibly not shown far in 2D space.