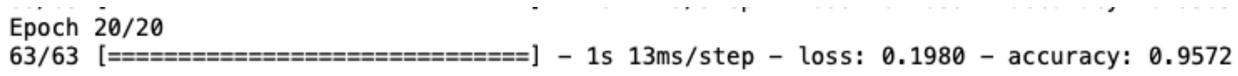
SCS 3546 Deep Learning Assignment 3 - Word2vec
https://www.tensorflow.org/tutorials/text/word2vec

Prepared by: Kerim Terzioglu

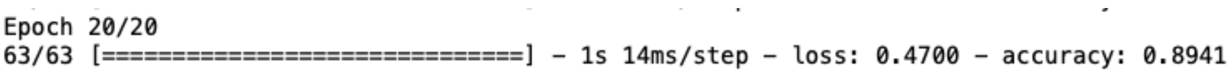
Question 1
Increase the skip window. Looking at the training error and the closest words, does the model seem to get better or worse? Explain why.

The initial value of the skip window is 2 which yields the following metrics after 20 epochs:



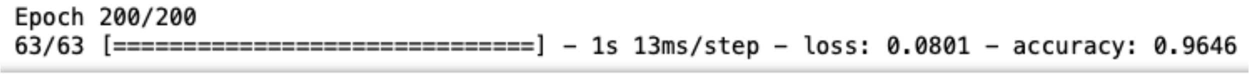
```
Epoch 20/20
63/63 [==============================] - 1s 13ms/step - loss: 0.1980 - accuracy: 0.9572
```

If the skip window is increased to a value of 10, the model yield the following metrics after 20 epochs:



```
Epoch 20/20
63/63 [==============================] - 1s 14ms/step - loss: 0.4700 - accuracy: 0.8941
```

As the skip window size is increased, the average loss increases (.1980 vs .4700) while the models accuracy decreases (.9572 vs .8941). This parameter is the number of context words. The larger the skip window, the more paired skip-grams are fed to the model. The training objective of the skip-gram model is to maximize the probability of predicting context words given the target word hence, the would model would require more calculations to get less average loss when increasing the skip window. Increasing the epochs to 200 demonstrates this:



```
Epoch 200/200
63/63 [==============================] - 1s 13ms/step - loss: 0.0801 - accuracy: 0.9646
```

Question 2
Research and explain NCE loss.

Noise Contrastive Estimation is a way of learning a data distribution by comparing it against a noise distribution, which we define. This allows us to cast an unsupervised problem as a supervised logistic regression problem.

The Noise Contrastive Estimation loss function is an efficient approximation for a full softmax. With an objective to learn word embeddings instead of modelling the word distribution, NCE loss can be simplified to use negative sampling.

Question 3
Why replace rare words with UNK rather than keeping them?

'UNK' represents an out-of-vocabulary issue. The very low frequency words tend not to get very good word-vectors from a low number of training examples. With too few and insufficiently varied examples, rare words are essentially ‘noise' and will impact the model’s performance. For example, in order to accurately translate a word in any particular context, the model needs to see as many examples as possible during the training stage. By definition the training data contains very few occurrences of rare words, so the model doesn't have enough information to learn their translation properly.

Question 4
If you run the model more than once the t-SNE plot looks different each time. Why?

t-SNE works by taking a group of high-dimensional vocabulary word feature vectors, then compresses them down to 2-dimensional x,y coordinate pairs. The idea is to keep similar words close together on the plane, while maximizing the distance between dissimilar words.

Unlike methods like PCA, t-SNE is non-convex, meaning it has multiple local minima and is therefore much more difficult to optimize, therefore t-SNE is non-deterministic. You can run it multiple times and get a different result each time

Question 5
What happens to accuracy if you set vocabulary_size to 500? Explain why.

The vocabulary size will be the number of unique tokens seen in the training corpus. If set 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?

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.