Note: The file paths in assignment 1 & 2 are defined as where they exist in my google drive. They may need changing.

The indentation is also off in places. This is because colab was throwing up strange errors about it and I had to put them out of whack to stop it.

Answers:

1. The vertical axis represents the vocabulary (each row represents a word). The horizontal axis represents the euclidean space and each column of a row represents a dimensional value in euclidean space. What this means is each row of the matrix is a vector associated with a word. In practice this allows us to represent the words in a euclidean space where we can assess their position relative to other words/vectors.



1. A feed forward network requires a weight for every dimension value of every word. As we increase the number of words in our vocabulary then the number of weights required increases exponentially. This is inefficient if we want to use large language libraries.  
   There is also the problem of inflexibility (the model's ability to interpret a word based on the context it is in within the sentence).   
   RNNs solve this by using previous words in a sentence as inputs for determining the next word. This means it can interpret the context of the whole sentence when predicting the next word, and can do so with relatively low computational complexity.
2. /6) The encoder first processes the french sentence given. Outputting some encoded representation of it. The decoder receives this along with the english vocabulary and decodes the encoded french sentence by associating each the sequence with the english euclidean space.

So when a word is received in french, its translated to english, and the decoder LSTM associates this (along with any previous words seen) with the most likely next word in english.

1. RNNs can have trouble with discerning attention (e.g. they can forget things that may have been important when given a large sequence). They are also not able to be parallelised because of their sequential nature.   
   Transformers are an architecture that uses an encoder and decoder but instead of recurrent networks making them up they use a feedforward architecture that analyses the importance of each word in relation to others. As a feedforward architecture that processes entire sentences at once, it is more efficient than an RNN and is much better at understanding attention.