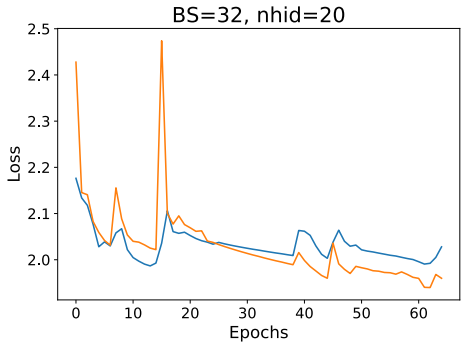
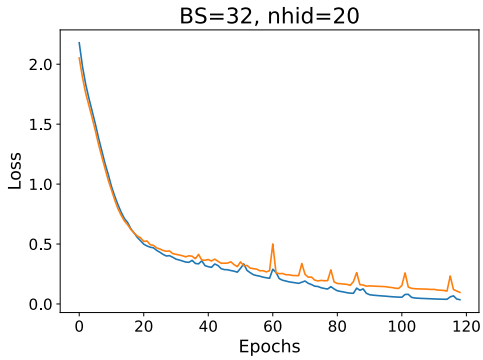
291-G assignment 2 report
Fatemehsadat Mireshghallah

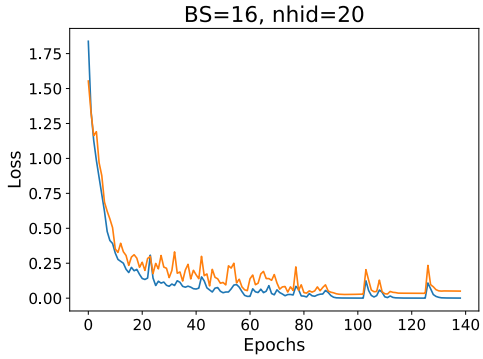
1. Hyper parameter tuning:
   I tried different batch size and hidden size it really did not make that much difference, they usually converged to a training loss of 0.001-0.008 by the 130th epoch (with learning rate fixed at 0.01). Increasing this lr, however, caused oscillations in the accuracy, for instance, you can see below the loss plot for lr of 0.1:



   And it was still not converging. Lr of 0.001 was also too slow:



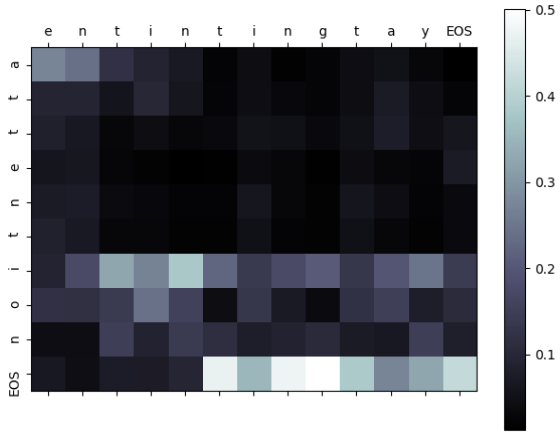
   At the end, I decided to go with BS of 16 and hidden size 20, with lr=0.01 which yields this plot (here the BS 16 and 32 didn't really make a difference):



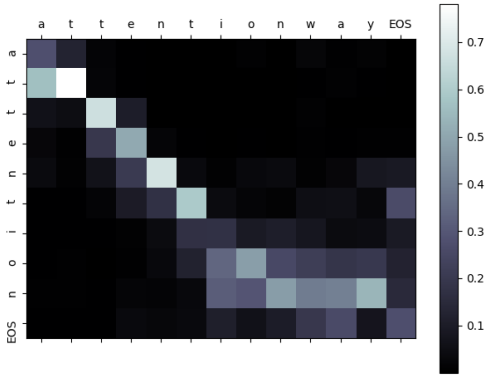
2. Trying different words and sentences
   For the 'I love deep learning' sentence, the sentence is translated to the correct 'iway ovelay eepday earninglay' sentence in around epoch 30, where the loss is 0.085.
   I also tried the sentence: attention is all you need which translated to "attentionway isway allway youway eednay" by the 38 th epoch. The sentence "attention brings more

insight to the model” is translated to “attentionway ingsbray oremay insightway otay ethay odelmay”

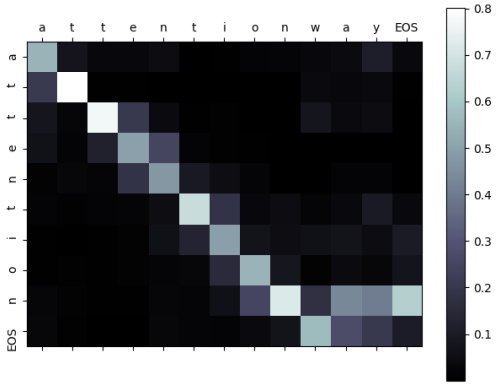
for the attention word, the starting maps are like this:



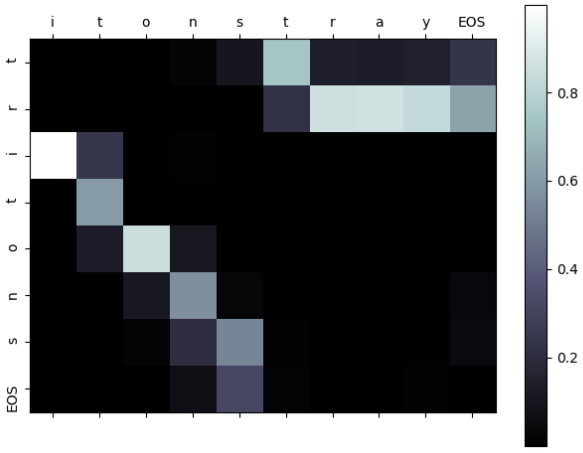
Which by epoch 10, turn into:



And by epoch 32, yields:

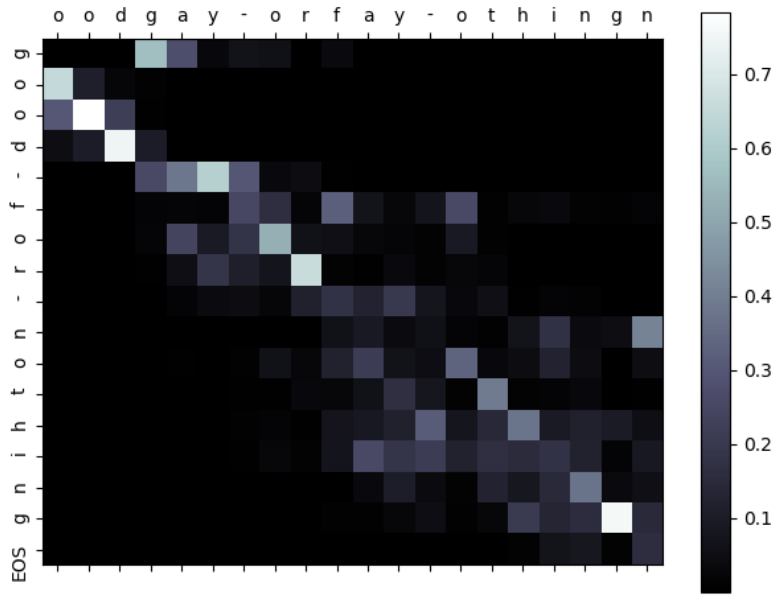


Which makes sense, since it coordinates the correct parts to each other. Now, attention starts with a vowel. If we try the word ‘tritons’, the map would look like this:



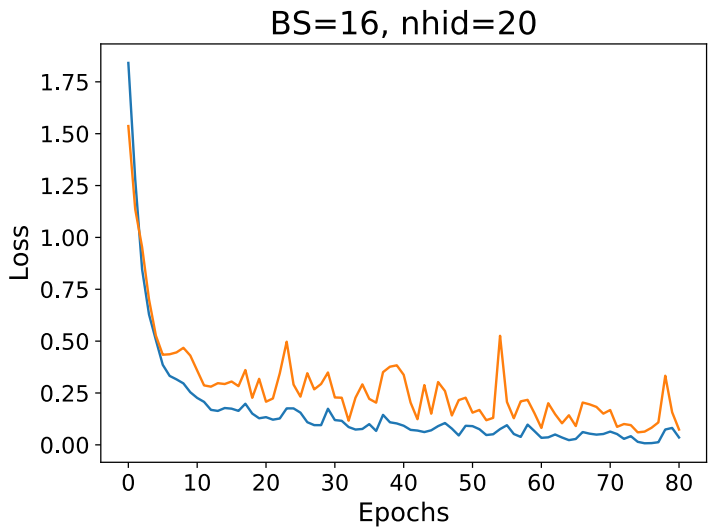
Which makes sense, since it corresponds the ‘itons’ part of the translated one to ‘ritons’ of the original word, and also the ‘tray’ part at the end of translated word to the ‘tr’ at the beginning of the original word. This also makes sense, because it is the first letter of the word that determines this.

For a word with dash ‘good-for-nothing’ the learning took a lot longer, around the 50th epoch, this is the map:

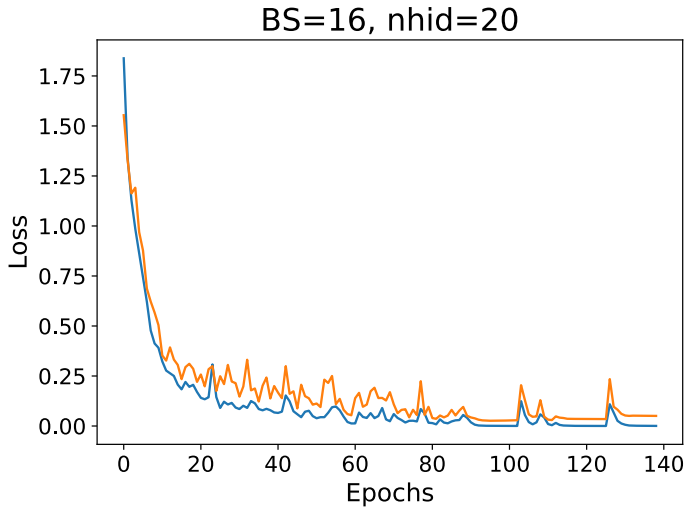


It can be seen that the words between dashes are found to be correlated, but yet, the diagonal line is not as straight as it is with non-compound words.

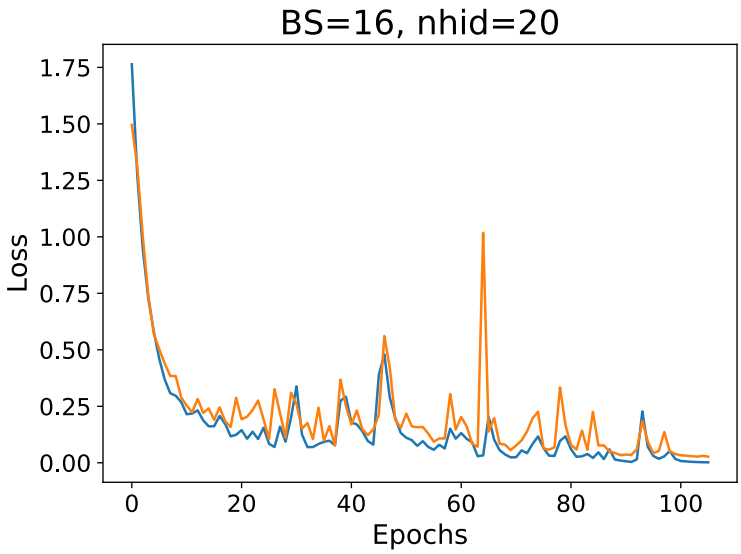
3. The f function:
I tried changing the activation function to leaky_relu and also tanh, results of which you can see below (leaky_relu and tanh respectively, lr=0.01):



The tanh one actually nearly zeroed out the loss:



I also tried making the model deeper by adding another hidden layer, with size hidden/2, which yielded this:



I didn’t really notice that much difference while changing the functions. The tanh seemed to be worse in the 60-70 epochs, but in the end, I felt it stabilized better than even ReLU. I expected the deeper model to be a lot better than the other two, but it was not.