Homework	Name: Mi Yu

Homework 4

Due Tuesday, Week 4 *Stanford 2018*

Question 1:

1(g): The enc_mask sets all the entry that represents 'pad' to 1 and non-pad to be 0, and then in our function, we assign all the 1 with negative infinity attention distribution score. Then, if we use this to calculate the soft max, we will have that the word 'pad' will be weighted towards 0 as $exp(-\infty) = 0$. We basically removed the effect of 'pad' token in our calculation for attention output. This makes sense because 'pad' tokens are originally not in the sentence, we are using these to fill the sentence such that all the sentences have the same length. Therefore, those should not have effects on our attention output, which will later affect the probability distribution over the target words.

1(i):

1(j):