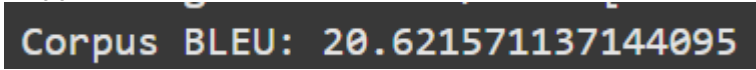# Assignment 4

Srihari Inukurthi - 110712807

November 29, 2024

Question G. The enc_masks tensor is used to mask the padded elements in the attention computation, so that the decoder does not attend to them. This is done by setting the values of e_t for the corresponding padded elements to -∞. The softmax operation on the e_t makes sure that the attention scores for the corresponding padded elements will tend to become 0 so that they will get ignored in the attention calculation.

Question H. The BLEU score is 20.621571137144095.
Proof:

Corpus BLEU: 20.621571137144095

Question I. (i)
The main advantage of the dot product attention is that it is computationally feasible and efficient because it requires only the dot product between the query and the key vectors.
The one disadvantage is that it will not allow the model to learn the different weights for the query and the key vectors, which makes it very difficult for the model to learn anamolies and complex patterns.

Question I. (ii) The advantage of using the additive attention is that it combines the encoder hidden state and the decoder hidden state through a feedforward neural network, allowing it to learn a more flexible and sophisticated scoring function. This can be beneficial for capturing complex relationships in the data.
The disadvantage is that it involves more operations due to the feedforward neural network, which can make it slower and more computationally expensive compared to multiplicative attention, especially for long sequences. Also, multiplicative attention is generally more efficient computationally, as it relies on matrix multiplications, which are highly optimized in modern deep learning frameworks. However, it might be less flexible in capturing nuanced relationships between sequences.