1 Question 1

The basic self-attention mechanism is a common approach to capture semantic representations of words. This consist on creating a single attention vector. Doing though, this method cannot handle tasks like sentiment classification, as there is not much information as input.

A first approach to handle this is to add a max (or average) pooling layer at all time step of the forward path. however, this could be hard to implement in a RNN structure.

Another good approach to improve basic self-attention mechanism is to modify his input; to have more dependency in the $2D$ matrix of embeddings, before going throught the attention mechanism, it may be useful to use a bidirectional LSTM that outputs H, a $2D$ matix of size $n \times 2u$ (instead of having a unidirectionnal approach wich outputs a vector fo size n). This produces a more relevent sentence representation.

Additionally, a common issue with basic self-attention is that it tends to focus on redundant information, where attention heads often concentrate on similar words or parts of the input. To mitigate this, a penalization term can be introduced to encourage diversity among attention heads and discourage redundancy. This penalization term, often implemented using the Frobenius norm, promotes the model to attend to different parts of the input, leading to better and more diverse sentence representations.

2 Question 2

The paper *Attention Is All You Need* introduced the Transformer model, which replaced recurrent operations (like those in RNNs and LSTMs) with self-attention mechanisms. The motivations for this change were leaded in addressing several limitations of recurrent models, particularly for tasks like machine translation and other sequence-based problems.

- from one time step to all time step (efficient in computing)
- motivation for parallel processing
- Long-Range Dependencies
- number of computational steps it takes for information from one token to influence another token. difficulty to compute gradients in backward path.
- more simple architecture

3 Bonus (*purpose of* `my_patience` *parameter*)

In the context of training a neural network, the `my_patience` variable refers to the patience parameter used in early stopping during training. Early stopping is a technique used to prevent overfitting and stop training when a model's performance on the validation set stops improving. It can prevent also from overfitting.

4 Question 3

For this question, I've decided to choose the last document of the test dataset. We can have a look at the `sentence attention coefficients`, and also to the `words attention coefficients`.

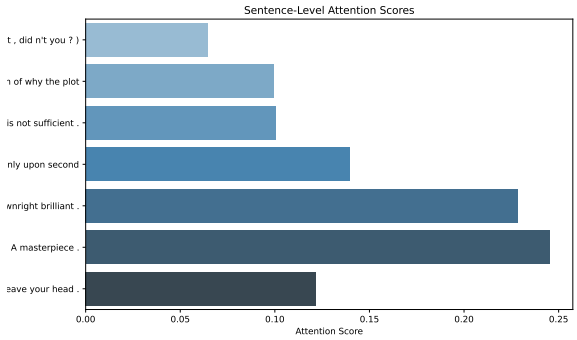

Figure 1: Sentence attention scores for the last document of the test dataset.

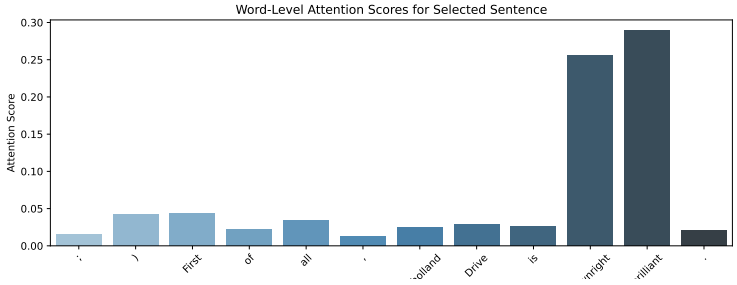

Figure 2: Word attention scores for the last sentence of the last document of the test dataset.

In figure 1, we can interpret that the 5^{th} and 6^{th} sentences are the most determinents in the choice of the model, helping it to classify if it's a good or bad review. In this case, the words *brillant* and *masterpiece* might be gamechanger in the classification.

[1]

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.