1 Question 1

- A possible problem with the self-attention mechanism is that nothing prevents the weights from being similar.

  As a consequence, [1] suggests to add a regularization term to prevent this effect: if A is the weight matrix, and $\|\cdot\|_F$ the Frobenius norm, then the authors consider the penalization $P = \|AA^T - I\|_F^2$.

  The idea is that if $AA^T = I$, then the different rows of A do not overlap (i.e., if the i-th coefficient of a row is non-zero, then the i-th coefficient of the other rows will be zero), hence diversity.

- Another possible improvement comes from the fact that in the basic self-attention mechanism, there is only one weight vector, which will usually focus on a single aspect of the sentence. Thus, for complicated sentences, the model may not be able to capture all relevant parts in the sentence.

  To solve this, the authors of [1] suggest to use multiple weight vectors, each being computed using a different attention distribution. The idea is that each of these weight vectors may focus on a different aspect of the sentence.

2 Question 2

- As explained in [3], the advantage of self-attention compared to recurrent operations is that it can easily be parallelized. Indeed, for recurrent operations, the computation of x_{t+1} requires to have already computed x_t, preventing efficient parallelization. For self-attention, all tokens in the sentence can attend to each other at the same time. As a consequence, self-attention based methods take much less time to train.

- Another advantage is the fact that RNNs struggle with learning long-range dependencies: in long sequences, the model can forget earlier information.

- Finally, as explained in the article, self-attention can offer more interpretability, by looking at the dependecies between tokens.

3 Question 3

In Fig. 1, we plot attention coefficients (inspired by Figure 3 in [1]).

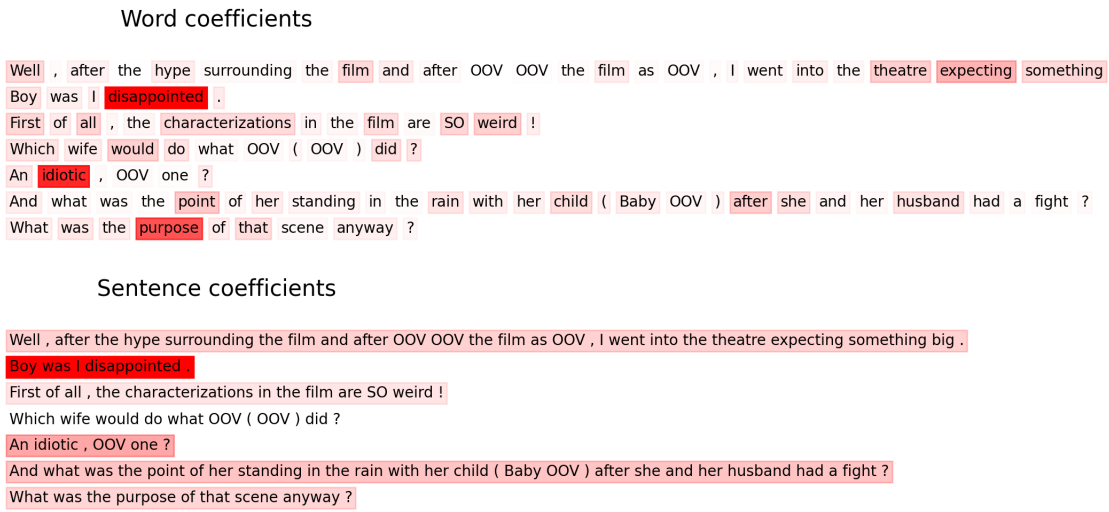

Figure 1: Coefficients for words and sentences

We see that ”disappointed” and ”idiotic” present rather high coefficients, which makes sense considering the fact that these words are negative, hence help the model understand the global sentiment of the review.
The word ”expecting” also receives a coefficient not too small, though obviously it can be used negatively (which is the case here) or positively, just like ”purpose”.

4 Question 4

A main limitation with HAN architecture, and pointed out in [2], is the fact that during the encoding process, each sentence is processed independently from the other sentences. As a consequence, it may be difficult for the model to understand the logic in a document made of several sentences.
[2] also points out the fact that if several sentences contain the same words with positive (or negative) value, then these words will be counted for each of the sentences they are in, and their effect will be multiplied. In practice however, it would be preferable to dampen this effect.

The authors of [2] suggest to use an architecture where an additional context vector is used, enabling interaction between sentences. Using a bidirectional encoder seems to yield even better results.

References

[1] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[2] Jean-Baptiste Remy, Antoine Jean-Pierre Tixier, and Michalis Vazirgiannis. Bidirectional context-aware hierarchical attention network for document understanding. *arXiv preprint arXiv:1908.06006*, 2019.

[3] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.