Rosalie MILLNER
rosalie@millner.fr

1 Question 1

How can the basic self-attention mechanism be improved? You may read [1] to find ideas.

In the paper 'A structured self-attentive sentence embedding' [1], it is explained that the interpretability of the basic self-attention mechanism is limited by the way it can only place its focus on one component of each sentence. For instance, it is difficult to interpret the semantics or classify sentences like *"The acting was not great, but i still really enjoyed the movie."* when one only looks at one part of the sentence (given that the beginning of the sentence sounds contradictory to the opinion it is actually conveying). The authors of the paper explain that the traditional self-attention mechanism can be improved by using multiple attention heads to prevent the model from focusing on the same positions repeatedly, which could help in better capturing the various contributions in meaning of words in a sentence. This problem is all the more significant when the sentences get longer. In the paper, this is done by transitioning from a vector to a matrix representation of the embedding.

2 Question 2

Read the transformer paper [2]. What are the main motivations for replacing recurrent operations with self-attention?

In the famous transformer paper, 'Attention is all you need' [2], the model introduces self-attention as a replacement for recurrent operations such as recurrent neural networks, for various reasons.
First, self-attention allows for significantly more parallelisation compared to recurrent operations. Indeed, there is an inherent constraint in recurrent operations which stems from its sequential nature. At each step t, the annotation h_t depends on the previous annotation h_{t-1} — so at every step, one needs to have computed the previous one beforehand (as is illustrated in the Lab 1 handout). These sequential computations make it very difficult to parallelise the operations during the training phase of the model. Furthermore, the longer the sequences, the bigger problem this is, as memory constraints limit batching. This enhanced parallelisation also helps in significantly reducing the training costs: as mentionned in the paper [2], "*The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs*".
It is also mentionned in the paper that using self-attention only, instead of recurrent operations, helps in handling long sequences (in the sense that it makes it easier to learn dependencies between distant positions), it improves representation learning, and does all of this with a better time complexity.

3 Question 3

Paste (or even better, plot) your attention coefficients for a document of your choice. Interpret your results.

In Figure 1 are plotted the attention coefficients for a document containing a review classified by the model as negative (probability of the review being positive = 1.9147e-06).

We can observe several words that have higher values for their attention coefficients. We can especially see that sentence 5 ("Now THAT was horrifically bad") has attention coefficients that are way higher than any other word in the other sentences. From an interpretative point of view, one could argue that this sentence is indeed the strongest one across the document in conveying the author's opinion. Other words, such as "bad" in sentence 1, also have relatively higher values for their attention coefficients, as for the words in the beginning of sentence 6 ("I could stand watching ten minutes"), for instance.

4 Question 4

What are some limitations of the HAN architecture? You may read [3] to find ideas.

According to the paper 'Bidirectional context-aware hierarchical attention network for document understanding' [3], a major weakness of the HAN architecture is that at level 1, each sentence is encoded independently, without considering the context of other sentences in the document. This means that when computing the representation of a sentence, the HAN architecture does not take into account the information from the surrounding sentences. This lack of communication can be considered to be suboptimal. Indeed, this can lead to the HAN architecture to spend most of its attention budget on redundant information from one sentence to another. The solution offered in the paper is to put in place a context-aware sentence encoder, which allows it to extract complementary information instead of repetitive ones, resulting in a richer document representation. The independent encoding of the sentences in level 1 can limit the HAN's ability in obtaining a deeper understanding of the relationships between sentences and thus extracting important aspects within a document.

5 Bonus question

What is the purpose of the parameter `my_patience`?

In the parameters section of the code, we set `my_patience` to the value 2. This is used for the early stopping strategy, which helps prevent overfitting. If the model's performance (in terms of validation accuracy) doesn't improve for 2 consecutive epochs, training will stop early, in order to save time and resources.

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. 2017.

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

[3] Jean-Baptiste Remy, Antoine Jean-Pierre Tixier, and Michalis Vazirgiannis. Bidirectional context-aware hierarchical attention network for document understanding. 2019.

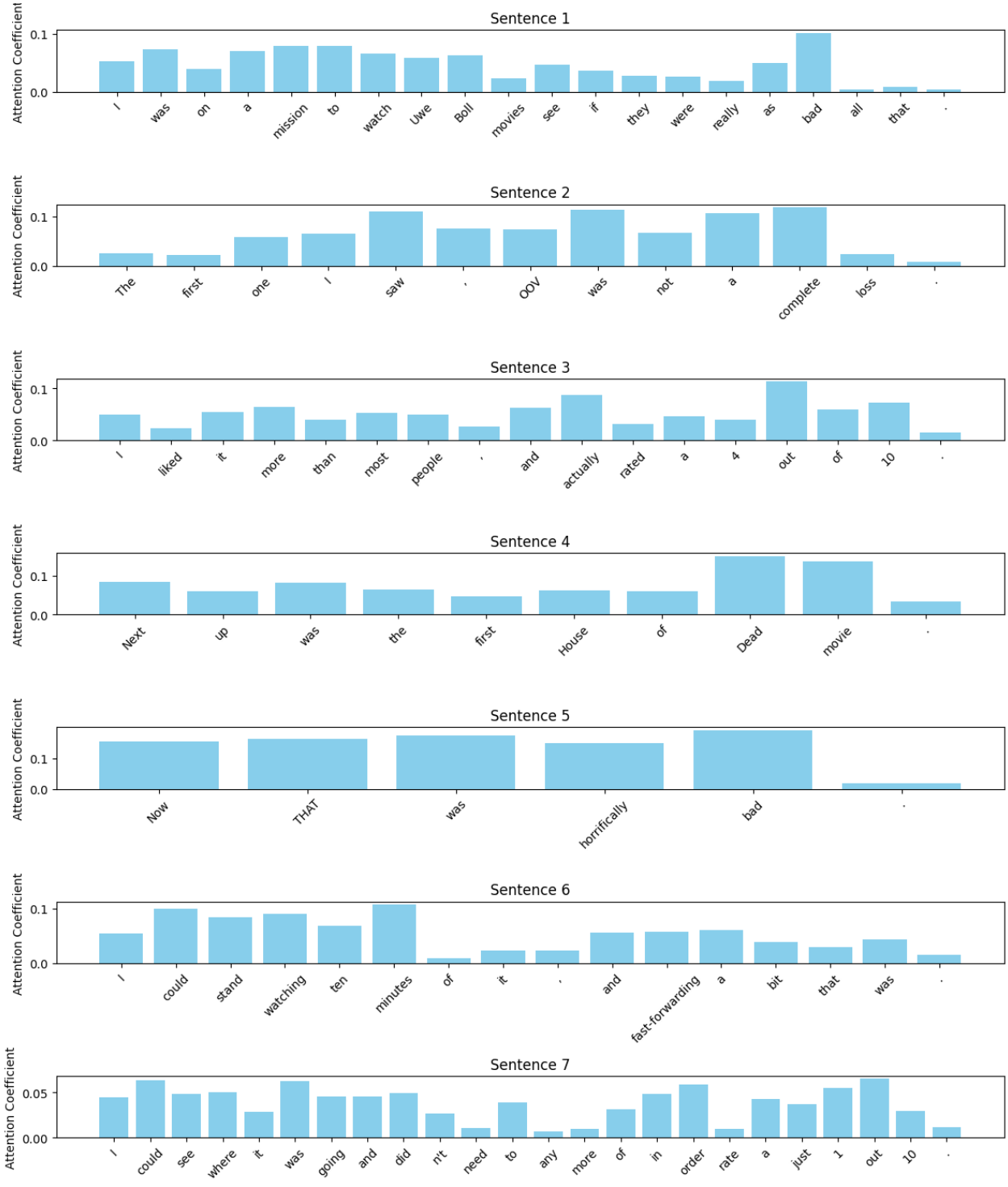

Figure 1: Attention coefficients for a document containing a review classified as negative by the model.