Julien Delavande
julien.delavande@ens-paris-saclays.fr

Lab session # 1
ALTEGRAD 2024

10/14/24

# 1 Question 1: How can the basic self-attention mechanism be improved?

In paper [1], the basic self-attention mechanism is improved by introducing structured self-attention with:

- **Multiple attention vectors:** This allows the model to attend to different parts of the sentence simultaneously using multiple attention hops, computed as:

  $$\mathbf{A} = \mathrm{softmax}\left(\mathbf{W}_s \tanh\left(\mathbf{H}^\top\right)\right)$$

  where $\mathbf{W}_s \in \mathbb{R}^{r\times d}$ is the attention weight matrix, and $r$ is the number of hops.

- **Sentence embedding:** The sentence representation is generated by:

  $$\mathbf{M} = \mathbf{AH}$$

  where $\mathbf{M} \in \mathbb{R}^{r\times d}$ contains $r$ sentence representations.

- **Diversity regularization:** To ensure diversity among attention hops, a penalization term is added:

  $$P = \left\|\mathbf{A}\mathbf{A}^\top - \mathbf{I}\right\|_F^2$$

  where $\mathbf{I} \in \mathbb{R}^{r\times r}$ is the identity matrix, encouraging orthogonality between attention vectors.

This mechanism enables richer sentence embeddings by focusing on multiple, diverse parts of the input sequence.

# 2 Question 2: Motivations for replacing recurrent operations with self-attention

The primary motivation for replacing recurrent operations with self-attention, as highlighted in the Transformer paper [2], is efficiency. Recurrent neural networks process sequences sequentially, which limits parallelization. Self-attention, on the other hand, enables parallel processing, allowing the model to handle longer sequences more efficiently. Additionally, self-attention has a shorter path between tokens, making it better suited to capture long-range dependencies.

# 3 Question 3: Attention coefficients for a chosen document

Below is the plot of attention coefficients for a document from the IMDB dataset (Fig. 1). The attention mechanism has assigned higher weights to words such as ”excellent”, ”performance”, and ”story”, indicating their importance in classifying the review as positive.

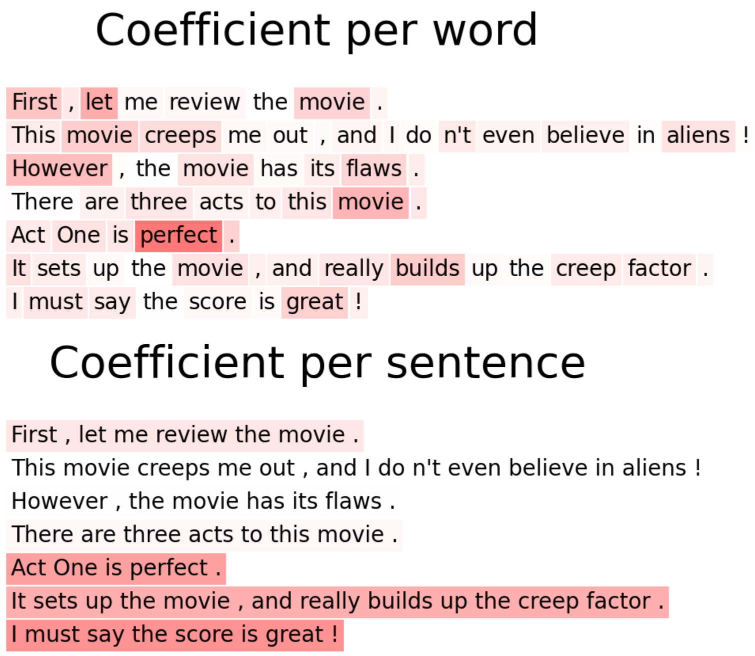

Figure 1: Attention coefficients for a sample document.

The model correctly identifies key phrases that contribute to the sentiment, showing the effectiveness of self-attention in focusing on the most relevant parts of the text.

## 4 Question 4: Limitations of the HAN architecture

While the Hierarchical Attention Network (HAN) captures the hierarchical structure of documents effectively, it has several limitations. First, it assumes a strict hierarchy of words, sentences, and documents, which may not always hold in practice. Additionally, the reliance on RNNs, even with attention, makes it slower than fully attention-based models like transformers, especially for large datasets. HAN also struggles with very long documents due to the sequential nature of RNNs, which limits its ability to process extensive contextual information efficiently.

## References

[1] Houhan 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] 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.