GRAVIER Thomas
thomas.gravier@gadz.org

Lab session # 1
ALTEGRAD 2024

10/15/24

1 Question 1

The self-attention mechanism can be improved through the introduction of multi-head attention, as seen in Transformer architectures. Multi-head attention applies the attention mechanism multiple times with different sets of weights, allowing the model to focus on different parts of the input sequence simultaneously. This leads to a richer representation of the input, as the model can attend to various dependencies in the data, improving performance on tasks like machine translation or document classification [1]. The multi-head structure also prevents information bottlenecks by offering the model multiple ”views” of the sequence.

Another improvement is the hierarchical application of attention, as in the Hierarchical Attention Network (HAN), which matches the natural structure of documents (words to sentences, sentences to documents). HAN first applies attention at the word level to determine which words are most important in a sentence, and then at the sentence level to highlight the key sentences in a document. This multi-level attention strategy allows the model to capture both local (within a sentence) and global (across sentences) context [1]. Further refinements could include dynamic context windows or the incorporation of external knowledge, making the attention process more flexible.

2 Question 2

Self-attention mechanisms replace recurrent operations, such as those used in RNNs or GRUs, because they allow for better long-range dependency modeling. In recurrent models, information is passed sequentially, which means dependencies between distant tokens can degrade over time, as they are stored in hidden states. Self-attention resolves this issue by creating direct connections between all tokens in a sequence, making it easier to capture relationships between distant words [3]. This is particularly beneficial for tasks like machine translation, where understanding long-range dependencies is critical for accurate translations.

Another key motivation for using self-attention is the ability to parallelize computation. In RNNs, tokens are processed one at a time, which makes training slow and less efficient. Self-attention, however, processes all tokens simultaneously, allowing the model to take full advantage of modern GPU architectures for faster training. This parallelization improves scalability, making self-attention more suitable for large-scale datasets and long sequences [3]. The combination of these benefits has made self-attention the backbone of state-of-the-art models like Transformers and BERT.

3 Bonus (*Purpose of the* `my_patience` *parameter*)

During the training of a neural network, the `my_patience` parameter refers to the patience threshold used for early stopping. Early stopping is a strategy designed to halt the training process when the model’s performance on the validation set ceases to improve, thereby preventing unnecessary training and reducing the risk of overfitting. This technique is particularly useful in ensuring that the model generalizes well to unseen data.

4 Question 3

In this task, I have selected the final document from the test dataset, where the model predicts a positive sentiment (”yes”). The sentence-level attention scores (Figure 1) reveal that the model places higher emphasis on specific sentences that clearly convey a positive tone. For instance, sentences such as *”a masterpiece”* and *”[...] downright brilliant”* receive higher weights, as they are crucial in guiding the model’s decision towards a positive classification.

At the word level, the attention scores (Figure 2) for the sentence *”:) First of all, Mulholland Drive is downright brilliant.”* highlight key words such as ”downright” and ”brilliant.” These words, when combined, strongly contribute to the overall positive sentiment and therefore attract more attention from the model.

To conclude, both sentence- and word-level attention scores demonstrate the model’s ability to focus on the most sentiment-relevant parts of the review. This explains its confident prediction of a positive (”yes”) classification, as it successfully identifies and prioritizes key elements of the text.

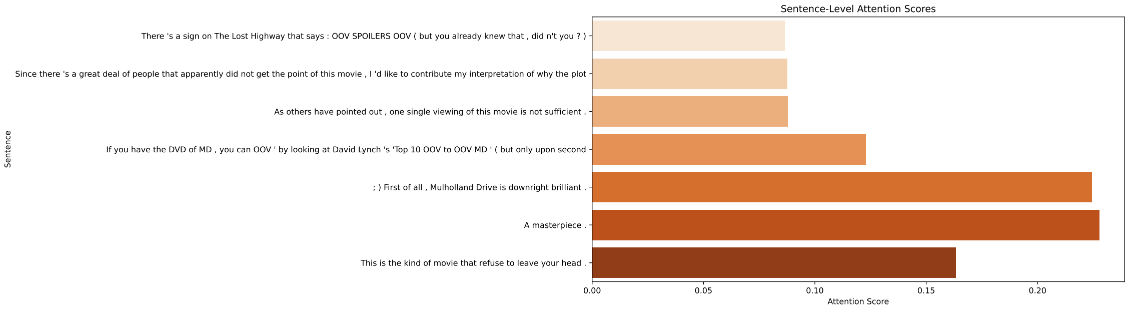

Figure 1: Sentence-level attention scores for the final document from the test dataset.

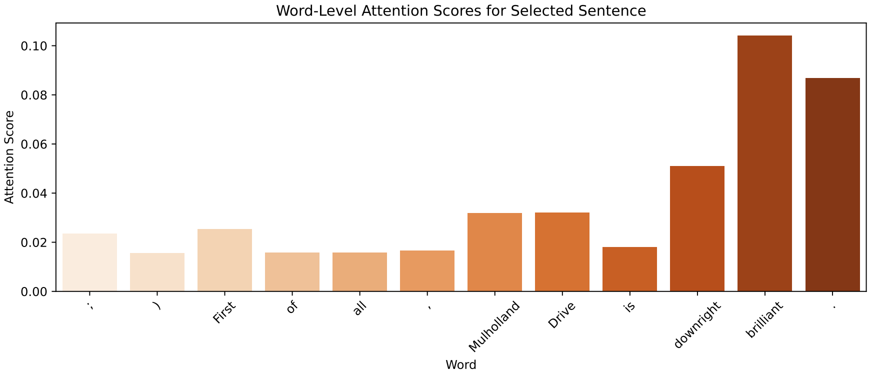

Figure 2: Word-level attention scores for the final sentence of the final document in the test dataset.

5 Question 4

The article [2] highlights several limitations of Hierarchical Attention Networks (HAN) in terms of architecture and training.

Firstly, the model processes sentences independently during the initial step of its architecture. As a result, while the model analyzes one sentence, it ignores the others, which can hinder its ability to capture the overall meaning of the document. This approach may impair the understanding of context and the relationships between sentences.

Moreover, the embedding representation for each sentence is uniform across all instances, limiting the extraction of diverse and complementary information. By not differentiating embeddings for various instances, HAN may overlook important nuances that could enhance its performance in document understanding tasks.

References

[1] Zhouhan Lin and al. A structured self-attentive sentence embedding. *ICLR*, ibm(1703.03130), 2017.

[2] Jean-Baptiste Remy and al. Bidirectional context-aware hierarchical attention network for document understanding, 2019.

[3] Ashish Vaswani and al. Attention is all you need, 2023.