Solal Danan
solal.danan@polytechnique.edu

Lab session # 1
ALTEGRAD 2023

10/15/24

1 Question 1

In the basic self-attention mechanism, a single attention vector is used to weigh the importance of tokens in a sentence. However, a single attention head might not capture the entire context of a sentence, especially in longer or more complex sentences. A key improvement is to use multiple attention heads (or hops), which allow the model to focus on different parts of the input sequence. Instead of relying on a single attention vector that might miss important details, these multiple heads create an attention matrix where each row captures a distinct aspect of the sentence's meaning. This enhancement creates a 2D matrix representation (instead of a single vector) where each row corresponds to the output of one attention head. This improvement, inspired by the work of Zhouhan Lin et al. [1], also improves the interpretability of the model, making it easier to visualize which parts of the sentence each head is focusing on. These improvements help the model better capture long-term dependencies while reducing redundancy in the representations. That said, without some form of constraint, there's a risk that the attention heads will overlap and focus on the same parts of the sequence, which can lead to redundancy. To prevent this, a penalization term can be introduced.

2 Question 2

The decision to replace recurrent operations with self-attention in the Transformer model [3] relying entirely on an attention mechanism came from several motivations:

One of the biggest drawbacks of recurrent models is their inherently sequential nature. In RNNs and LSTMs, each computation step depends on the result of the previous one, which forces the model to process tokens in order. This severely limits the ability to parallelize training, slowing down training, especially with longer sequences, as the number of sequential operations scales with the length of the input. Self-attention, however, doesn't have this limitation. It allows the model to process all tokens in the input sequence simultaneously, since the attention mechanism relates all tokens to each other in parallel. This results in much faster training times.

Another key issue with recurrent models is their difficulty in capturing long-range dependencies. As the sequence length grows, the number of steps the information has to travel increases, and signals from earlier tokens can degrade or get lost, making it harder for the model to retain long-term information. This is partly due to the vanishing gradient problem, where gradients diminish as they are propagated back through many time steps, causing the model to "forget" earlier tokens. Self-attention overcomes this by allowing every token to directly attend to every other token, regardless of their distance in the sequence. This means the model can easily capture relationships between distant tokens, which is crucial for tasks in NLP, where words at opposite ends of a sentence may still be related.

3 Question 3

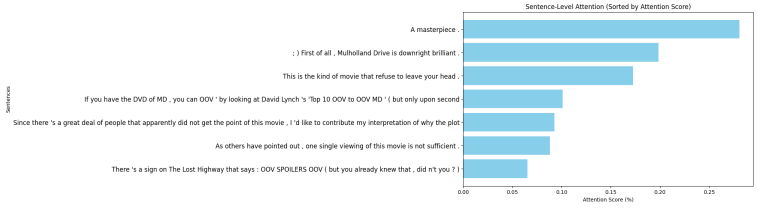

Figure 1: Sentence-Level Attention Plot.

The sentence-level attention plot (Figure 1) highlights the most important sentences in the review. Sentences like "A masterpiece." and "First of all, Mulholland Drive is downright brilliant." receive the highest attention scores, indicating that these strong, positive statements are crucial for the model's sentiment classification. On the other hand, sentences with neutral or less relevant information, such as "There 's a sign on The Lost Highway that says: OOV SPOILERS OOV...", are assigned lower attention. The model effectively manages out-of-vocabulary (OOV) words, reducing their impact on the prediction by assigning them low attention. This

shows that the attention mechanism is well-aligned with human intuition, focusing on the most meaningful content for sentiment analysis, while handling unfamiliar words without overemphasizing them.

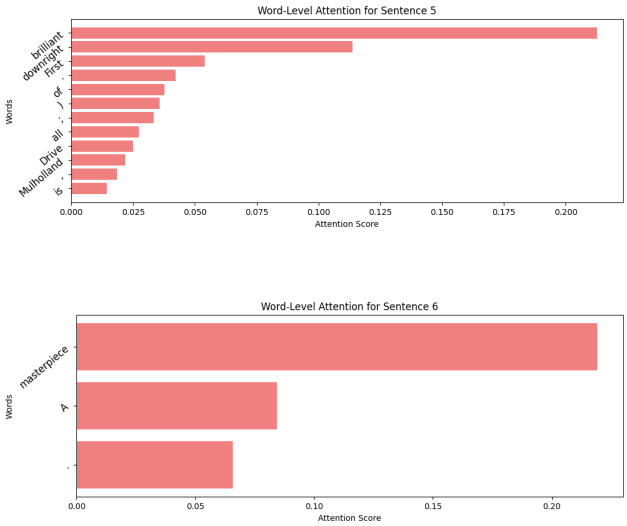

Figure 2: Word-Level Attention Plot highlighting words sorted by their attention score.

At the word level (Figure 2), the model focuses on sentiment-heavy words like ”masterpiece”, ”brilliant”, and ”downright”, which receive the highest attention scores. Words that are contextually important but carry less sentiment, such as ”Mulholland” or ”is”, are given lower scores, showing that the model prioritizes sentiment-bearing terms.

4 Question 4

The limitation of the original HAN (Hierarchical Attention Network) architecture is that at level 1, each sentence is encoded in isolation, meaning it is encoded separately and does not consider the surrounding sentences. This lack of context results in suboptimal sentence representations, as the encoder cannot capture how sentences influence each other. As a result, contextual dependencies between sentences, which are crucial for understanding the document's full meaning, are ignored. While the level 2 document encoder in HAN assigns importance scores to sentences, by that point, it is too late to modify the sentence representations themselves. The sentence vectors have already been formed, and the model cannot address high redundancy or ensure that important subtopics are represented. Thus, the document encoder only ranks the fixed sentence representations, leading to a loss of detail and potential misinterpretation of the document's overall meaning.

The CAHAN model [2] addresses the limitations of HAN by allowing the sentence encoder at level 1 to make its attentional decisions based on contextual information from surrounding sentences. Unlike HAN, where each sentence is encoded in isolation, CAHAN incorporates a bidirectional document encoder. This encoder processes the document in two directions: one RNN processes the document forwards, using the preceding sentences as context, and another RNN processes it backwards, using the following sentences as context. By leveraging both past and future sentences, CAHAN enables the model to create more context-aware sentence representations, improving document understanding and overcoming the redundancy and lack of detail that affect HAN. This modification ensures that important subtopics and sentence-level interactions are better captured during encoding, leading to more accurate document-level predictions.

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 J.-P. Tixier, and Michalis Vazirgiannis. Bidirectional context-aware hierarchical attention network for document understanding. *arXiv preprint arXiv:1908.06006v1*, 2019.

[3] 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.