Soël MEGDOUD
soel.megdoud@ens-paris-saclay.fr

Lab session # LAB1
ALTEGRAD 2024

10/14/24

Question 1

The basic self-attention mechanism can be improved by representing the sentence embedding in a 2D matrix instead of a simple vector.

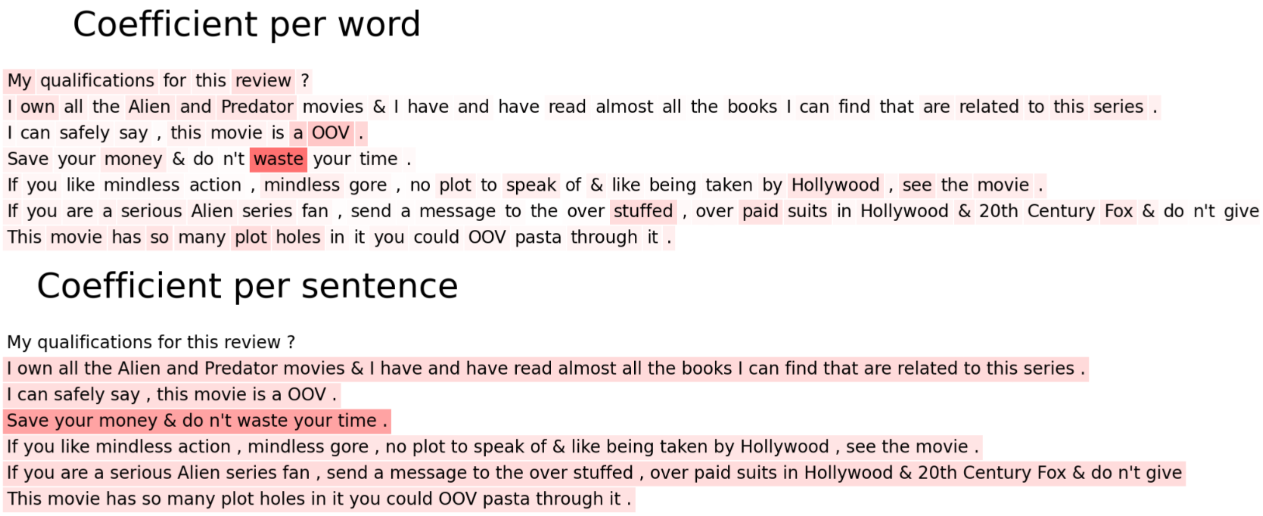
- The paper uses a 2D matrix to represent a sentence, where each row focuses on different parts of the sentence, and applies multiple passes of attention to capture various important components. This enables to represent sentences with multiple semantic aspects.
- A regularization term is introduced to ensure that attention is distributed in different parts of the sentence, preventing over-focus on a single part.
- This 2D matrix also allows to an helpful visualization making the model more explainable by showing which parts of the sentence are encoded in the embedding. This can be useful for tasks like sentiment analysis and text classification.

Question 2

The main motivations for replacing recurrent operations with self-attention are computational complexity and better long-range dependencies.

- Recurrent models like RNN process sequences step by step, limiting parallelization since each step depends on the previous one. The sequential operations are in O(n) whereas for Self-Attention layers there are constant (O(1)). Therefore, self-attention layers are faster than recurrent layers when the sequence length is smaller than the representation dimensionality. Self-attention layers in Transformers offer significant advantages in parallelization, the attention operations are essentially matrix products with the Query, Key and Value matrices and these calculations can easily be optimised for the GPU.
- Recurrent models struggle to capture long-range dependencies due to their sequential nature. Self-attention overcomes this by allowing each token in the sequence to directly attend to every other token, enabling the model to capture long-range relationships more effectively. The path length—the number of operations required to connect two distant tokens—is in self-attention is constant (O(1)) whereas a recurrent layer requires O(n) sequential operations. This makes it easier to capture long range dependencies.

Question 3



Coefficient per word

My qualifications for this review ?
I own all the Alien and Predator movies & I have and have read almost all the books I can find that are related to this series .
I can safely say , this movie is a OOV .
Save your money & do n't waste your time .
If you like mindless action , mindless gore , no plot to speak of & like being taken by Hollywood , see the movie .
If you are a serious Alien series fan , send a message to the over stuffed , over paid suits in Hollywood & 20th Century Fox & do n't give
This movie has so many plot holes in it you could OOV pasta through it .

Coefficient per sentence

My qualifications for this review ?
I own all the Alien and Predator movies & I have and have read almost all the books I can find that are related to this series .
I can safely say , this movie is a OOV .
Save your money & do n't waste your time .
If you like mindless action , mindless gore , no plot to speak of & like being taken by Hollywood , see the movie .
If you are a serious Alien series fan , send a message to the over stuffed , over paid suits in Hollywood & 20th Century Fox & do n't give
This movie has so many plot holes in it you could OOV pasta through it .

Figure 1: Attention coefficients.

This plot allows to understand that the model based its prediction mostly on the sentence ’Save your money & don’t waste your time’ which is indeed very important to understand that this is a bad review. The word ’waste’ has been the most usefull to make the prediction.

Question 4

The Hierarchical Attention Network has notable limitations. The critical drawback is its isolated sentence encoding at level 1. In HAN, each sentence within a document is processed independently by the sentence encoder, which means it lacks contextual awareness from neighboring sentences. This approach leads to suboptimal performance in scenarios where sentences share similar content or when understanding the broader context of the document is crucial.
For instance, in cases where a document contains repetitive phrases or themes across sentences, HAN tends to allocate attention to the same prominent features repeatedly, neglecting other important aspects of the document. This limitation can result in incomplete document representations and less accurate predictions, particularly in tasks requiring nuanced understanding such as abstractive summarization or detailed topic coverage.

Two key improvements are proposed in the paper:

- The use of context vectors that allow the model to account for preceding and following sentences.
- A bidirectional document encoder that processes the document both forwards and backwards.

These modifications help reduce redundancy, improve coverage of subtopics, and lead to richer document representations.