Paul Londres
paul.londres@free.fr

Lab session # Lab 1
ALTEGRAD 2024

10/15/24

1 Question 1

Here are some ways to improve the basic self-attention mechanism :

- One of the main limitations of basic attention is that it can have trouble capturing multiple relationships between sequence elements (whether it's words or sentences) in a single attention vector. To remedy this, multi-head attention divides representations into several subspaces and applies attention independently in each of these subspaces. This makes it possible to capture several aspects of the relationships between elements. This concept of multi head attention was introduced by this paper [6].
- Another thing that could be improved in the basic self attention mechanism is the combination formula of the hidden states, mentionned in [4] the simple weighted sum formula could be changed to correspond to a more complex relationship of the hidden states.
- One of the main drawbacks of the basic self attention mechanism is mentioned in [4] and is the fact that it cannot be trained in an unsupervised way. This issue was adressed with BERT [2] and GPT [1] models which are pre-trained in an unsupervised way.

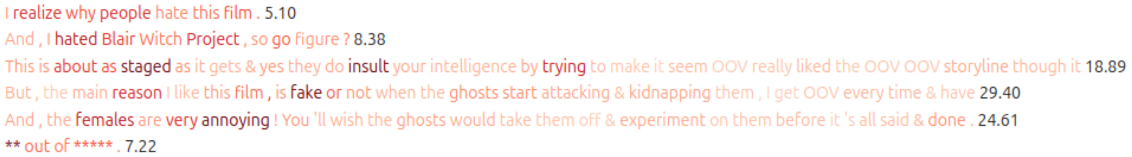
2 Question 2

Here are the three main motivations cited in [6] :

1. The first motivation is the total complexity per layer which is directly related to the computation efficiency, the RNN used to have an $\mathcal{O}(n.d^2)$ complexity with n, the length of the sequence and d the embedding size while self attention has an $\mathcal{O}(n^2.d)$ and as $n < d$ in most models, self-attention is more efficient per layer than RNNs
2. The second reason is the fact that during training, more operations can be parallelized is self-attention mechanism than in RNNs, this is due to the fact that each part of a sequence can be treated in parallel while in RNNs even in training, the sequence has to be treated sequentially
3. The last motivation is the fact that long-term dependencies are learnt more easily in self-attention mechanism than in RNNs. The length of the path that signals need to do back and forth in the network are directly related to the ease of learning long term dependencies [5], for RNNs, the length of this path is of order $\mathcal{O}(n)$ while in self-attention, it is $\mathcal{O}(1)$.

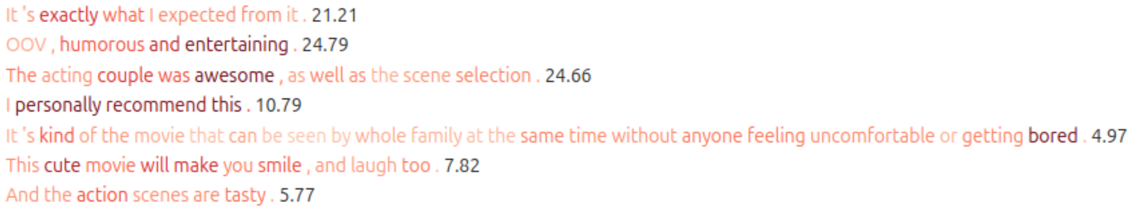
3 Question 3

Here are two examples of reviews in the IMDB movie review database, the word attention are showed in color (red means high coefficient) and the sentence attention coefficient is showed at the end of each sentence :



I realize why people hate this film . 5.10
And , I hated Blair Witch Project , so go figure ? 8.38
This is about as staged as it gets & yes they do insult your intelligence by trying to make it seem OOV really liked the OOV OOV storyline though it 18.89
But , the main reason I like this film , is fake or not when the ghosts start attacking & kidnapping them , I get OOV every time & have 29.40
And , the females are very annoying ! You 'll wish the ghosts would take them off & experiment on them before it 's all said & done . 24.61
** out of ***** . 7.22

Figure 1: Attention coefficients of a negative review



It 's exactly what I expected from it . 21.21
OOV , humorous and entertaining . 24.79
The acting couple was awesome , as well as the scene selection . 24.66
I personally recommend this . 10.79
It 's kind of the movie that can be seen by whole family at the same time without anyone feeling uncomfortable or getting bored . 4.97
This cute movie will make you smile , and laugh too . 7.82
And the action scenes are tasty . 5.77

Figure 2: Attention coefficients of a positive review

We can see that in both reviews, the attention coefficient of words seems to really correlate with the impact of the word on the review, for example we can see in fig. 1 the words ”insult”, ”fake”, ”annoying” with high attention coefficient are word with high semantic meaning. In the positive review (fig. 2) the words ”entertaining”, ”awesome”, ”recommend” also have high attention coefficients. However not all words with important semantic meaning have high attention coefficient such as ”hate” in the negative review or ”laugh” in the positive review. Both reviews were correctly classified and one thing we can observe is that words that could be interpreted as positive or negative depending on the context were correctly understood by the classifier such as ”bored” in the positive review.

The sentence attention coefficients however, do not seem to be truly correlated with the impact of a sentence in the review. For example, sentences like ”I realize why people hate this film” (fig. 1) or ”I personally recommend this” (fig. 2) have a low attention coefficient.

4 Question 4

Here are some limitations of the HAN model :

- One of the limitation of this model is that it can only be trained in a supervised way which can cause issues if we want to train it on large datasets.
- Another limitation discussed in [3] is that each sentence is encoded independently which means that attention coefficients of words are computed without knowing the neighbouring sentences. This can be a problem, especially when repeating parts of a sentence on which the attention budget will only focus on, leaving aside other words that should have more impact.
- Han uses RNNs in his architecture which as explained in question 2 are costly networks in terms of computational resources.

5 Bonus section

What is the purpose of the parameter $my\_patience$?
This parameter allows to stop the training if the test accuracy has not been improving for $my\_parience$ epochs, hence saving computer resource if the model is not training correctly anymore.

References

[1] url=https://cdn.openai.com/research-covers/language-unsupervised/language$_{u}nderstanding_{p}aper.pdf AlecRadfordKa$ 2018.$Improvinglanguageunderstandingbygenerativepre-training.$

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Michalis Vazirgiannis Jean-Baptiste Remy Antoine J.-P. Tixier1. Bidirectional context-aware hierarchical attention network for document understanding, 2019.

[4] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding, 2017.

[5] Sepp Hochreiter Yoshua Bengio Paolo Frasconi J¨urgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[6] 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.