CS475 Spring 2021 Homework 3

Team 5: Assem Zhunis, Ern Chern Khor, Bella Godiva, Bauyrzhan Tokenov

May 9, 2021

1 Our BERT Pooler

We have implemented the model based on **K-max pooling** (2). For every input sentence, we obtain the top k tokens by weight of their embeddings and then calculate their average. This pooling method is chosen to preserve more information compared to max pooling. The idea may be represented by the following formula:

$$C_{\text{KMAX}} = \left[\frac{\sum_{i=0}^{k} T_i}{k} \right], \quad C_{\text{KMAX}} \in \mathbb{R}^H, \quad T_i \in \mathbb{R}^H. \tag{1}$$

Where T_i's correspond to top k tokens by weight. Note that in case where number of tokens is smaller then k our formula will simply compute the average of all tokens, or mean-pooling, and in case $k = 1$ the method converges to the max-pooling.

2 Experiments and Results

For the experiments we evaluated BERT with different pooling methods. For the experiments we compared results with same batch size (16) and epoch number (3). We took CLS method as a benchmark. You can find the summary f the results in the Table 1

From the Figure 1 you can find the results of experiments with different k-values on the Cola task. As we can see the best performance is achieved with $k = 2$. Interestingly, increasing k-value doesn't improve the result for the Cola task. We found that our method outperforms the default CLS method as well as MEANMAX pooling method.

Figure 2 represents the results that we got on the MRPC task. Contrary to the Cola task here we can see that k value has some positive correlation with the performance. Both F1 score and accuracy are increasing with higher k values. With k value equal to 5, KMAX pooling method has higher performance than both CLS and MEANMAX methods.

3 Discussion

From the Table 1 you can see that our KMAX method is better than the CLS and MEANMAX on the Cola task with $k = 2$. Increasing the k value did not improve the performance on this task. We may conclude that top 2 tokens carry the most crucial information for this task. Considering more tokens might blur their effect and thus decrease the performance.

We also see that in case of the MRPC task, different k value than Cola task is needed to get better results. The reason can be that in the MRPC task, more information needs to be included to give a better accuracy of checking whether sentences are paraphrases from each other (1). So, more tokens are needed, which means we need higher k.

To conclude, **K-max** method can outperform CLS method by adjusting the k value for each task. We have shown that optimal k values differ according to the dataset, which might also give additional information on the properties of the dataset. However, more experiments need to be done to achieve the best results.

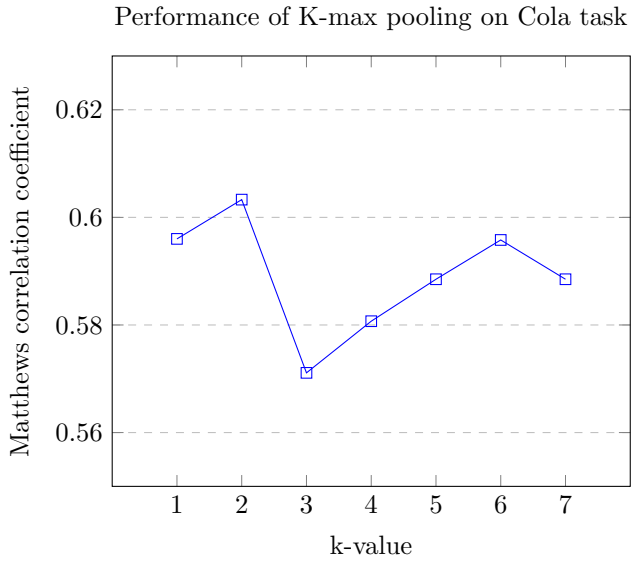

Figure 1: Experiments on Cola task

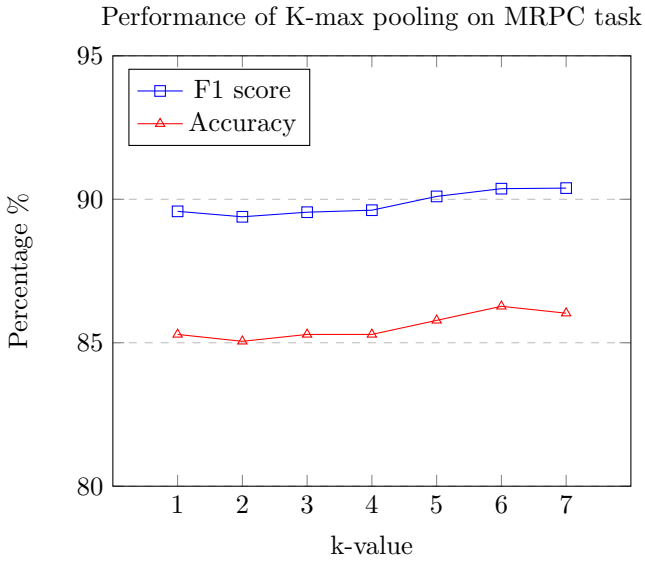

Figure 2: Experiments on MRPC task

GLUE Task	Evaluation Metric	CLS	MEANMAX	KMAX k=2	KMAX k=6
Cola	Matthews correlation coefficient	0.5760	0.5729	**0.6033**	0.5958
MRPC	Accuracy	86.02%	85.05%	85.05%	**86.27%**
	F1 score	90.09%	89.57%	89.39%	**90.38%**

Table 1: Results of different pooling methods on 2 GLUE tasks.

References

[1] Microsoft research paraphrase corpus. URL: https://www.microsoft.com/en-us/download/details.aspx?id=52398.

[2] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014. URL: http://goo.gl/EsQCuC.