AttentiveCLS Pooler

Team 1: Seungil Lee, Jongchan Park, Eugene Seo, and Dohyung Kim

November 5, 2021

1 Introduction

BERT [1] has become a standard architecture for NLP research ever since it was published. BERT computes the representation for every token, but it uses the output representation of the special token `[CLS]`, for sentence-level tasks (e.g., sentiment analysis). However, various strategies were introduced to get sentence-level representation.

We have designed new pooler named `AttentiveCLS` pooler and evaluated its performance on CoLa, MNLI, RTE and SST-2 tasks, which are the representatives of sentence-level tasks. The results were compared with `MeanMaxTokens` pooler, which is suggested in the homework description, also with the original `BERTPooler` in `huggingface` library (`https://huggingface.co/`).

2 AttentiveCLS Pooler

Though original BERT pooler simply adopts last output of `[CLS]` token, we tried to exploit information from other tokens as well. Nowadays, attention mechanism is widely used to get the importance of the given sequence, so we added a attention pooling layer on the top of the tokens other than `[CLS]` token. Then, the output of this attention layer is concatenated with the last output of `[CLS]` token. We also apply linear transformation with $W \in \mathbb{R}^{H \times 2H}$ and tanh activation just like as the `BERTPooler` in `huggingface` implementation.

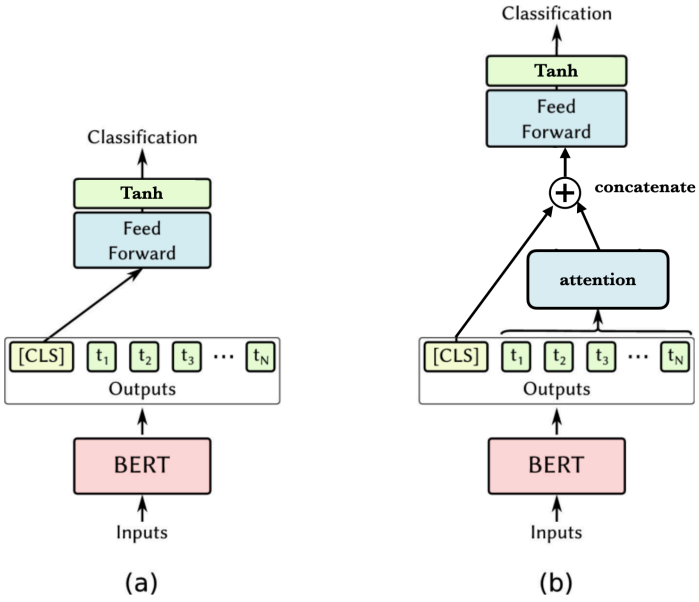

Figure 1: Schematic visualization of AttentiveCLS. This figure is modified from the paper [2]

3 Result

	CoLA		MRPC			RTE		SST-2	
	Loss	Matthews	Accuracy	F1	Loss	Accuracy	Loss	Accuracy	Loss
AttentiveCLS	0.4588	**0.6044**	**0.8544**	**0.9012**	**0.3952**	0.6137	0.7498	0.9289	0.3272
MeanMax	**0.4208**	0.5375	0.8162	0.8777	0.4424	0.5379	**0.6920**	**0.9312**	0.2711
BERTPooler	0.4388	0.5934	0.8456	0.8934	0.4000	**0.6462**	0.7058	0.9300	**0.2387**

Table 1: Table showing the result of 4 experiments. Bold numbers indicate the best performance among 3 poolers.

In this experiment, we use 4 different text classification tasks (CoLA, MRPC, RTE, SST-2) in the GLUE [3] benchmark to evaluate 3 pooler strategies. Each dataset has its own evaluation metrics (e.g. Accuracy, F1).

Table 1 shows our experiment results. For CoLA task, AttentiveCLS results highest performance for Matthews correlation and MeanMax has the best performance for loss value. Interestingly, for MRPC task, AttentiveCLS is superior to other strategies for all evaluation metrics. However, for RTE and SST-2 tasks, AttentiveCLS doesn't have the highest performance for any metrics. For RTE task, BERTPooler has the best accuracy and MeanMax has the best loss value. On the other hand, for SST-2 task, MeanMax has the best accuracy and BERTPooler has the best loss value.

This result shows that each pooler can show different performance depending on the dataset. In our dataset, AttentiveCLS has good performance on most metrics. Interestingly, only AttentiveCLS shows the best performance for all metrics in the specific dataset(MRPC).

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.

[2] Jan Lehečka, Jan Švec, Pavel Ircing, and Luboš Šmídl. Adjusting bert's pooling layer for large-scale multi-label text classification. In *International Conference on Text, Speech, and Dialogue*, pages 214–221. Springer, 2020.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.