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

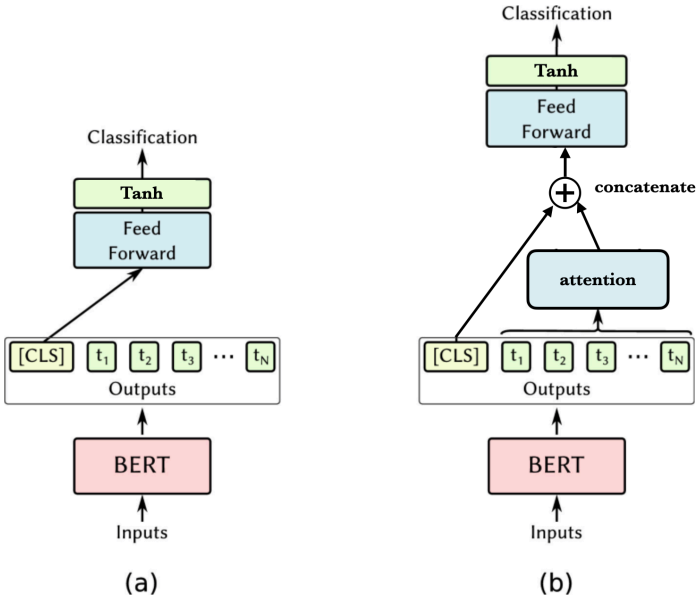

Figure 1: Schematic visualization of AttentiveCLS. This figure is modified from the paper [2].

	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	**0.4246**	**0.6060**	0.8382	0.8907	0.4570	0.6101	**0.6818**	**0.9300**	**0.2206**
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.

3 Result

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

Table 1 shows our experiment results.

1. CoLA: `MeanMax` reached the highest performance in terms of Matthews correlation score and validation loss. `AttentiveCLS` beated `BERTPooler`.
2. MRPC: Interestingly, `AttentiveCLS` was superior to other strategies in terms of all evaluation metrics.
3. RTE: `AttentiveCLS` did not perform better than others. `BERTPooler` had the best accuracy and `MeanMax` had the lowest loss.
4. SST-2: Both `MeanMax` and `BERTPooler` showed the best accuracy and `BERTPooler` had the best loss value.

This result showed that each pooler can make different performances depending on the dataset. `AttentiveCLS` took a first place in MRPC, while it was not performed well in CoLA, RTE and SST-2. However, the differences were insignificant. There was no big difference in performance according to the pooler, so we concluded that the pooler strategy had no critical effect on text classification task.

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.