# BERT for Multiple Choice Machine Comprehension

**Chenglei Si**
River Valley High School, Singapore
sichenglei1125@gmail.com

## Abstract

Recently introduced pre-trained language encoders such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) have achieved impressive empirical results on a number of NLP tasks. With such pre-trained encoders as feature extractors, we are able to build more powerful models for tasks such as machine comprehension. In this paper, we describe a simple but effective way to apply BERT for multiple choice machine comprehension. We tested our method on RACE dataset and achieved the best result on the leaderboard. Our codes are available at: https://github.com/NoviScl/BERT-RACE.

## 1 Introduction

The task of machine reading comprehension requires the model to understand natural languages and to do reasoning and inference based on the text information. An effective way of assessing the model's ability of reading comprehension is through question answering. Various question answering datasets have been made in recent years, such as MCTest (Richardson et al., 2013), Children's Book Test (Hill et al., 2015), CNN/Daily Mail (Hermann et al., 2015), NewsQA (Trischler et al., 2017), SQuAD (Rajpurkar et al., 2016), MS MARCO (Nguyen et al., 2016), and RACE (Lai et al., 2017). Among them, datasets like RACE are particularly challenging because most of the answers are not directly extracted from the passage and many questions require the model to perform reasoning and inference. This difficulty is also reflected by the significant gap between the current SOTA and human performance [1].

In this paper, we describe a simple way to apply pre-trained BERT model on the RACE dataset that achieved SOTA results.

## 2 Model

In this section, we describe how we apply BERT for RACE dataset [2].

### RACE dataset

RACE is a dataset for multiple choice machine comprehension task. Each example in RACE consists of a passage, a question and four candidate answers for the question. The task is to choose the correct the candidate answer from the four based on the information from the passage. We refer readers to Lai et al. (2017) for more details about the dataset.

### BERT for RACE

In the original BERT model, the input sequence is either a single text sentence or a pair of text sentences. In RACE, there are three components of each example: passage, question and answers. We construct four input sequences for each example, one for each option among the four candidate answers. Following the notation of Devlin et al. (2018), we input the passage as sentence A and the concatenation of the question and the candidate answer as sentence B. The input sequence can be denoted as:

[[CLS] Passage [SEP] Question + Option [SEP]]

We follow the same fine-tuning procedure as Devlin et al. (2018). We apply a linear layer and a softmax layer on the final hidden state for the class token for the four input sequences of each example. We maximize the log-probability of the correct label.

[1] http://www.qizhexie.com/data/RACE_leaderboard

[2] http://www.cs.cmu.edu/~glai1/data/race/

| Model | RACE-M | RACE-H | RACE |
|---|---|---|---|
| Co-Match (Wang et al., 2018) | 55.8 | 48.2 | 50.4 |
| DFN* (Xu et al., 2017) | 55.6 | 49.4 | 51.2 |
| Bi-Attention* (Tay et al., 2018) | 60.2 | 50.3 | 53.3 |
| OpenAI GPT (Radford et al., 2018) | 62.9 | 57.4 | 59.0 |
| Reading Strategies (Sun et al., 2018) | 69.2 | 61.5 | 63.8 |
| Reading Strategies* | 72.0 | 64.5 | 66.7 |
| BERT$_{base}$ | 71.7 | 62.3 | 65.0 |
| BERT$_{large}$ | **75.6** | **64.7** | **67.9** |
| Turkers | 85.1 | 69.4 | 73.3 |
| Human Ceiling | 95.4 | 94.2 | 94.5 |

Table 1: Experiment Results on RACE. *ensemble model

## 3 Experiment

After some tuning, we find the following hyper-parameters to work reasonably well.

BERT base: batch size 32, learning rate 5e-5, training epoch: 3.

BERT large: batch size: 8, learning rate 1e-5, training epoch: 2.

We find that it is important to use small learning rate for BERT large, otherwise the training will not be effective.

We evaluate our model on the test set, which consists of two sets: RACE-M from middle school examinations and RACE-H from high school examinations. In 1, we show our experiment results on RACE-M, RACE-H and the entire RACE test set. BERT large model achieves state-of-the-art result.

## 4 Conclusion

We described a simple but effective way of applying BERT on multiple choice machine comprehension dataset RACE. Our empirical result shows that BERT large achieves the current best result on RACE.

## Acknowledgment

We thank Shuohang Wang from Singapore Management University for his advice.

## References

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/languageunderstandingpaper.pdf`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, 2013.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *CoRR*, abs/1511.02301, 2015.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*, 2017.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*, 2017.

Shuohang Wang, Mo Yu, Shiyu Chang, and Jing Jiang. A co-matching model for multi-choice reading comprehension. In *ACL*, 2018.

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. Dynamic fusion networks for machine reading comprehension. 2017.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Multi-range reasoning for machine comprehension. *CoRR*, abs/1803.09074, 2018.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Improving machine reading comprehension with general reading strategies. *CoRR*, abs/1810.13441, 2018.