Improving Machine Reading Comprehension with Single-choice Decision and Transfer Learning

Yufan Jiang[1]*, **Shuangzhi Wu**[1]*, **Jing Gong**[2]*, **Yahui Cheng**[2]*,
Peng Meng[2], **Weiliang Lin**[2], **Zhibo Chen**[2] and **Mu Li**[1]
[1]Tencent Cloud Xiaowei
[2]Tencent Cloud TI-ONE
{frostwu,garyyfjiang,jennygong,huecheng}@tencent.com
{pengmeng,weilianglin,ruibobchen,ethanlli}@tencent.com

Abstract

Multi-choice Machine Reading Comprehension (MMRC) aims to select the correct answer from a set of options based on a given passage and question. Due to task specific of MMRC, it is non-trivial to transfer knowledge from other MRC tasks such as SQuAD, Dream. In this paper, we simply reconstruct multi-choice to single-choice by training a binary classification to distinguish whether a certain answer is correct. Then select the option with the highest confidence score. We construct our model upon ALBERT-xxlarge model and estimate it on the RACE dataset. During training, We adopt AutoML strategy to tune better parameters. Experimental results show that the single-choice is better than multi-choice. In addition, by transferring knowledge from other kinds of MRC tasks, our model achieves a new state-of-the-art results in both single and ensemble settings.

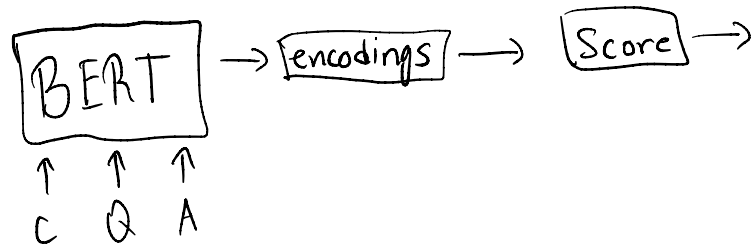
1 Introduction

The last several years have seen a land rush in research on machine reading (MRC) comprehension and various dataset have been proposed such as SQuAD1.1, SQuAD2.0, NewsQA and CoQA (Rajpurkar et al., 2016; Trischler et al., 2016; Reddy et al., 2019). Different from the above which are extractive MRC, RACE is a multi-choice MRC dataset (MMRC) proposed by (Lai et al., 2017). RACE was extracted from middle and high school English examinations in China. Figure 1 shows an example passage and two related questions from RACE. The key difference between RACE and previously released machine comprehension datasets is that the answers in RACE often cannot be directly extracted from the passages, as illustrated by the two example questions (Q1 & Q2) in Table 1. Thus, answering these questions needs inferences.

Passage: For the past two years, 8-year-old Harli Jordean from Stoke Newington, London, has been selling marbles . His successful marble company, Marble King, sells all things marble-related - from affordable tubs of the glass playthings to significantly expensive items like Duke of York solitaire tables - sourced, purchased and processed by the mini-CEO himself. "I like having my own company. I like being the boss," Harli told the Mirror....Tina told The Daily Mail. "At the moment he is annoying me by creating his own Marble King marbles - so that could well be the next step for him."
Q1: Harli's Marble Company became popular as soon as he launched it because ___ . **A**: it was run by "the world's youngest CEO" **B: it filled the gap of online marble trade** **C**: Harli was fascinated with marble collection **D**: Harli met the growing demand of the customers
Q2: How many mass media are mentioned in the passage? **A**: One **B**: Two **C: Three** **D**: Four

Table 1: An example passage and two related multi-choice questions. The ground-truth answers are in **bold**.

Recently, pretrained language models (LM) such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) have achieved great success on MMRC tasks. Notably, Megatron-LM (Shoeybi et al., 2019) which is a 48 layer BERT with 3.9 billion parameters yields the highest score on the RACE leaderboard in both single and ensemble settings. The key point to model MMRC is: first encode the context, question, options with BERT like LM, then add a matching network on top of BERT to score the options. Generally, the matching network can be various (Ran et al., 2019; Zhang et al., 2020; Zhu et al., 2020). Ran et al. (2019) proposes an option



*Equal contribution. Correspondence to {*frostwu*, *garyyfjiang*, *jennygong*, *huecheng*}@tencent.com.*

comparison network (OCN) to compare options at word-level to better identify their correlations to help reasoning. Zhang et al. (2020) proposes a dual co-matching network (DCMN) which models the relationship among passage, question and answer options bidirectionally. All these matching networks show promising improvements compared with pretrained language models. One point they have in common is that the answer together with the distractors are jointly considered which we name multi-choice models. We argue that the options can be concerned separately for two reasons, 1) when human works on MMRC problem, they always consider the options one by one and select the one with the highest confidence. 2) MMRC suffers from the data scarcity problem. Multi-choice models are inconvenient to take advantage of other MRC dataset.

In this paper, we propose a single-choice model for MMRC. Our model considers the options separately. The key component of our method is a binary classification network on top of pretrained language models. For each option of a given context and question, we calculate a confidence score. Then we select the one with the highest score as the final answer. In both training and decoding, the right answer and the distractors are modeled independently. Our proposed method gets rid of the multi-choice framework, and can leverage amount of other resources. Taking SQuAD as an example, we can take a context, one of its question and the corresponding answer as a positive instance for our classification with golden label 1. In this way many QA dataset can be used to enhance RACE. Experimental results show that single-choice model performs better than multi-choice models, in addition by transferring knowledge from other QA dataset, our single model achieves 90.7% and ensemble model achieves 91.4%, both are the best score on the leaderboard.

2 Task Description

Multi-choice MRC (MMRC) can be represented as a triple $< P, Q, A >$, where $P = s_1, s_2, ..., s_m$ is an article consist of multiple sentences s, Q is a question asked upon the article and $A = \{A_1, ...A_n\}$ is a set of candidate answers. Only one answer in A is correct and others are distractors. The purpose of the MMRC is to select the right one. RACE is one kind of MMRC task,

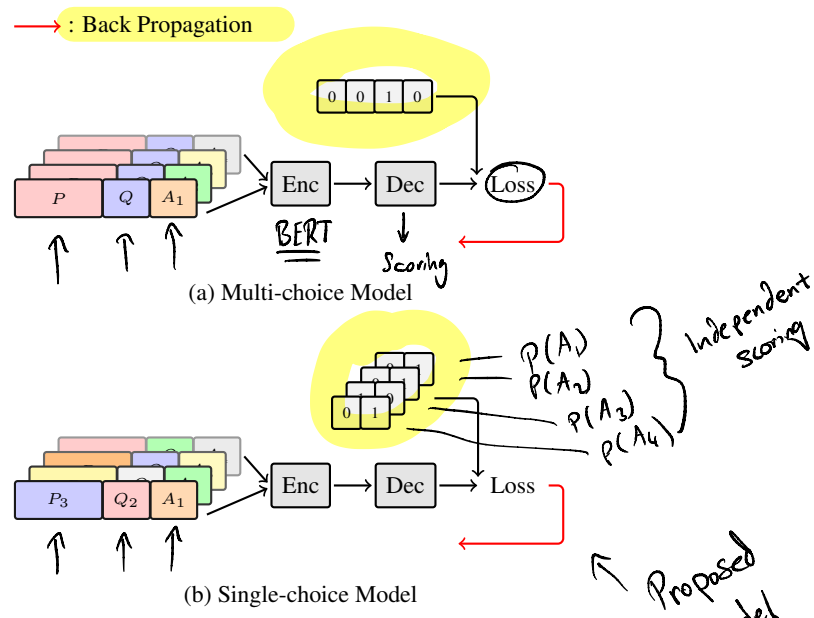


Figure 1: An overview of Standard Model and Single-choice Model.

which is created by domain experts to test students' reading comprehension skills, consequently requiring non-trivial reasoning techniques. Each article in RACE has several questions and the questions always have 4 candidate answers, one answer and three distractors.

3 Methods

Previous works have verified the effectiveness of Pretrained language models such as BERT, XLNet,Roberta and Albert in Multi-choice MRC tasks. Pretrained language models are used as encoder to get the global context representation. After that, a decoder is employed to find the correct answer given all the information contained in the global representation. Let P, Q, and $\{A_1, ...A_n\}$ denote the passage, question and option set separately. The input of Pretrained encoder is defined as $(P \oplus Q \oplus A_i)$, the concatenation of P, Q and A_i one of the option in candidate set. Moreover, for the same question, the inputs with different options are concatenated together as a complete training sample, which is more intuitive and similar to human that select the correct answer compared with other options. After encoding all the inputs for a single question, the contextual representations $T = \{T_{CLS1}, ..., T_{CLSn}\}$ is used to classify which is the correct answer given passage and question(see Figure2(a)). An single full connection layer is added to compute the probability

$p(\{A_1, ...A_n\}|P, Q)$ for all the answers and the ground truth y is the index of correct answer in candidates. $T \in \mathbb{R}^{n \times h}$, n denotes the number of the options in candidate set. We define the score to be:

$$p(\{A_1, ...A_n\}|P, Q) = \sigma(WT + b) \quad (1)$$

where $W \in \mathbb{R}^{h \times 1}$ is the weight and b is the bias. Parameter matrices are finetuned based on pre-trained language model with the cross entropy loss function which is formulated as:

$$\text{loss} = -\sum y \log(p) \quad (2)$$

3.1 Single-choice Model

As all input sequences with the same passage and question are tied together, each training sample contain much duplicate content. For example, the passage with multiple sentences repeat n times in a single training sample which may degrade the diversity in each training step. Moreover, this method need to fix the data format that each question must have the same number of options which is also inconvenient to take advantage of other MRC datasets.

Alternatively, we reconstruct the multi-choice to single-choice. We just need to distinguish whether the answer is correct without considering other options in the candidate set. By this way, we keep the diversity in training batches and relax the constraints on multi-choice framework.

Instead of concatenate all inputs with the same question together, we just encode a single input and use its contextual representations T_{CLSi} to classify whether the answer is correct (see Figure2(b)). The ground truth is $y \in \{0, 1\}$. Thus we re-define the score $g(P, Q, A_i)$ as:

$$g(P, Q, A_i) = \sigma(WT_{CLSi} + b) \quad (3)$$

where $W \in \mathbb{R}^{h \times label}$. Correspondingly, the cross entropy loss function can be re-formulated as:

$$\text{loss} = -\sum y \log(g(P, Q, A_i)) + (1 - y) \log(1 - g(P, Q, A_i)) \quad (4)$$

In the end, to get the correct answers, we select the top-n answers with respect to score. Here n denotes the number of correct answers. E.g., n=1 in RACE.

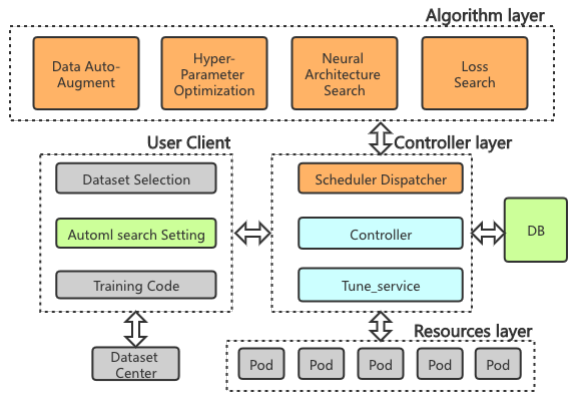


Figure 2: Overview of our AutoML architecture.

3.2 Transfer Learning

In this section, We propose a simple yet effective strategy to transfer knowledge from other QA dataset. As the single-choice model relax the constraints on multi-choice framework, more QA datasets such as SQuAD2.0, ARC, CoQA and DREAM can be used to enhance RACE. It consists of three steps:

(1) we preprocess data with different formats to the same input type as mentioned in section 3. For multiple-choice MRC datasets, like DREAM and ARC, we concatenate each option with corresponding context and question. And for extractive MRC datasets like SQuAD2.0 and CoQA, we take the context (passage or dialog), one of its question and corresponding answer as a positive instance for the binary classification.

(2) We collect and corrupt the preprocessed data from different QA datasets and then train the binary classification on this mixed data. We find that the model benefits a lot from the large amount of MRC datasets.

(3) Finally, we further finetune the model from step 2 on the raw RACE data to adapt the model parameters to the task.

4 Tencent TI-ONE Platform[1]

Our systems are built on TI-ONE, which is a deep learning platform built on the Tencent GPU Cloud. It is an industrial platform with advanced technologies and rich features including popular deep learning frameworks, automated machine learning, large scale distributed training as well as ser-

[1]https://cloud.tencent.com/product/tione

Dataset Specs

Models	RACE	DREAM	SQuAD2.0	CoQA	ARC	Crawl
#Article	27,933	6,444	130,319	8,399	-	-
#Question	97,687	10,197	-	127,000	7787	-
#Answer per Question	4	3	1	-	4	-
#Word per Article	321.9	85.9	-	271	-	-
#Instance	351,464	15,470	86,835	90,000	20,784	446,095

Table 2: Details of different MRC resources. "#Instance" refers to the number of true training samples built by different resources.

vice platforms. We adopt the AutoML algorithm to select better hyper parameters and accelerate the process by distributed training.

4.1 AutoML

Finetuning pretrained language models on downstream tasks is sensitive to the selection of hyper parameters. A good set of hyper parameter affects the final performance to a great extent. However, it is impossible to manually search an optimize set from the huge amount of hyper parameter combinations. To alleviate this problem, we take advantage of automated machine learning (AutoML)(He et al., 2019; Zöller and Huber, 2019; Elshawi et al., 2019) to automatically adapt hyper parameters.

Our AutoML system named TianFeng is a lightweight, extensible, and easy-to-use framework. TianFeng incorporates most current state-of-the-art algorithms and can make good use of resources. It consists of three parts, internal layers, algorithm layer, controller layer and resources layer, as shown in Figure2.

(1) Algorithm layer. The algorithm layer has 4 sub modules, which are used to search model parameters from different perspectives.

(2) Controller layer. Responsible for docking Client, issuing algorithm logic and exception handling.

(3) Resources layer. Effective management of resource pools.

In general, the controller layer receives request from client and select a proper algorithm from algorithm layer. Then uses the idle GPU computing resources in the resource pool to perform multiple tasks in parallel.

4.2 Distributed Training

Due to the huge amount parameters of pretrained language models, it is very time-consuming to conduct AutoML training. We take advantage of the distributed training techniques on Tencent Cloud TI-ONE platform to make the training more efficient. The mixed precision training is used, which greatly accelerate training on single-machine. When training on multiple machines, TI-ONE's fast communication framework can fully leverage more GPUs. Two main advantages on multi-machine communication are: 1) Use optimized all-reduce algorithm and multi-stream to make full use of the bandwidth of VPC network 2) Support gradient fusion of multiple strategies to improve communication efficiency. Our best model is trained on 4 machines with 32 V100 GPUs. The training time can be shortened to 33 percent of the original single machine with 8 cards.

5 Experiments

5.1 Dataset

RACE RACE (Lai et al., 2017) is dataset collected from middle and high school Englis exams in China. RACE has a wide variety of question type such as summarization, inference, deduction and context matching. It contains articles from multiple domains (i.e. news, ads, story) and most of the questions need reasoning.

In the transfer learning stage, we also consider other MRC tasks. Specifically, we consider SQuAD2.0 (Rajpurkar et al., 2016), ARC (Clark et al., 2018), CoQA (Reddy et al., 2019) and DREAM (Sun et al., 2019). We give a brief description of these datasets.

SQuAD2.0 and CoQA SQuAD2.0 and CoQA are extractive MRC tasks, the articles of which are wiki passages and dialogs. Their questions do not have candidate answers, instead participants are asked to extract the answer from the passage.

ARC ARC is the largest public-domain multiple-choice dataset that consist of natural and grade-school science question. It is partitioned into a Challenge Set and an Easy set.

Models	Test
Roberta (Liu et al., 2019)	83.2
ALBERT (single) (Lan et al., 2019)	86.5
ALBERT (ensemble) (Lan et al., 2019)	89.4
ALBERT + DUMA (single) (Zhu et al., 2020)	88.0
ALBERT + DUMA (ensemble) (Zhu et al., 2020)	89.8
Megatron-BERT (single) (Shoeybi et al., 2019)	**89.5**
Megatron-BERT (ensemble) (Shoeybi et al., 2019)	**90.9**
ALBERT baseline	87.1
ALBERT single-choice	87.9
+ transfer learning	88.3
+ AutoML	**90.0**
+ Crawled corpus	**90.7**
Ensemble	**91.4**

Table 3: Results on RACE dataset.

DREAM DREAM is multiple-choice dialogue-based Reading comprehension examination dataset. It article is a dialog and each question has only three options.

Although we have transferred as much data as we can, the MMRC task still suffers data insufficiency problem. Thus we crawl different kind of MRC data from website. Table 2 lists all the resources we use.

5.2 Experimental Settings

Our implementation was based on Transformers[2]. We use the ALBERT-xxlarge as encoder. For hyper parameters, we follow Table 15 in (Lan et al., 2019), except that we set the learning rate to 1e-5 and the warmup steps to 2000. Because we find this is better for the huggingface ALBERT-xxlarge model. After adding the other resources, we do not use a fixed "Training Steps", the training steps after two epochs and the warm up step is 10% of the total training steps. All the models are trained on 8 nVidia V100 GPUs. The training takes about 2 days.

Baseline Our baseline is the original huggingface ALBERT-xxlarge model with the default multi-choice strategy. The hyper parameters follow the description above. In addition, we compare our model with many other public results from both papers or the leaderboard.

5.3 Results

Table 3 shows the results of our models and the baselines. The top part of the table lists the results from the current leaderboard [3] and papers. Megatron-BERT (Shoeybi et al., 2019) achieves the best single and ensemble results. It is a variant of BERT(Devlin et al., 2018) with 3.9 billion parameters which is almost 40 times bigger than ALBERT-xxlarge.

The results of our models are listed below the table. Our ALBERT baseline yields better result than original ALBERT due to the different choice of hyper parameters illustrated in 5.2 showing that the task is sensitive to hyper parameters. Compared with the baseline, our single-choice model achieves 0.8 more score, which shows that single-choice is better than multi-choice under the ALBERT-xxlarge model. After transferring knowledge from other MRC dataset, we get another 0.4 more score. With the help of autoML, our single model achieves 90% which surpasses Megatron-BERT (Shoeybi et al., 2019) and becomes the new state-of-the-art single model results. When adding the web crawl corpus into transfer learning, our single model get the final score as high as 90.7%. This illustrates that single-choice model is easy to incorporate other resources and we achieve this by a simple transfer learning strategy. Our ensemble model gets the best score of 91.4%.

6 Conclusion

In this paper, we propose a single-choice model for MMRC that consider the options separately. Experiments results demonstrate that our method achieves significantly improvements and by taking advantage of other MRC datasets, we achieve a new state-of-the-art performance. We plan to consider the difference between two methods and if we can combine them together in future study.

[2]https://github.com/huggingface/transformers

[3]http://www.qizhexie.com/data/RACE_leaderboard.html

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, A. Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

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

Radwa Elshawi, Mohamed Maher, and Sherif Sakr. 2019. Automated machine learning: State-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287*.

Xin He, Kaiyong Zhao, and Xiaowen Chu. 2019. Automl: A survey of the state-of-the-art. *arXiv preprint arXiv:1908.00709*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option comparison network for multiple-choice reading comprehension.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. Dcmn+: Dual co-matching network for multi-choice reading comprehension.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Duma: Reading comprehension with transposition thinking.

Marc-André Zöller and Marco F Huber. 2019. Survey on automated machine learning. *arXiv preprint arXiv:1904.12054*, 9.