Explore whether the state-of-the-art pre-trained models in NLP can achieve general reading and reasoning abilities

NTU-AI6127: Project Final Report

Junyu Yin
School of Computer Science and Engineering
Nanyang Technological University
s210027@e.ntu.edu.sg

Abstract

Machine reading comprehension (MRC) is a fundamental task in Natural Language Processing (NLP) for testing the capability of natural language understanding (NLU). It seeks to construct an intelligence agent which is capable of understanding text as people. With the recent rising of unsupervised representation learning in NLP, large-scale pre-training models have shown saturated performance on most of the existing simple MRC datasets. It is time to challenge these models with more complex MRC tasks requiring more comprehensive understanding and reasoning of text. In this project, I examined the reading and reasoning abilities of various pre-training models by conducting experiments on recently introduced datasets requiring high-level logical reasoning. The experiment results show that they struggle on these datasets with poor performance, indicating the state-of-the-art machine readers still fall far behind human performance.

1 Introduction

As far as we humans are concerned, in order to achieve good performance on reading comprehension tasks, reading through a given passage and understanding what it says are just the preliminary requirements. We can see that the texts in the reading comprehension tests from the elementary school level to the postgraduate examination level gradually changed from simple to complex, and the question types also changed from simple text extraction to complex logical reasoning. And in this process, we humans have developed some general reading comprehension abilities through continuous learning and training. As a result, it is natural to ask that whether a computer, or specifically an AI system, can also acquire this combination of natural language understanding and logical reasoning.

Machine reading comprehension (MRC) [1, 2] is such a challenging but hot sub-field in natural language processing (NLP), which is very useful for multiple downstream tasks such as open domain question answering [3] and information retrieval [4]. Based on Chen et al. [5], machine reading comprehension is the machine's ability to read a text, process it, understand its meaning, and answer related questions. It is also a useful benchmark to evaluate natural language understanding (NLU) [6] of machines.

Just like the language tests of humans, the most common way to test whether an intelligent agent (a person or an AI system) can fully understand a piece of text is to require her/him/it to answer related questions according to the given text. This task is defined as Question Answering (QA) [7] formally. In a typical task setting [1], an intelligent system is given a passage (context), a question, and asked to select a most appropriate answer from a list of candidate answers or to generate an answer directly.

Due to the persuasiveness and convenience of QA-based evaluation, almost every dataset in the field of MRC is constructed in this format. Two prominent datasets in early period were the MCTest [8] with 500 fictional stories and 2000 questions and the ProcessBank [9] with 585 questions over 200 paragraphs related to biological processes. In 2015, the introduction of large datasets such as CNN / Daily Mail [2] and SQuAD [10] opened a new window in the MRC field by allowing the development of deep models.

However, with the booming of self-training methods in NLP, the emerging large-scale pre-trained language models have brought significant performance gains over the past simple MRC datasets. In fact, BERT [11] has outperformed humans on those datasets, such as SQuAD [10] and MCTest [8]. And recently proposed pre-trained models, such as RoBERTa [12], XLNet [13] and ALBERT [14], further surpass BERT in performance. Now it seems the time to challenge these large-scale pre-trained language models with more complex datasets.

In this project, I mainly focused on the reasoning-based tasks in machine reading comprehension. More specifically, I introduced two highly challenging datasets, ReClor [15] and LogiQA [16], which aim at testing the logical reasoning ability of models. This ability is a significant component of human intelligence and is essential in negotiation, debate and writing etc. To demonstrate the hardness of these datasets, I list two examples in ReClor (See Fig. 1) to express how humans would solve such questions.

Context:
If the purpose of laws is to contribute to people's happiness, we have a basis for criticizing existing laws as well as proposing new laws. Hence, if that is not the purpose, then we have no basis for the evaluation of existing laws, from which we must conclude that existing laws acquire legitimacy simply because they are the laws

Question: The reasoning in the argument is flawed in that the argument

Options:
A. takes a sufficient condition for a state of affairs to be a necessary condition for it
B. draws a conclusion about how the world actually is on the basis of claims about how it should be
C. infers a causal relationship from the mere presence of a correlation
D. trades on the use of a term in one sense in a premise and in a different sense in the conclusion

Answer: A

Reasoning Process of Humans:
We may first look at the question to understand the specific task of the question – identify a flaw. We then analyze the argument in the context. The conclusion 'existing laws acquire legitimacy simply because they are the laws.' is based on the argument (*purpose is **NOT** happiness*) → (***NOT** basis for criticizing laws*), which is obtained from the first statement: (*purpose is happiness*) → (*basis for criticizing laws*). However, we know $\neg A \rightarrow \neg B$ cannot be obtained from $A \rightarrow B$. Therefore, we should choose option A that describes this flaw. The distractors here are different types of reasoning flaws. Prior knowledge of basic logical rules is needed to correctly answer this question.

Context:
Psychologist: Phonemic awareness, or the knowledge that spoken language can be broken into component sounds, is essential for learning to read an alphabetic language. But one also needs to learn how sounds are symbolically represented by means of letters; otherwise, phonemic awareness will not translate into the ability to read an alphabetic language. Yet many children who are taught by the whole-language method, which emphasizes the ways words sound, learn to read alphabetic languages.

Question: Which one of the following can be properly inferred from the psychologist's statements?

Options:
A. The whole-language method invariably succeeds in teaching awareness of how spoken language can be broken into component sounds.
B. Some children who are taught by the whole-language method are not prevented from learning how sounds are represented by means of letters.
C. The whole-language method succeeds in teaching many children how to represent sounds symbolically by means of letters.
D. When the whole-language method succeeds in teaching someone how to represent sounds by means of letters, that person acquires the ability to read an alphabetic language.

Answer: B

Reasoning Process of Humans:
Looking at the question and we know that it is asking about implication. From the first two sentences in context, we know that there are two necessary conditions to *read an alphabetic language*: *phonemic awareness* and *symbolic letters*. We also learn [(***NOT** symbolic letters*) ***AND*** (*phonemic awareness*)] $\nrightarrow$ *read an alphabetic language* (denoted as Formula 1). The last sentence in the context says that many children are taught by the whole-language method to learn a language. As for option A, from the context, we only know the whole language method works for 'many' children, which cannot be inferred to 'invariably' works. As for option B, combing three sentences in the context, we know that the whole-language method meets the two necessary conditions to learn a language, especially the last sentence mentions 'learn to read alphabetic languages'. Children learn to read alphabetic languages means that they must recognize symbolic letters that represent sound because *symbolic letters* is a necessary condition of *read an alphabetic language*; otherwise, they cannot read because of Formula 1 mentioned above. Therefore, option B is correct. As for option C, from the context we only know the whole-language method teaches *phonemic awareness* and *read an alphabetic language*. *Symbolic letters* may be taught by other methods, so C is wrong. As for D, similar to C, *symbolic letters* may be taught by other methods and we also cannot obtain: *symbolic letters* → *read an alphabetic language*.

Figure 1: Two examples in ReClor showing how humans would solve the questions.

Then I conducted experiments to see whether the current state-of-the-art pre-trained models, which outperform humans on simple MRC datasets, can also achieve good performance when facing such hard datasets targeting logical reasoning. Experimental results demonstrate a significant gap between machine and human ceiling performance. But since the model architecture used here is just a simple combination of a pre-trained model plus a softmax classification layer, how to increase the metrics on logical reasoning datasets is still a potentially promising research direction.

2 Related Work

2.1 Machine Reading Comprehension

MRC is a long-standing goal of NLU that aims to teach a machine to read and comprehend textual data [1]. As one of the major and challenging problems of NLP concerned with question answering, semantic analysis, and reasoning [17], MRC stimulates great research interests in the last decade. Early MRC task was simplified as requiring systems to return a sentence that contains the right answer. The systems are based on rule-based heuristic methods, such as bag-of-words approaches [18], and manually generated rules [19]. After the introduction of deep neural networks and effective architecture like attention mechanisms in NLP [20, 2], MRC becomes a hot research topic recently and experiences rapid development. Recent Large-scale pre-trained models [11, 12, 14, 13] lead to a new paradise of contextualized language representations. It greatly strengthened the capacity of language encoder, the benchmark results of MRC were boosted remarkably, which stimulated the progress towards more complex reading, comprehension, and reasoning systems [21]. As a result, the researches of MRC become closer to human cognition and real-world applications.

2.2 Logical Reasoning in NLP

One important aspect of human reading comprehension and question answering is logical reasoning, which was also one of the main research topics of early AI [22, 23]. Natural Language Inference (NLI), also known as recognizing textual entailment [24], is a typical task requiring logical reasoning, which aims to construct models to take a pair of sentence as input and classify their relationship types, i.e., ENTAILMENT, NEUTRAL, or CONTRADICTION. SNLI [25], MultiNLI [26], QNLI [27] and SciTail [28] are widely used benchmarks in evaluating the performance of NLI. However, this task only focuses on sentence-level logical relationship reasoning and the relationships are limited to only a few types. Another task related to logical reasoning in NLP is argument reasoning comprehension task introduced by [29] with a dataset of this task. Given an argument with a claim and a premise, this task aims to select the correct implicit warrant from two options. Although the task is on passage-level logical reasoning, it is limited to only one logical reasoning type, i.e., identifying warrants.

In recent years, an increasing number of tasks and datasets have been introduced targeting on complicated logical reasoning of text in NLP area. Several multiple-choice question answering datasets, which need to select an answer from candidate options given a context and a question, have been proposed for promoting the development of logical reasoning. ReClor [15] and LogiQA [16] are such the datasets which integrate various logical reasoning types into reading comprehension, with the aim to promote the development of models in logical reasoning not only from sentence-level to passage-level, but also from simple logical reasoning types to the complicated diverse ones. Compared with factual question answering [10], lexical overlap between the paragraph and the candidate answers plays a relatively less important role. Compared with commonsense reading comprehension [30], such questions do not rely heavily on external knowledge. In this project, I conduct experiments on these two datasets for investigating logical reasoning skills of the large-scale pre-trained language models.

2.3 Pre-trained Language Model

Language modeling is the foundation of deep learning methods for natural language processing. How to learn good word representations has been an active research area, and many methods were proposed for decades. There are two main approaches: static embedding and contextual embedding. All meanings of a word are represented with a fixed vector in static embeddings, while contextual embeddings move beyond word-level semantics and represent each word considering its context. The most used approach nowadays is contextual embedding method, including Word2Vec [31], GloVe [32], ELMo [33], GPT [34], BERT [11] and so on.

Word2Vec and GloVe use sliding window to cut fixed-length sentence fragments as the input to learn word embeddings. As a result, they cannot capture the complete sentence-level context. It is a common sense that sentence is the least unit that delivers complete meaning as human uses language and sentence-level information especially matters when we have to handle passages in MRC tasks, where the passage always consists of a lot of sentences. In other words, MRC, as well as other application tasks of NLP, needs a sentence-level encoder, to represent sentences into embeddings, so as to capture the deep and contextualized sentence-level information.

ELMo, GPT, BERT and nearly all other latest pre-trained contextualized language models use the whole sentence as the input to learn word embeddings, so they can inject the sentence-level context into learned word embeddings. The ELMo method obtains the contextualized embeddings by a 2-layer Bi-LSTM, while BERT and GPT are bi-directional and uni-directional transformer-based [35] language models, respectively. These models are usually pre-trained on huge unlabeled corpuses firstly and the number of parameters of them is very large as well. As a consensus of limited computing resources, the common practice is to fine-tune the model using task-specific data after the public pre-trained sources. And that's why people refer to them as the large-scale pre-trained models.

3 Methdology

3.1 Formal Description

In this project, I study the problem of logical reasoning of text on the multiple-choice question answering (MCQA) task.

The task can be described formally as following: Given the context C, the question Q and a list of candidate answers $A = \{A_1, A_2, \ldots, A_n\}$, the multiple-choice task is to select the correct answer A_i from A by learning the function F such that $A_i = F(C, Q, A)$.

3.2 Solution

Generally speaking, I use a combination of a pre-trained model and a softmax classification layer to choose the answer from four options. The context, the question and each candidate answer are concatenated as one sequence and a pre-trained contextualized embedding model is used to encode the sequence for calculating its score.

In detail, I first concatenate the context, the question and each option as the input. Given four candidate answers, four concatenated sequences are constructed by pairing each candidate answer with the context and question. Then I feed the four concatenated sequences respectively to the pre-trained model to get four hidden vectors. After that, the four vectors are feed respectively into a linear layer plus a softmax layer. It generates a probability distribution of the candidate answers and the one with the highest score will be chosen as the result.

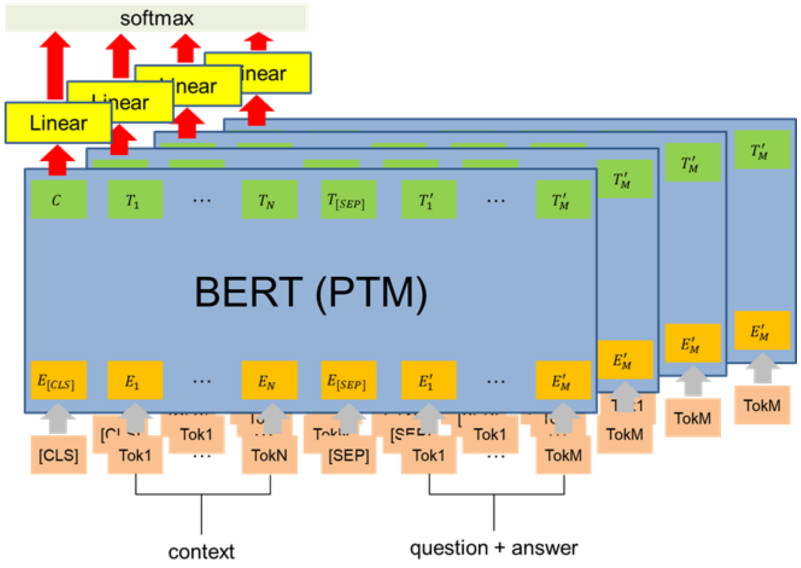


Figure 2: Model Architecture.

Take BERT [14] as an example. It treats the paragraph as sentence A and the concatenation of the question and each candidate as sentence B, and further concatenates them as $[CLS]$ A $[SEP]$ B $[SEP]$ for encoding. Then the final hidden vector of the classification token $[CLS]$ is used as the input to the classifier (a linear layer plus a softmax layer). During fine-tuning, we optimize the model's parameters by maximizing the log-probability of the correct answer.

To give a better sense of my solution, I show the model architecture in Fig. 2. I take BERT as a representative example in this figure and it can be switched to other pre-trained models. Specifically, I experiment with four pre-trained models, BERT, ALBERT [14], RoBERTa [12] and XLNet [13], I briefly introduce them below.

BERT is a bi-directional transformer pre-training over a lot of unlabeled textual data to learn a language representation that can be used to fine-tune on specific machine learning tasks. It uses novel pre-training tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to do the self-supervised training.

ALBERT is a lite BERT using two parameter reduction techniques, factorized embedding parameterization and cross-layer parameter sharing, to lower memory consumption and increase the training speed. And it replaces the next sentence prediction loss with the inter-sentence coherence loss.

RoBERTa is a robustly optimized BERT with more training data, longer training time and bigger batches, which uses a more dynamic sentence masking method and removes the next sentence prediction loss.

XLNet is trained with Permutation Language Modeling (PLM), where all tokens are predicted but in random order, rather than MLM. In addition, it removes NSP and introduces more data for pre-training. Instead of basic transformer, transformer-XL is used as the base architecture in XLNet, which shows good performance even in the absence of permutation-based training.

Model	Loss	2^{nd} Loss	Direction	Encoder Arch.	Tokenizer
BERT	MLM	NSP	Bi	Transformer	WordPiece
ALBERT	MLM	SOP	Bi	Transformer	SentencePiece
RoBERTa	MLM	-	Bi	Transformer	Byte-level BPE
XLNet	PLM	-	Bi	Transformer-XL	SentencePiece

Table 1: A brief summary of the pre-trained models used in this project.

4 Experiments

4.1 Datasets

In this project, I conduct experiments with two datasets, LogiQA and ReClor. They are both MCQ type datasets with 4 options–only one of them is correct. Their details are as follows.

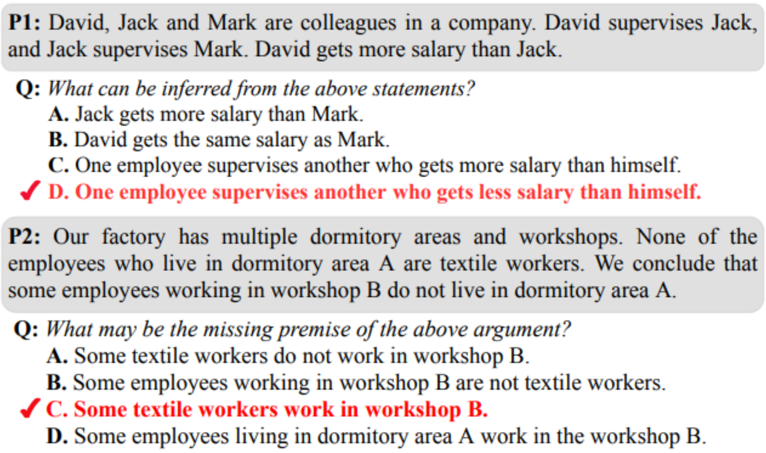
4.1.1 LogiQA

LogiQA [16], which contains 8,678 paragraph-questions pairs, is constructed by collecting the logical comprehension problems from publically available questions of the National Civil Servants Examination of China, which are designed to test the civil servant candidates' critical thinking and problem solving. Its English version is translated from the Chinese original by professional English speakers. To ensure the quality of the translation, the authors of the original dataset also hired proofreaders to do further proofreading.

The dataset covers a wide range of logical reasoning types, including categorical reasoning, conditional reasoning, disjunctive reasoning, and conjunctive reasoning. Thus, a model needs to acquire strong reasoning ability to achieve good performance on this dataset. To better understand this, I list two samples from this dataset in Fig. 3.

4.1.2 ReClor

ReClor [15] is constructed by collecting the reading comprehension problems in some high-level standardized tests, such as GMAT and LSAT. To be specific, it contains 6,138 data points, in which



P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

Q: *What can be inferred from the above statements?*

A. Jack gets more salary than Mark.
B. David gets the same salary as Mark.
C. One employee supervises another who gets more salary than himself.
✓ **D. One employee supervises another who gets less salary than himself.**

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: *What may be the missing premise of the above argument?*

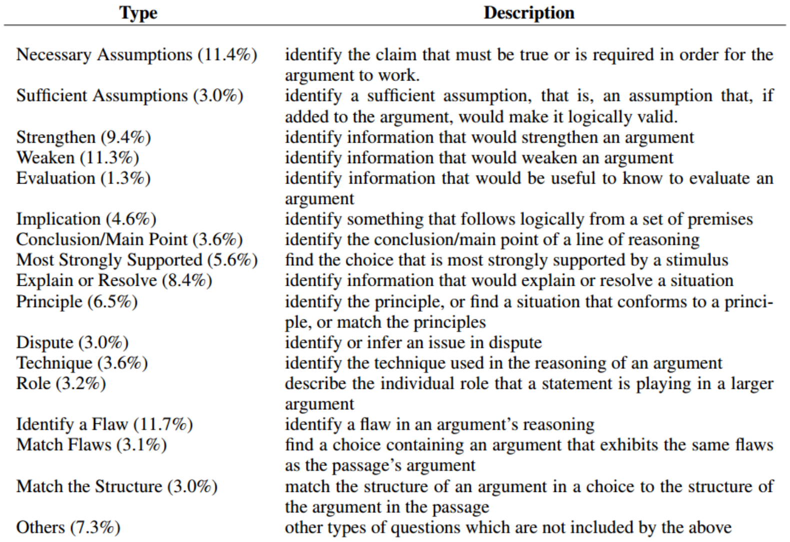
A. Some textile workers do not work in workshop B.
B. Some employees working in workshop B are not textile workers.
✓ **C. Some textile workers work in workshop B.**
D. Some employees living in dormitory area A work in the workshop B.

Figure 3: Two examples in the LogiQA dataset.

91.22% are from actual exams of GMAT and LSAT while others are from high-quality practice exams.

The texts in the ReClor dataset cover large vocabularies (nearly 26,576) as well as complex grammatical structures. In ReClor, every sentence in the contexts is useful when selecting the correct answer. And it covers much more types of reasoning skills than LogiQA whose percentages and descriptions are shown in Fig. 4. This means that high-level reading comprehension ability and high-level logical reasoning skills are required to perform well on this dataset. To better understand this, I show a sample from this dataset in Fig. 5.

What's more, the authors of the original dataset leveraged some techniques to find out the biased data points in ReClor and divided the whole dataset into two subsets named EASY and HARD respectively.

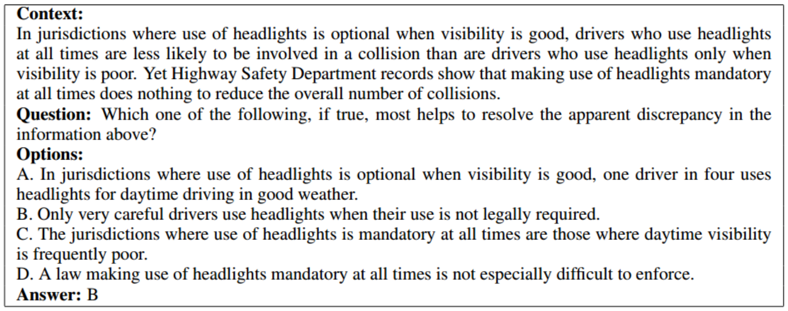


Type	**Description**
Necessary Assumptions (11.4%)	identify the claim that must be true or is required in order for the argument to work.
Sufficient Assumptions (3.0%)	identify a sufficient assumption, that is, an assumption that, if added to the argument, would make it logically valid.
Strengthen (9.4%)	identify information that would strengthen an argument
Weaken (11.3%)	identify information that would weaken an argument
Evaluation (1.3%)	identify information that would be useful to know to evaluate an argument
Implication (4.6%)	identify something that follows logically from a set of premises
Conclusion/Main Point (3.6%)	identify the conclusion/main point of a line of reasoning
Most Strongly Supported (5.6%)	find the choice that is most strongly supported by a stimulus
Explain or Resolve (8.4%)	identify information that would explain or resolve a situation
Principle (6.5%)	identify the principle, or find a situation that conforms to a principle, or match the principles
Dispute (3.0%)	identify or infer an issue in dispute
Technique (3.6%)	identify the technique used in the reasoning of an argument
Role (3.2%)	describe the individual role that a statement is playing in a larger argument
Identify a Flaw (11.7%)	identify a flaw in an argument's reasoning
Match Flaws (3.1%)	find a choice containing an argument that exhibits the same flaws as the passage's argument
Match the Structure (3.0%)	match the structure of an argument in a choice to the structure of the argument in the passage
Others (7.3%)	other types of questions which are not included by the above

Figure 4: The percentage and description of each logical reasoning type.

4.2 Evaluation method

For the MCQA task, we can just use accuracy, i.e., the proportion of correctly answered questions, as our evaluation metric.



Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Figure 5: An example in the ReClor dataset.

4.3 Experimental details

I take BERT-large, RoBERTa-large, XLNet-large and ALBERT-xxlarge-v2 respectively as the backbone model. For implementation, I modify the codes in HuggingFace [36] to complete this MCQA task. For the training of RoBERTa-large and ALBERT-xxlarge-v2, I take AdamW [37] with β_1 = 0.9 and β_2 = 0.98 as the optimizer and the learning rate is set to 1e-5. For the training of BERT-large and XLNet-large, I take AdamW with β_1 = 0.9 and β_2 = 0.999 as the optimizer and the learning rate is set to 2e-5. For all four models, I use a linear learning rate scheduler with 10% warmup proportion and weight decay is set to 0.01. The maximum input sequence length for all models is 256. And each model is fine-tuned for 10 epochs with the batch size 24. I train the models on one 3090 GPU and it takes about 20 hours to finish one experiment.

4.4 Results

In this section, BERT, XLNet, RoBERTa and ALBERT respectively refer to BERT-large, XLNet-large, RoBERTa-large and ALBERT-xxlarge-v2.

The results are shown in Tab. 2.

Model	ReClor				LogiQA	
	Val	Test	Test-E	Test-H	Val	Test
BERT	53.8	49.8	72.0	32.3	33.8	32.1
ALBERT	**62.6**	**59.5**	69.1	**51.9**	32.1	31.9
RoBERTa	62.6	55.6	75.5	40.0	**35.8**	**35.3**
XLNet	62.0	56.0	**75.7**	40.5	34.5	34.4

Table 2: Experimental results of different models (accuracy %). The results in **bold** are the best performance of each column.

4.5 Analysis

From Tab. 2, it can be seen that ALBERT performs best on ReClor and RoBERTa performs best on LogiQA. I think the success of ALBERT on ReClor may attribute to the SOP task used in ALBERT. This task may help the model to capture some useful relationships between sentences which are helpful for doing logical reasoning. However, ALBERT doesn't perform well on LogiQA and RoBERTa performs best on it. I don't know what the exact reason for this.

Besides, the overall accuracy of ReClor is higher than which of LogiQA. The best result of ReClor can achieve about 60% accuracy, which is high enough to some extent since people who have not studied the GMAT and LSAT generally cannot achieve such scores. But the best result of LogiQA can only achieve 35% accuracy, which is just a little better than random guessing (25%). I think this is caused by the dataset itself. The data in ReClor is selected from the GMAT and LSAT, which is constructed by native English speakers, so the data conforms to English language habits. But the data points in LogiQA are constructed by native Chinese speakers. Despite being translated by professional English speakers, they still do not conform to the expression habits of native English speakers. Since the models here are all pre-trained on the English corpus, it is natural that they perform better on ReClor.

5 Conclusion

In this project, I empirically study the different behaviors of state-of-the-art models on two datasets, ReClor and LogiQA, which require logical reasoning and aim to push research progress on logical reasoning in NLP forward from sentence-level to passage-level and from simple logical reasoning to multiple complicated one. Experimental results show that though recent powerful transformer-based pre-trained language models have an excellent ability to exploit the biases in the dataset, they still have difficulty in conducting complicated reasoning over text. To sum up, there is still a significant gap between the powerful transformer-based pre-trained models and human ceiling performance. However, the solution I used is just a simple combination of a pre-trained model plus a softmax classification layer, which means that even without understanding symbolic logic, the models can still achieve a promising performance. There is a long way to equip them with real logical reasoning abilities, and it is a potentially promising research direction. For example, we can combine symbolic logic and neural model or use graph neural network to capture more fine-grained relations.

References

[1] Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. A survey on machine reading comprehension systems. *Natural Language Engineering*, pages 1–50, 2020.

[2] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

[3] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018, 2015.

[4] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[5] Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*, 2016.

[6] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.

[7] Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011.

[8] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203, 2013.

[9] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, 2014.

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

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

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

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

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

[15] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*, 2020.

[16] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

[17] Xin Zhang, An Yang, Sujian Li, and Yizhong Wang. Machine reading comprehension: a literature review. *arXiv preprint arXiv:1907.01686*, 2019.

[18] Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 325–332, 1999.

[19] Ellen Riloff and Michael Thelen. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000.

[20] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[21] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

[22] John McCarthy. Artificial intelligence, logic and formalizing common sense. In *Philosophical logic and artificial intelligence*, pages 161–190. Springer, 1989.

[23] Alain Colmerauer and Philippe Roussel. The birth of prolog. In *History of programming languages—II*, pages 331–367. 1996.

[24] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

[25] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[26] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[27] 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. *arXiv preprint arXiv:1804.07461*, 2018.

[28] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*, 2017.

[30] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*, 2019.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.