Good evening, everyone.

What I do in this project is to explore whether the state-of-the-art pre-trained models in NLP can achieve general reading and reasoning abilities.

Here is the outline of my presentation.

Introduction

As we all know, constructing an intelligent agent which is capable of understanding text as people is an utimate aim in NLP community.

To test the reading ability of a machine, machine reading comprehension task can serve as a useful benchmark.

During the research process, many datasets have been proposed, and the question types also have changed from simple text extraction to complex logical reasoning

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.

It seems the time to challenge these large-scale pre-trained language models with datasets requiring the logical reasoning ability.

This ability is a significant component of human intelligence.

So in this project, I examined the reading and reasoning abilities of various pre-training models by conducting experiments on recently introduced datasets, ReClor and LogiQA, which require high-level logical reasoning.

To demonstrate the hardness of these datasets, I list two examples.

The first one comes from LogiQA and the second one comes from ReClor.

I also show how humans would solve such question.

We can see even for our humans this question is hard to solve.

The overlap between the paragraph and the candidate answers plays a less important role.

We need to fully understand it and use complicated reasoning skills to solve it.

Related work

MRC

MRC is a long-standing goal of NLU that aims to teach a machine to read and comprehend textual data.

Early MRC systems are based on rule-based heuristic methods, such as bag-of-words approaches and manually generated rules.

Now with the introduction of deep neural networks, especially the large-scale pre-trained models, the pre-training plus fine-tuning framework is used everywhere.

This method boosted the benchmark results of MRC, thus stimulated the research progress towards human cognition and real-world applications

LG

Logic reasoning is one important aspect of human reading comprehension, which was also one of the main research topics of early AI.

Natural language inference and argument reasoning comprehension are early representative tasks.

The former takes a pair of sentence as input and classify their relationship types into ENTAILMENT, NEUTRAL, or CONTRADICTION.

The latter aims to choose the right implicit warrant for a claim.

They only focus on sentence-level logical relationship reasoning and the relationships are limited to only a few types.

In recent years, several multiple-choice question answering datasets are proposed.

They require the passage-level reasoning and contain more complicated and diverse reasoning types.

PTM

Language modeling is the foundation for natural language processing.

The early used method includes static embedding, Word2Vec and GloVe.

Static embedding uses a fixed vector to represent a word without using its context.

Word2Vec and GloVe use sliding window to learn word embeddings.

They either don't use contextual information, or they don't have complete sentence-level context.

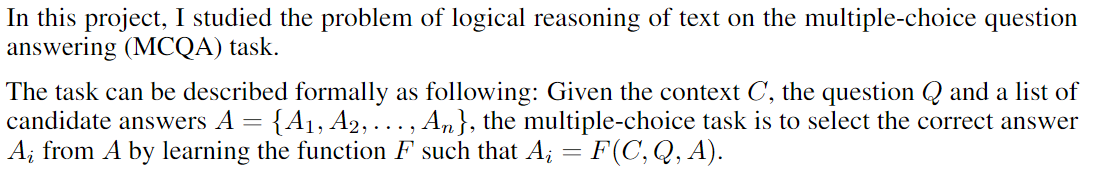
ELMo, GPT, BERT and nearly all latest language models use the whole sentence as the input to learn word embeddings.

So they capture the complete sentence-level context.

These models are usually pre-trained on huge unlabeled data and use transformer-based architecture to learn better representations.

Methodology

FD



Solution

In general, I use a combination of a pre-trained model and a softmax classification layer to do this task.

In details, I first concatenate the context, the question and each option as the input.

Then feed the four concatenated sequences to pre-trained model to get four hidden vectors of the [CLS] token.

The four vectors are then feed into a linear layer plus a softmax layer.

The one with the highest score are choosen as the answer.

To give a better sense of my solution, I show the model architecture here.

The BERT can be change to other pre-trained model.

I experimented with BERT, ALBERT, ROBERTA and XLNet, here is a summary of them.

Results

As mentioned before, I use the following four pre-trained models and experiment on ReClor and LogiQA datasets. ReClor is constructed from GMAT and LSAT, while LogiQA is from the National Civil Servants Examination of China.

They results show that ALBERT performs best on ReClor and Roberta performs best on LogiQA.

Conclusions

To sum up, there is still a significant gap between the powerful transformer-based pre-trained models and human ceiling performance.

Though they performs better than random guess, there is a long way to equip them with real logical reasoning abilities.

However, the model used here is just a simple combination of a pre-trained model plus a softmax classification layer, it is still 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.

That’ all. Thanks for the listening.