Multiple Choice Reading Comprehension with Small Transformer Models

Jiayue Guo, Bo Wu, Yilun Wu

Language Technologies Institute, Carnegie Mellon University

Abstract

Multiple Choice Reading Comprehension is a language comprehension task where given an article and a question about the article, the machine must select the correct answer from a fixed number of options. We re-implement a baseline model suited for English language understanding to answer such multiple choice questions, and explore different options to improve upon the model. Our models and experiments are based on ALBERT, DUMA, and end-to-end memory network, and it is trained and evaluated on the RACE dataset.

Introduction

Machine reading comprehension (MRC) aims to teach machine to answer question automatically according to the content of the input passages/texts. Since reading comprehension requires both understanding of natural language and ability to extract contextual knowledge, it is a hard but meaningful task for machines.

Most existing MRC models are trained on the SQuAD dataset, where the span of answer is marked in the passage. Such models could only handle the questions with expected answers in the form of a segment or short consecutive pieces of the corresponding passage. Multiple choice reading comprehension questions like those present in standardized tests such as TOEFL, SAT, and GRE typically require further reasoning or passage summarization, and shows the insufficiencies of previous models.

Fig. 1: Example of a MC question that requires reasoning [4]

Passage: *Runners in a relay race pass a stick in one direction. However, merchants passed silk, gold, fruit, and glass along the Silk Road in more than one direction. They earned their living by traveling the famous Silk Road. ...* ***The Silk Road was made up of many routes, not one smooth path.*** *They passed through what are now 18 countries. The routes crossed mountains and deserts and had many dangers of hot sun, deep snow and even battles...*

Question: *The Silk Road became less important because _ .*

A. *it was made up of different routes*
B. *silk trading became less popular*
C. ***sea travel provided easier routes***
D. *people needed fewer foreign goods*

Dataset

We choose *RACE: Large-scale ReAding Comprehension Dataset From Examinations* as our target dataset for model training and evaluation. [3]

- Multiple-choice reading comprehension problems from English exams for middle school Chinese students.
- Near 28,000 passages and 100,000 questions generated by English instructors.
- Various carefully designed topics from news, short stories, to ads

Model Architectures

1. **ALBERT** [1]
   - Baseline
   - A transformer architecture based on BERT that shares parameters across layers.
   - The same model complexity can be achieved using 9x less parameters

2. **End-to-End Memory Network** [5]
   i. the module employs hierarchical attention. For each hop, it updates the intermediate embedding as the new query, and attends it with all sentences embeddings. The embedding matrices weights are shared across different hops
   - Refer to Fig. 2

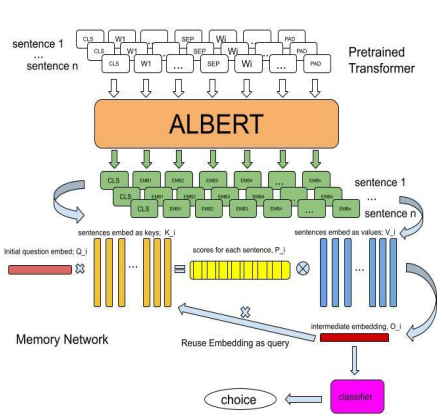

Fig. 2: Architecture of End-to-end Memory Network

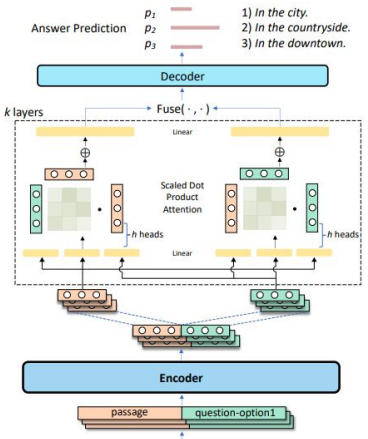

Fig. 3: Architecture of DUMA model

3. **DUMA** [2]
   - DUal Multihead co-Attention
   - The DUMA Model first attends question by answer and passage, and then attends answer by question and passage, and take the concatenation of two contexts as final embedding.
   - Refer to Fig. 3

4. **Sentence Selection**
   - [4] analyzes that 87% of the questions can be answered using only N=2 sentences from the article as context, so eliminating useless sentences should help
   - Pre-selects key sentences in the article that are likely useful to derive the answer, before passing them into the transformer. Algorithm is from [4].
   - Each sentence is scored by the mean over all words in question/answers of the max over all words in the sentence of the cosine similarity between each pair of words.
   - Sentences with top N (hyperparameter) scores are chosen to be kept in the article

Results

Model	Test acc(RACE-all)	Notes
Baseline ALBERT$_{base}$	64.0	result from [1]
ALBERT$_{base}$+Sentence Selection	63.70	worse than baseline
ALBERT$_{base}$ (hidden sequence)	65.84	max-seq-len=384
ALBERT$_{base}$+DUMA	67.93	max-seq-len=384
ALBERT$_{base}$+DUMA	69.74	max-seq-len=512
ALBERT$_{base}$+HP Tuning	70.96	

Table 5: Results

Many of our models did not perform as well as expected. The sentence selection used in [4] actually worsened our results, and the best results were obtained through using an updated ALBERT checkpoint and hyperparameter tuning. Otherwise, DUMA worked well to increase the accuracy. Overall, most of the models still performed above baseline.

Error Analysis

- During the tokenization, some information about context will be lost due to the length limit.
- Due to implementation issues, we didn't incorporate the temporal embedding and positional embedding components in end-to-end memory network.
- Since each article only has different number of sentences, the batch size is currently set to one to avoid dimension alignment issue in data loading.

Future Work

- Incorporate knowledge bases such as Wikipedia
- Try ALBERT$_{xxlarge}$(ALBERT with more parameters)+DUMA
- Try ensemble methods
- Try dimensional reduction techniques, including PCA

References

[1] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. CoRR, abs/1909.11942, 2019.
[2] Pengfei Zhu, Hai Zhao, and Xiaoguang Li. Dual multi-head co-attention for multi-choice reading comprehension. CoRR, abs/2001.09415, 2020.
[3] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. CoRR , abs/1704.04683, 2017.
[4] Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. DCMN+: dual co-matching network for multi-choice reading comprehension. CoRR, abs/1908.11511, 2019.
[5] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-To-End Memory Networks. CoRR, abs/1503.08895, 2015.