A BERT-based model for Multiple-Choice Reading Comprehension

Kegang Xu Jing Jie Tin Jung Youn Kim

{tosmast, jjtin1, jyk423}@stanford.edu

Stanford University

Overview

Problem

Machine reading comprehension can be applied to a wide range of commercial applications such as the understanding of financial reports, customer service, and healthcare records. We focused on analyzing and improving the accuracy in predicting the correct answers of the automated multiple-choice reading comprehension task on RACE dataset [1].

Approach

Our ensemble model consists of the following models:

- Pre-trained Bidirectional Encoder Representations from Transformers (BERT)
- Easy Data Augmentation (EDA)
- Deep Comatch Network (DCN)

and was able to achieve a 6% increase in test accuracy.

Dataset

Dataset	RACE-Middle			RACE-High		
Subset	Train	Dev	Test	Train	Dev	Test
#Passages	6,409	368	362	18,728	1,021	1,045
#Questions	25,421	1,436	1,436	62,445	3,451	3,498

Data Input

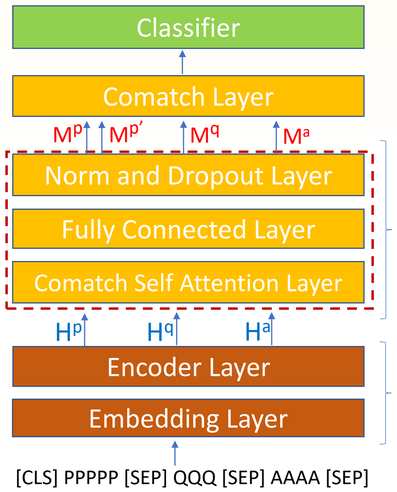
A passage, question, and option are concatenated together with special tokens CLS and SEP as one input sequence. Each of 4 input sequences is then labeled with a correct option number.

Input: [CLS] passage [SEP] question [SEP] option 1 [SEP]
[CLS] passage [SEP] question [SEP] option 2 [SEP]
[CLS] passage [SEP] question [SEP] option 3 [SEP]
[CLS] passage [SEP] question [SEP] option 4 [SEP]

Data Augmentation

- EDA was used to extend the dataset by 10%.
- Synonym Replacement, Random Insertion, Random Swap, and Random Deletion techniques were applied on passages with an augmentation parameter $\alpha = 0.2$

Model

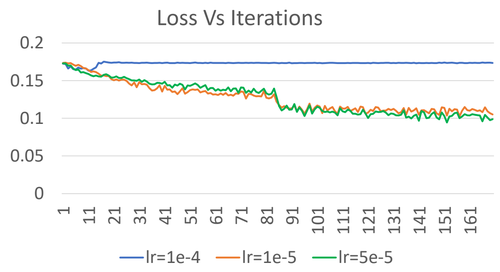


Deep Comatch Network (DCN)

- Output the hidden states, H^p, H^q, H^a at the last layer of BERT.
- Create n Deep Comatch Attention blocks. Each block contains 3 layers.
- In the Comtach layer, we calculate:
  $S^x = RELU([M^x - H^x; M^x \cdot H^x]W_1)$
  $C^x = Maxpooling(S^x)$
  and concatenate all C^x.
- In the last layer, we change 4 outputs as one predict.
- Loss of the model is calculated with Cross Entropy loss.

Experiments

- **Learning rate**: The best LR was 5e-5. LR of 1e-4 was too large that the loss gets stuck.
- **Freeze layers:** Impacts the performance, but speeds up the training and reduces GPU memory.
- **Max sequence length:** The accuracy with 450 length improved by 2.1% compared to 320.
- **Regularization**: 0.1, 0.01, and 0.001 had no obvious difference in performance.



Freeze Layers	Dev Accuracy
No Freeze layers	62.2%
Freeze first 1 layer	61.2%
Freeze first 3 layers	60.4%
Freeze first 5 layers	57.9%
Freeze first 7 layers	55.8%

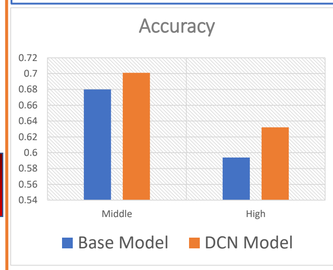
Model	Accuracy
Base BERT	61.6%
Fine-tuned Base BERT	62.6%
Large BERT	65.0%
Easy Data Augmentation	65.6%
Deep Comatch Network	66.2%
Ensemble of models	67.9%

- Baseline BERT was constructed with a pretrained model with default parameters.
- Our models are implemented upon Large BERT.
  - Through EDA, performance improved by 0.6%.
  - Deep Comatch Network resulted in 1.2% improvement.
  - After ensembling, we gained 2.9% more accuracy.

Analysis

1. Questions: "In which part of a newspaper can you likely read this passage?"
Options: "***Today's News***", "Culture", "Entrainment", "*Science*"
Passage: "Three cattle farms in Andong...were infected with ...disease,Nov.2 2010,Thursday....On Monday, today showed that all … infected with the disease," an official said. Two newly infected cattle farms...indicating the disease will likely continue... has culled 33,000 animals ...No suspected cases ...

2. Question: From the passage we can know that __
Choice: "*Stewart was depressed at one time*", "**Stewart lost his left arm 22 years ago**", "Stewart never complained about the unfairness of life", "Stewart was persuaded to kayak through the Grand Canyon"
Passage: For most of his life, the **45**-year-old man has lived with only his right arm. He lost his left arm ... when he was **18**. He became a bitter young man, angry at the unfairness of what had happened, and often got into fights.



DCN correctly answers "***Today's News***" in Q1 while Base BERT chooses "**Science**". It means DCN learns more than Base model, but both of them failed in Q2 which requires simple math calculation (45-18=27).
1) DCN performs better than Base in both high and middle dataset.
2) DCN performance is even much better in long and complex passage in high dataset.

Conclusion and Future works

- Using larger max sequence was helpful although the average length of passage is about 350 words.
- Gain in accuracy with Data Augmentation wasn't dramatic since the # of original training dataset was already huge.
- After applying ensemble model, we gained a total of 1.7% increase in accuracy compared to DCN.
- Our models and datasets were too large that it took a very long time to train and test each experiment. If we had more time, we would have spent more time on parameter tuning with Large model which our implementations are based upon.
- In the future works, we could explore more layers of DCN and use higher max sequence. We could also train on another dataset first and conduct a transfer learning.

References:

- Lai etc,RACE: Large-scale Reading Comprehension Dataset from Examinations. arXiv preprint arXiv:1704.04683.
- Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, Xiang Zhou, Dual Co-Matching Network for Multi-choice Reading Comprehension. arXiv:1901.09381, 2019.
- 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.