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)

Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RACE-Middle</th>
<th>RACE-High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset</td>
<td>Train</td>
<td>Dev</td>
</tr>
<tr>
<td>#Passages</td>
<td>6,409</td>
<td>368</td>
</tr>
<tr>
<td>#Questions</td>
<td>25,421</td>
<td>1,436</td>
</tr>
</tbody>
</table>

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] CLS passage [SEP] question [SEP] option 1 [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 $\alpha = 0.2$

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.

Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base BERT</td>
<td>61.6%</td>
</tr>
<tr>
<td>Fine-tuned Base BERT</td>
<td>62.6%</td>
</tr>
<tr>
<td>Large BERT</td>
<td>65.0%</td>
</tr>
<tr>
<td>Easy Data Augmentation</td>
<td>65.6%</td>
</tr>
<tr>
<td>Deep Comatch Network</td>
<td>66.2%</td>
</tr>
<tr>
<td>Ensemble of models</td>
<td>67.9%</td>
</tr>
</tbody>
</table>

Deep Comatch Network (DCN)
- Output the hidden states, $H_P^0, H_P^1, H_A$ at the last layer of BERT.
- Create $n$ Deep Comatch Attention blocks. Each block contains 3 layers.

In the Comatch block, we calculate:
- $S^x = \text{RELU}(M^x - H^2; H^x, W^j)$
- $C^x = \text{Maxpooling}(S^x)$ 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.

Analysis

1. **Question**: "In which part of a newspaper can you likely read this passage?"
   **Options**: "Today's News", "Culture", "Entertainment", "Science"
   **Passage**: "Three cattle farms in Andong... were infected with __...disease, Nov 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 the context:
   **Context**: “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 walk 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.

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: