Increasing Robustness of DistilBERT QA System with Few Sample Finetuning and Data Augmentation via Back-Translation

Anjali Sukhavasi and Alina Chou
Stanford CS224N Default Final Project

Problem + Background

• Humans can quickly learn new words and generalize information to new contexts or domains by learning the true meaning of a word rather than correlations between words.
• NLP systems often cannot accurately generalize information beyond their training domain because they learn superficial correlations between words rather than understanding their meaning.
• Building NLP systems that are robust to data outside their training distribution is integral to building accurate systems that can interact with natural language in real-world scenarios, which will rarely match training data.
• Our project sought to build upon the existing DistilBERT model to increase robustness on out-of-domain reading comprehension tasks by implementing few sample finetuning and data augmentation via back-translation using different pivot languages.

Methods

1. Trained baseline DistilBERT model on all training data by minimizing our loss function.
2. Implemented few sample finetuning by adjusting hyperparameters to values that generated high performance in Zhang et al.’s research.
3. Implemented data augmentation via back-translation in German, Russian, and Chinese.

Analysis

• Finetuning results show that the number of layers and learning rate will greatly affect the model’s training as well as evaluation performances.
• The training loss, F1 score, and EM score are approximately the same for the baseline model and the data augmentation model indicating little improvement from the data augmentation.
• One potential reason that data augmentation via back-translation does not improve the model is that back-translation generates sentences that are of the same meaning and, usually, sentence structure as the original sentence. While data augmentation via back-translation has proven to improve NMT models that generate text, it might not be useful for our model that has to select the correct span of text for reading comprehension. The information that is fed into the training “answer” and “context” is too similar, which does not lead to a significant difference in answering questions given contexts by selecting spans of text.

Experimental Results

• Training and evaluating results compared across baseline, finetune implementations, and data augmentation as shown in figures 1, 2, and 3.
• The best performing model had 8 layers, 32 batches, 3e-5 learning rate.
• The model that converged the fastest in minimal training time had 4 layers, 32 batches, and a 5e-5 learning rate.

Conclusion

While our findings have shown improved results in few-shot finetuning, data augmentation via back-translation has produced negligible improvement on the reading comprehension task. We believe that the latter is due to the high similarity between the original language and back-translated language, which does not significantly contribute to linguistic understanding.

References


Figure 1: Training Loss

Figure 2: F1 Score

Figure 3: EM Score

Figure 4: EM/F1 Score Table

<table>
<thead>
<tr>
<th>Experiment</th>
<th>EM Score</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>30.628</td>
<td>47.718</td>
</tr>
<tr>
<td>Finetuning #1 (minimal training time)</td>
<td>31.214 (+0.585)</td>
<td>48.440 (+1.723)</td>
</tr>
<tr>
<td>Finetuning #2 (best performing)</td>
<td>34.555 (+3.141)</td>
<td>49.881 (+1.442)</td>
</tr>
<tr>
<td>Data Augmentation</td>
<td>30.65 (+0.002)</td>
<td>47.72 (+0.004)</td>
</tr>
</tbody>
</table>

(values in parenthesis represent improvements compared to the baseline)