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**

Example of Data Augmentation via Back-Translation

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.

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 language understanding**.

References

1. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2019.

2. Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In ICLR, 2020.

3. Amane Sugiyama and Naoki Yoshinaga. Data augmentation using back-translation for context-aware neural machine translation. In Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 35–44, 2019. ACL.

4. Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. In ICLR, 2021.

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 6 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**

Legend
Baseline
Best Performance
Fastest Convergence w/ Minimal Training Time
Data Augmentation via Back-Translation

Figure 1: Training Loss

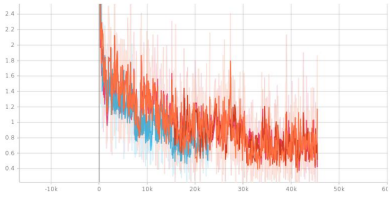

Figure 2: F1 Score

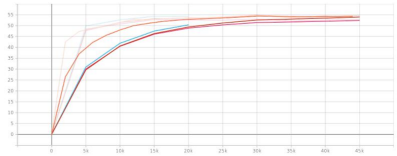

Figure 3: EM Score

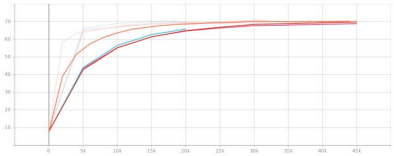

Figure 4: EM/F1 Score Table

Experiment	EM Score	F1 Score
Baseline	30.628	47.716
Finetuning #1(minimal training time)	31.414 (+0.785)	48.440 (+0.723)
Finetuning #2 (best performing)	34.555 (+3.141)	49.881 (+1.442)
Data Augmentation	30.63 (+0.002)	47.72 (+0.004)

(values in parenthesis represent improvements compared to the baseline)