Improving Out-of-Domain Question Answering Performance with Adversarial Training

Stanford CS224N Default Project

Jack Lee
Department of Computer Science
Stanford University
jack9766@stanford.edu

Abstract

In this project, we aim to investigate the effectiveness of adversarial training on improving out-of-domain performance of question answering tasks. We fine-tune a pre-trained transformer model with a variety of adversarial training configurations. We then evaluate and compare the out-of-domain performance between the configurations.

1 Key Information to include

- Mentor: N/A
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

In recent years, we have seen significant progress on natural language understanding problems. However, there is increasing research showing models' performance degrade substantially beyond the training distribution. Investigating ways to improve model robustness is crucial for us to demonstrate models have generalized language understanding similar to humans and guarantee the their effectiveness when deployed in the real world. In this project, we investigate the effectiveness of adversarial training on improving model robustness for question answering (QA) tasks. We show that finetuning a pretrained transformer with adversarial examples generated with Fast Gradient Method (FGM) [1] using in-domain training data consistently improves the out-of-domain performance of the model.

3 Related Work

3.1 BERT

BERT [2] is a stacked bidirectional Transformer encoder that is pretrained on Wikipedia and BooksCorpus, has given state-of-art results on a wide variety of NLP tasks. Models based on BERT and BERT variants are also shown to have better out-of-domain performance compared to other language models [3].

3.2 DistilBert

DistilBert [4] is a BERT variant which reduced the parameters of the original BERT model by 40% while retaining 97% of its language understanding capabilities and being 60% faster by leveraging knowledge distilation [5] during pretraining.

3.3 Adversarial Training

Adversarial Training is a data augmentation techniques in which neural networks are trained on generated adversarial examples in addition to the training data. Fast Gradient Sign Method (FGSM) [6], the Fast Gradient Method (FGM) and Projected Gradient Descent (PGD) [7] are first proposed to perform adversarial attacks on neural networks; those methods are then employed to perform adversarial training. Adversarial Training is shown to produce models that are robust against adversarial attacks and significantly improve in-domain performance for tasks with limited training data [8].

4 Approach

4.1 Task

Our model will perform question answering tasks. The input include a paragraph and a question and the output is the start and end indices indicating the span of text containing the answer within the paragraph. An example of a question, paragraph and answer triplet is shown in Figure 1.

Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

Figure 1: A example sample for performing question answering tasks

4.2 Architecture

We use DistilBert augmented with a dense classifier for answer span prediction as the model architecture. The Question Answering DistilBert implementation is provided by HuggingFaces [9] library.

4.3 Training

During fine-tuning, the pre-trained transformer is trained on both the original input samples and the generated adversarial input samples. The model is trained on original input samples with Cross-Entropy loss defined as follows where y_j^i and s_j^i are ground truth probability distribution and predicted probability distribution respectively, C is the number of categories.

$$L = -\sum_{i}^{T}\sum_{j}^{C} y_j^i \log s_j^i$$

The model is trained on adversarial input samples using the following min-max loss where θ represents the model parameters, x and y represents the input embedding and the corresponding target, r_{adv} represent some input perturbation in the perturbation space S.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[\max_{r_{adv}\in S} L(\theta, x + r_{adv}, y)\right]$$

Intuitively, the loss tries to find a input perturbation to maximize the loss L and model parameters to minimize the perturbed loss simultaneously.

We generate adversarial perturbation r_{adv} using the FGM. Specifically, the adversarial perturbation is defined as follows where ϵ is a hyperparameter.

$$g = \nabla_x L(\theta, x, y)$$

$$r_{adv} = \epsilon \cdot \frac{g}{\|g\|_2}$$

Intuitively, we point the input perturbation in the same direction as the input gradient respect to the loss we want to maximize.

The overall architecture and the adversarial training process is shown in Figure 2.

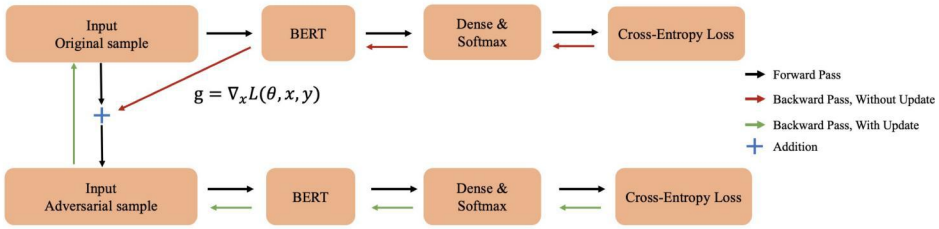

Figure 2: Overall architecture and the adversarial training process of AT-BERT

The adversarial training process is implemented from scratch as a data augmentation step during training.

4.4 Configurations

4.4.1 Baseline

The baseline model is a pretrained DistilBert for Question Answering finetuned using in-domain training data without any adversarial examples.

4.4.2 Adversarial Ratio

Adversarial Ratio denotes the ratio between original training samples and generated adversarial samples. We train models with 4 different adversarial ratios - 4-to-1, 2-to-1, 4-to-3 and 1-to-1. For each ratio, we randomly mix original samples and adversarial samples; new adversarial samples are generated for each epoch.

4.4.3 Baseline Finetuning

In addition to finetuning the pretrained DistilBert with adversarial training, we also investigate the effects of funetuning with adversarial training starting from the baseline model.

4.4.4 Gradient Reuse

To improve memory efficiency, instead of generating adversarial samples from randomly selected training samples, we reused the gradient produced by the current training sample for each training step. The resulting adversarial samples are no longer i.i.d and are strongly correlated to the original samples. However, this method reduce the computation cost and memory usage significantly.

4.4.5 Ensemble

We produce an ensemble model by performing majority voting using the top-3 models in terms of F1 scores. The prediction of the model with the best F1 score are used for tie-breaking.

5 Experiments

5.1 Data

The data is split between in-domain datasets containing Natural Questions [10], NewsQA [11] and SQuAD [12] and out-of-domain datasets containing RelationExtraction [13], DuoRC [14] and RACE [15]. The sizes of each dataset and the overall train, dev, test split in shown in Figure 3.

Dataset	Question Source	Passage Source	Train	dev	Test
		`in-domain datasets`			
SQuAD	Crowdsourced	Wikipedia	50000	10,507	-
NewsQA	Crowdsourced	News articles	50000	4,212	-
Natural Questions	Search logs	Wikipedia	50000	12,836	-
		`oo-domain datasets`			
DuoRC	Crowdsourced	Movie reviews	127	126	1248
RACE	Teachers	Examinations	127	128	419
RelationExtraction	Synthetic	Wikipedia	127	128	2693

Figure 3: Dataset statistics

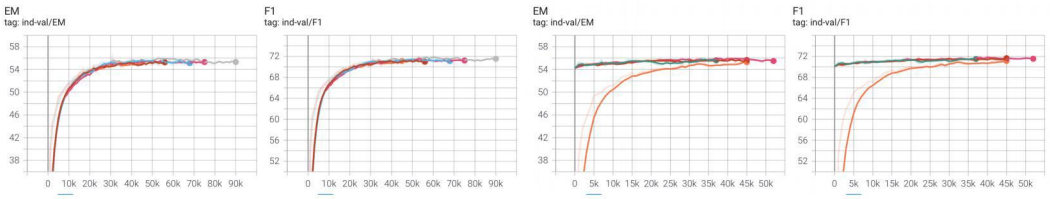

Figure 4: In-domain performance: Adversarial Ratio experiments (Left) with baseline (orange), 4-to-1 (red), 2-to-1 (blue), 4-to-3 (magenta) and 1-to-1 (grey); Baseline Finetuning experiments (Right) with baseline (orange), 4-to-1 (green), 2-to-1 (red) and 4-to-3 (magenta)

5.2 Evaluation method

We measure both in-domain and out-of-domain performance quantitatively using Exact Match (EM) and F1 metrics.

- **Exact Match** is a binary measure (i.e. true/false) of whether of model output (i.e. the answer span prediction) matches the ground truth exactly.
- **F1** is the harmonic mean of precision and recall defined as $2 \cdot \frac{precision \cdot recall}{precision + recall}$.

5.3 Experimental details

We used batch size of 16, learning rate of 3e-5 and the AdamW [16] optimizer across all of our experiments. All models are trained for 3 epochs and models finetuned from baseline are trained for 2 epochs.

5.4 Results

5.4.1 In-domain Performance

As shown in Figure 4, adversarial training did not degrade the in-domain performance across all experiments. The F1 and EM scores of adversarially trained models matched the scores of the baseline model throughout the training process.

5.4.2 Adversarial Ratio

As shown in Figure 5, 2-to-1, 4-to-3, and 1-to-1 adversarial training produced comparable performance while 4-to-1 adversarial training produced slightly worse performance. Increasing the adversarial ratio increases the training time proportionally.

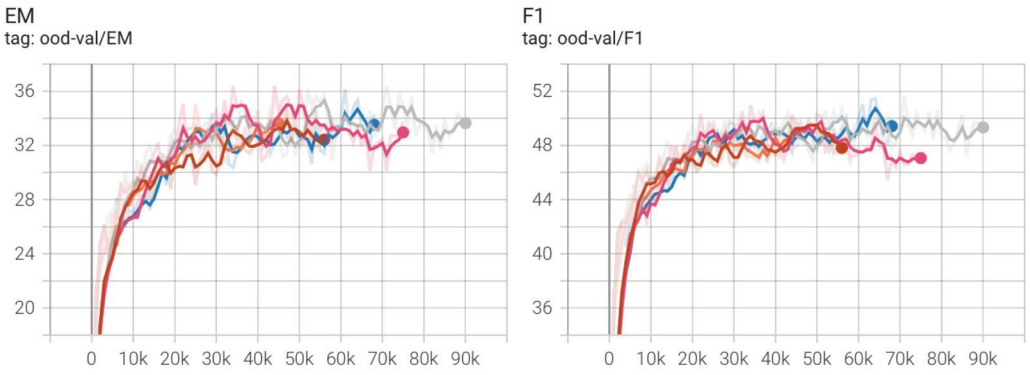

Figure 5: Out-of-domain performance for Adversarial Ratio experiments with baseline (orange), 4-to-1 (red), 2-to-1 (blue), 4-to-3 (magenta) and 1-to-1 (grey)

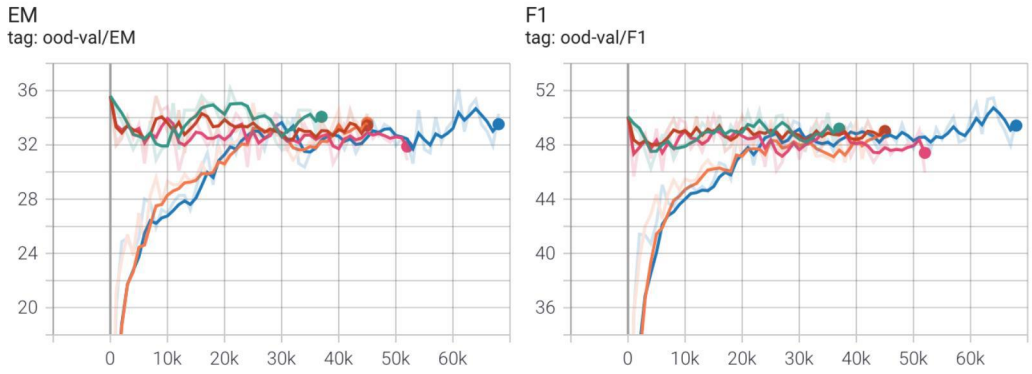

Figure 6: Out-of-domain performance for Baseline Finetuning experiments with baseline (orange), 2-to-1 baseline (blue), 4-to-1 (green), 2-to-1 (red) and 4-to-3 (magenta)

5.4.3 Baseline Finetuning

As shown in Figure 6, finetuning from baseline achieved slightly worse performance compared to finetuning from pretrained DistilBert with significantly lower training time. Baseline funetuning achieves similar performance across different adversarial ratios.

5.4.4 Gradient Reuse

As shown in Figure 7, reusing gradients produces lower performance than using generating adversarial samples from randomized selections. However, reusing gradients reduced the training time by 40%.

5.4.5 Overview

As shown in Fig 8, adversarially trained models consistently improve out-of-domain performance. Without using ensembles, 2-to-1 adversarial training produces the best out-of-domain performance. Ensembling the top-3 models improves the performance moderately even when the model architectures in the ensemble are identical.

Our best model - Ensemble of three of our models with the highest F1 scores - ranked 3rd on the *RobustQA* test leaderboard with 61.307 F1 and 43.165 EM scores and ranked 8th on the *RobustQA* validation leaderboard with 52.893 F1 and 36.649 EM scores. This result exceeded our expectation since our approach did not alter the model architecture; in addition, we performed significantly better on the test set than the validation set which suggests our technique is robust to overfitting.

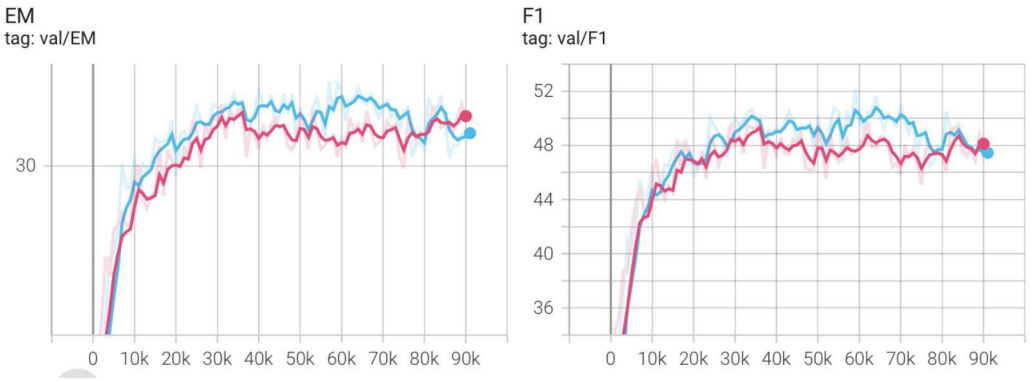

Figure 7: Out-of-domain performance for Gradient Reuse experiments with randomized (blue) and reused (magenta)

Model	Out-of-domain F1(%)	Out-of-domain EM(%)
Baseline	47.10	31.68
4-to-1	50.51	36.13
2-to-1	52.20	36.39
4-to-3	50.94	36.39
1-to-1	51.77	36.39
Baseline Finetuning 4-to-1	50.5	36.13
Baseline Finetuning 2-to-1	49.97	36.39
Baseline Finetuning 4-to-3	50.15	34.82
Baseline Finetuning 1-to-1	50.39	34.55
Gradient Reuse 2-to-1	49.75	34.55
Ensemble	**52.90**	**36.65**

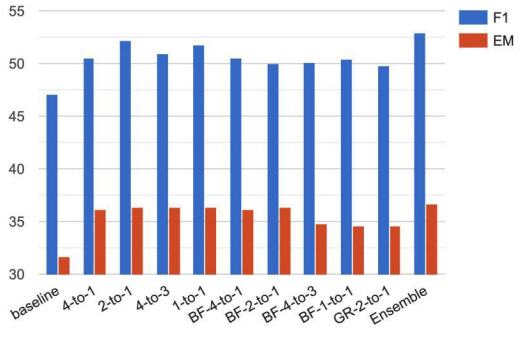

Figure 8: F1 and EM scores across all experiments

6 Analysis

6.1 Quantitative Error Analysis

As shown in Figure 9, we investigate the effects of answer length and context length on the performance of our model and the performance degradation moving from in-domain data to out of domain data for each category. The vast majority of in-domain (71%) and out-of-domain (93%) answer lengths are between 1 to 3 words, therefore the statistics for answer lengths greater than 3 could be unreliable. Nevertheless, the out-of-domain performance degrades greatly for answer lengths greater than 7 words. The vast majority of in-domain (74.5%) answer lengths are less than 200 words;

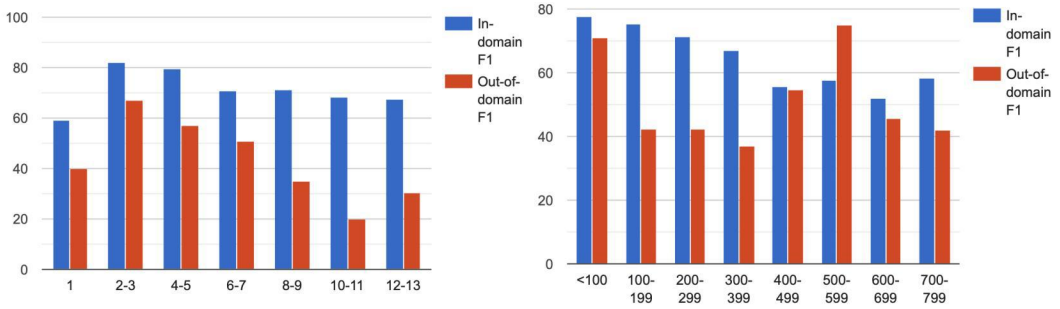

Figure 9: In-domain vs. Out-of-domain F1 scores by expected answer length (Left) and In-domain vs. Out-of-domain F1 scores by expected context length (Right)

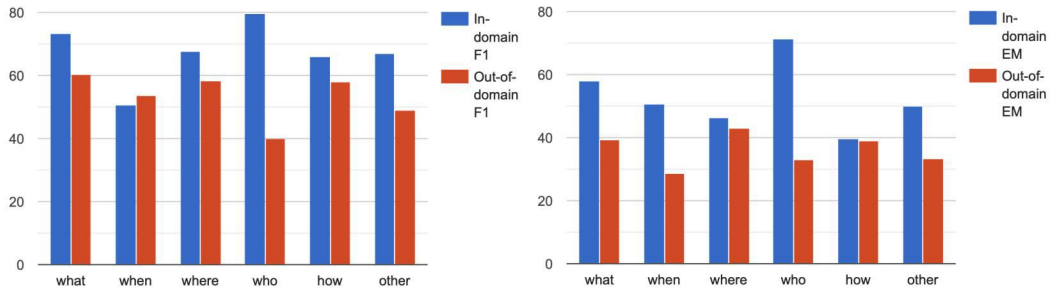

Figure 10: In-domain vs. Out-of-domain F1 scores by question type (Left) and In-domain vs. Out-of-domain EM scores by question type (Right)

Question: At the 2013 CES, which item drew the most attention?
Shortened Context Paragraph: Televisions were among the most talked about items at the 2013 International Consumer Electronics Show last week in Las Vegas, Nevada. Some employed the most advanced technology ever. Some of the TVs used a new technology called Organic Light Emitting Diodes, or OLED. ... There was even a fork that tells you when you are eating too fast. Cars, smart-phones, tablet computers and PCs also made news. And a 27-inch table computer drew quite a bit of attention. CEA President Gary Shapiro says there was much to see but not nearly enough time to see it all. "You cannot see the show in the four days that you have. We have over 3200 different industries showing over 20,000 new products. It's ly incredible."
Answer: Televisions
Prediction: a 27-inch table computer

Question: What can be the best title of this passage?
Shortened Context Paragraph: His name was Fleming, and he was a poor Scottish farmer. One day, while trying to make a living for his family, he heard a cry for help coming from a nearby bog. ... In time, Farmer Fleming's son graduated from St. Mary's Hospital Medical School in London, and went on to become known throughout the world as the noted Sir Alexander Fleming, the discoverer of penicillin. Years afterward, the nobleman's son was stricken with pneumonia. What saved him? Penicillin. The name of the nobleman? Lord Randolph Churchill. His son's name? Sir Winston Churchill. Someone once said, "What goes around, comes around."
Answer: 'What goes around, comes around'
Prediction: Sir Alexander Fleming,

Figure 11: Error samples from the out-of-domain validation set

however, 26.7% of out-of-domain samples has context greater than 600 words and 68.6% are less than 400 words. Overall, there is no reliable trend for out-of-domain performance degradation in terms of context length.

As shown in Figure 10, we investigate the effects of question type on in-domain and out-of-domain performance. The sample categories are classified using substring matching (e.g., checking if 'what' is a substring of the question). Note there are very little 'when' questions in the out-of-domain validation set, therefore the statistics for 'when' questions could be unreliable. Overall 'who' questions experienced the greatest out-of-domain performance loss; it is possible that out-of-domain distribution of named entities are significantly different compared to the in-domain distribution.

6.2 Qualitative Error Analysis

Figure 11 highlights several common errors made by the ensemble model on the out-of-domain validation set. The model often pattern match the question with the context naively to produce the answer; in the first example, the model matched 'most attention' with 'quite a bit of attention' instead of 'most talked about'. This category of mistakes require the model to recognize deeper semantic relationship over syntactic relationships. The model also fails at recognizing idioms and symbolic relationships in general; in the second example the model naively selected recurring entity 'Sir Alexander Fleming' instead the idiomatic phrase 'What goes around, comes around'.

7 Conclusion

In this project, we demonstrated FGM-based adversarial training can significantly and consistently improve the out-of-domain performance of DistilBert in the domain of Question Answering. Our approach does not alter the model architecture and can be easily generalized to other domains. However, adversarial training does increase training time and memory requirement.

In the future, we would like to evaluate alternative adversarial training methods such as Projected Gradient Descent (PGD), alternative models including BERT, RoBERTa [17] and ALBERT [18] and alternative domains including Token Classification and Sequence Classification.

References

[1] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness, 2020.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[8] Danqing Zhu, Wangli Lin, Yang Zhang, Qiwei Zhong, Guanxiong Zeng, Weilin Wu, and Jiayu Tang. At-bert: Adversarial training bert for acronym identification winning solution for sdu@aaai-21, 2021.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[10] T. Kwiatkowski, J. Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, C. Alberti, D. Epstein, Illia Polosukhin, J. Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Q. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[11] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset, 2017.

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[13] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension, 2017.

[14] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension, 2018.

[15] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations, 2017.