Generating Robustness: Exploring Various Ways to Adapt Question Answering to New Domains

Stanford CS224N Default Project - RobustQA

Helen Gu
Department of Statistics
Stanford University
helengu@stanford.edu

Quentin Hsu
Department of Statistics
Stanford University
qhsu@stanford.edu

Nicholas Lui
Department of Statistics
Stanford University
niclui@stanford.edu

Abstract

We implement a variety of techniques that boost the robustness of a QA model trained with domain adversarial learning and evaluated on out-of-domain data, yielding a 16% increase in F1 score in development and 10% increase in test. We find that the following innovations boost model performance: 1) finetuning the model on augmented out-of-domain augmented data, 2) aggregating Wikipedia-type datasets during adversarial training to simplify the domain discriminator's task, and 3) supplementing the training data with synthetic QA pairs generated with roundtrip consistency. We also ensemble the best-performing models on each dataset and find that ensembling yields further performance increases.

1 Introduction

Question Answering (QA), or the task of asking a model to answer a question correctly given a passage, is one of the most promising areas in NLP. However, state-of-the-art QA models tend to overfit to training data and do not generalize well to new domains, requiring additional training on domain-specific datasets to adapt. In this project, we aim to design a QA system that is robust to domain shifts and can perform well on out-of-domain (OOD) fewshot data.

We first implement domain adversarial training, inducing our QA model to learn features that are not domain specific so that it can use these domain-agnostic representations to make better predictions outside the domains on which it is trained. We initially find that domain adversarial training yields little improvement over the baseline; however, by experimenting with various techniques, including finetuning, out-of-domain data augmentation, changes to domain alignment, and ensembling, we achieve substantial improvements to model performance.

2 Related Work

Adversarial Learning. Lee et al. (2019) [1] introduce a framework that applies adversarial training to train a QA model, forcing the model to learn domain-invariant representations that enable it to better generalize to out-of-domain data. They ultimately validate their adversarial model for the MRQA Shared Task using 6 out-of-domain dataset and obtain a 1.5 point improvement in the average F1 score over the BERT baseline. This paper, however, has limitations in the context of this project: specifically, their adversarial model is trained on a large variety of data originating from six datasets, enabling the model to better learn domain-invariant features due to the diversity of domains on which it is trained. In our project, we are provided with only three in-domain datasets, two of which are both Wikipedia data. This relatively homogeneous training data further limits the ability of our model to learn domain-agnostic representations. In our project, we must find additional ways to bolster the QA model's ability to learn domain invariance.

Data Augmentation. Previous work by Wei and Zou (2019) [2] has found that easy data augmentation techniques (EDA) such as synonym replacement, random deletion, random swap, and random

insertion can be powerful methods for improving textual classification tasks on smaller datasets. We leverage synonym replacement to generate additional out-of-domain data for our model to train and finetune on, adapting the nlpaug library authored by Ma, 2019 to generate augmented data for our project.

Synthetic Question Answer (QA) Generation. Question generation has been found to improve QA models in low-resource settings with limited gold labeled samples (Yang et al. 2017 [3], Dhringa et al. 2018 [4]). Alberti et al. 2019 [5] develop a method of generating synthetic question-answer-context triplets and filter the results to ensure roundtrip consistency; their synthetic data generation model obtains state-of-the-art results on the Natural Questions [6] dataset. We adapt the synthetic question generation model that Alberti et al. developed to expand our out-of-domain training data, providing us with more varied training samples from which the model can learn domain-agnostic representations.

3 Approach

3.1 Baseline

Our baseline is the DistilBERT Question Answering model trained on the provided in-domain training data, as specified in the default project.

3.2 Main Approach: Domain Adversarial Training

The domain adversarial learning model is comprised of two components - a QA model and a domain discriminator. During training, the discriminator is trained to predict the domain (dataset) of the hidden representation produced by the QA model. In contrast, the QA model is penalized for the success of the discriminator, thus forcing the QA model to learn domain-invariant features such that it produces a hidden representation that is indistinguishable to the domain discriminator. At the same time, the domain discriminator also needs to learn what domain-invariant features to keep in order to maintain its performance on the samples generated by the QA model.

The discriminator is trained with a cross-entropy loss function. For a given training point, the loss function compares the discriminator's predicted probabilities (for all K domains) and the ground truth label (a one-hot vector which specifies the actual domain the data point belongs to).

The QA model is trained with a combined loss function comprised of a standard cross-entropy loss ($\mathcal{L}_{QA}$) plus a domain-invariance term ($\mathcal{L}_{adv}$) that measures the Kullback-Leibler divergence between the uniform distribution over all K domains and the discriminator's actual domain prediction ($\hat{\mathbf{d}}_i$). The final loss for the QA model is given by $\mathcal{L}_{QA} + \lambda\mathcal{L}_{adv}$ where λ is a hyper-parameter for controlling the importance of adversarial loss. We use $\lambda = 0.01$ as previous work finds this value of lambda performs best in ablation studies [1].

3.3 Improvements

3.3.1 Out-of-domain Fine-tuning

After training the adversarial model, we finetune the QA model (without the discriminator) on the out-of-domain data to give it a chance to learn domain-specific features from the out-of-domain samples. Lee et al. [1] do not finetune their model after adversarial training, but we find that finetuning greatly improves the performance of the model on out-of-domain data.

3.3.2 Data Augmentation

As we have limited out-of-domain data to train and finetune on, we hypothesize that out-of-domain data augmentation may help improve the performance of our model. Therefore, we implement two techniques to expand our out-of-domain data samples.

EDA: Synonym Swapping. We implement the synonym swap method from the nlpaug package for easy data augmentation. To adapt synonym swapping to our question answer dataset, we develop a way to account for random swaps that occur within the answer span. We first generate highlighting tags for the answer span within the context paragraph. After highlighting, we apply

synonym swap (and exclude the highlight tags as stop words), extract out the answer + starting index, and remove the highlight tags to ensure the compatibility of synonym swap with question answering. Using this approach, we generate 381 extra context-question pairs from the out-of-domain data.

Synthetic Question Answer Generation. We leverage the multitask T5 model finetuned on a SQuAD dataset to generate synthetic question answer pairs [5]. This model only supports processing paragraphs that are at most 16 sentences long, so we build a chunking process to split our context paragraphs into 16 sentence chunks with an 8 sentence overlap between chunks. Then, we run the question generation model to generate approximately 1 question and answer span per sentence per chunk. We extract the answer index by searching for the answer text in the nearest sentence. Finally, to ensure roundtrip consistency, we remove duplicate questions and then rerun the QA portion of the T5 model to repredict the answer given the generated question and context chunk. We only keep question answer pairs in which the QA model predicts the same answer as the generated answer to ensure we have high-quality question-answer pairs. Using this approach, we generate 1579 extra context-question-answer pairs.

3.3.3 Domain Alignment

The standard approach to alignment is multi-source alignment, where each dataset is treated as a different domain for the discriminator to predict. However, this poses three challenges. First, the domain boundaries are not well-defined: SQuAD and Natural Questions are both Wikipedia-based datasets, so the discriminator is trained to differentiate between relatively similar domains, which may impede the model's ability to learn features agnostic to more substantial domain shifts. Secondly, if we include out-of-domain training data, the number of domains that need to be identified increase from 3 to 6, impeding the discriminator's ability to effectively differentiate between domains, particularly when it has few samples to learn from in some domains. Thirdly, if we include out-of-domain training data, the discriminator faces major class imbalance as there are more than 3500 times more in-domain samples than out-of-domain samples. We hypothesize that these challenges make it difficult for the discriminator to learn to distinguish between domains, leading the discriminator to exert less pressure on the QA model and diminishing the QA model's ability to generalize to out-of-domain samples.

Thus, we explore Wiki alignment where the Wikipedia datasets (SQuAD, NaturalQuestions, RelationExtraction) are treated as one domain, and the non-Wiki datasets (NewsQA, DuoRC, RACE) are treated as another domain. This allows us to partition the sample space into fewer, better-balanced domains with well-defined boundaries.

3.3.4 Tuning Discriminator Architecture

We hypothesize that the QA model may fail to generalize well to out-of-domain samples due to weak discriminator performance during adversarial training as a result of limited out-of-domain samples, and thus explore two innovations that improve the stability of discriminator learning. First, we incorporate discriminator lambda annealing. The discriminator lambda starts at 0 and is gradually increased using a tanh function before plateauing at 0.01 at step 20,000. This prevents the discriminator from initially being overwhelmed with difficult examples, and allows it to progressively train on harder examples[7].

Secondly, we incorporate Wasserstein regularization [8] where the weights of the discriminator are clipped between -0.01 and 0.01 before backward propagation. Weight clipping can enforce the Lipschitz constraint, which regularizes adversarial training and improves stability.

3.3.5 Ensemble Methods

After experimenting with multiple approaches, we ensemble various models together with the goal of improving prediction accuracy by reducing overall variance. We do this by averaging logits for the start and end indices across our models prior to producing a final answer. We specifically choose to combine the models that produce the highest F1 score on each of the out-of-domain datasets (Relation Extraction, DuoRC, and RACE).

4 Experiments and Analysis

We use the datasets provided for the RobustQA track. These datasets are described in the appendix.

Models were trained for 3 epochs and fine-tuned for 5 epochs with evaluation every 50 steps. A batch size of 16 and Adam Optimizer with a learning rate of 3e-5 were used. For adversarial models, we use a discriminator lambda of 0.01. Overall results are presented in **Table 1**. Dataset-level results are presented in the Appendix A.1.

4.1 Baseline

Our baseline is a QA model that was trained on in-domain data. Without finetuning, the baseline achieves an F1 score of **49.88** and an EM score of **34.55**. After finetuning on out-of-domain training data, the baseline achieves an F1 score of **49.43** and an EM score of **33.25**.

Finetuning does not improve the baseline model's performance. One possible reason is that the baseline QA model learned features specific to in-domain datasets that may directly contradict features specific to out-of-domain datasets (e.g. different document structures). As such, finetuning becomes less effective as the training and finetuning "contradict" each other.

4.2 Experiment 1: Adversarial Learning

Description. Following Lee et al., we train a domain adversarial model on in-domain data. We train one variant without finetuning on out-of-domain data (M1) and one variant with finetuning (M2).

Results. We find that initially, without finetuning, the adversarial model underperforms the baseline model without finetuning (-4.7% change in F1). After finetuning, however, the adversarial model sees a large performance improvement from finetuning (+5.7% change in F1) and outperforms baseline, achieving an F1 score of **50.22**.

Analysis. Our results suggest that the adversarial model, by itself, does not perform well on out-of-domain data in this context. We hypothesize that our results may differ from Lee et al. because we have fewer distinct domains to train on, limiting the ability of the model to learn domain-invariant features that generalize well. Surprisingly, we notice a large improvement in F1 scores after finetuning, suggesting that finetuning is required to unlock the potential of adversarial models in this context. We hypothesize that the model learns some domain-invariant features from training, but it is only able to adapt these to out-of-domain data after finetuning.

4.3 Experiment 2: Expanding Out-of-domain Fine-tuning Samples

Description. Given the substantial performance increase of the QA model after finetuning, we want to see if expanding the finetuning dataset with augmented and synthetic out-of-domain samples can improve performance. We use the augmented data described in Section 3.3.2 to finetune on the following datasets:

1. out-of-domain data + 381 extra context-question pairs generated using synonym swapping (M3)

2. out-of-domain data + 1579 extra context-question pairs using synthetic question generation (M4)

3. out-of-domain data + 381 extra context-question pairs using augmentation + 1579 extra context-question pairs using synthetic question generation (M5)

Results. Finetuning on out-of-domain data supplemented with EDA samples yields large improvements in performance. M3 achieves the best results with an F1 score of **53.5**, outperforming M2 by 6.5%.

Analysis. Crucially, we find that expanding the finetuning dataset via EDA synonym swap significantly improves the model's performance. This builds upon our previous finding that finetuning improves performance, potentially because finetuning on more examples enables the model to adapt the domain-invariant features it has learned to out-of-domain data.

We note that the inclusion of synthetic examples (M4) leads to a slight degradation in performance relative to finetuning on out-of-domain data only (M2, -1.0% change in F1). Including synthetic

examples alongside augmented examples (M5) also leads to a slight degradation (-0.7% change in F1) relative to M4 which only has augmented examples.

We hypothesize that EDA works better than synthetic examples during finetuning because the model is very sensitive to the quality of out-of-domain samples as it is trying to extract precise features from the samples during finetuning. EDA synonym swap results in only small changes to the augmented data; therefore, these question-answer pairs are likely to be well-formed and be good representations of the out-of-domain data. Synthetic question generation, on the other hand, is a more difficult task; thus, it may result in lower-quality question-answer pairs that are nonsensical, degrading performance.

4.4 Experiment 3: Domain Alignment

Description. Thus far, we have been using multi-source alignment where each dataset is treated as a unique domain. However, as explained in Section 3.3.3, domain boundaries are not well-defined under multi-source alignment. We build upon Experiment 4.3's best model, M3, by using Wiki alignment and finetuning on out-of-domain + EDA augmented samples, giving rise to model M6.

Results. M6 achieves an F1 score of **55.12**, which is a 3.0% improvement in F1 over M3.

Analysis. This suggests that Wiki alignment is indeed valuable in helping the discriminator learn better, which in turn improves QA model performance. We note that this improvement largely arises from improvements on RelationExtraction, where F1 rises from **78.17** to **83.33** (+6.6%). RelationExtraction is Wikipedia data, suggesting that forcing the model to differentiate between two Wikipedia datasets worsens its ability to generalize well to out-of-domain Wikipedia data.

4.5 Experiment 4: Including Out-of-domain Data in Training under Wiki Alignment

Description. We note that our previous results have yielded large increases specifically on the RelationExtraction dataset (+25.4%), but a much smaller performance increase on RACE (+1.8%), and a sizeable performance *decrease* on DuoRC (-4.3%). Given that the training samples are dominated by Wiki datasets (2 out of 3 in-domain datasets are Wiki datasets), we can expect the model to be especially good at learning features that are generalizable across Wiki datasets, translating to robust performance improvements on the out-of-domain Wiki dataset, RelationExtraction. We hypothesize that the model is performing less well on the non-Wikipedia out-of-domain datasets because the discriminator is not trained on data from these domains; therefore, the model does not learn feature invariance that generalizes well to them. We thus try introducing out-of-domain samples during training.

Results. After including out-of-domain data, M7 achieves an F1 score of **53.41**, a -3.1% decline in F1 over M6. We find that the model's performance declines across DuoRC and RelationExtraction, but improves on RACE.

Analysis. We hypothesize that the model generally performs worse once out-of-domain data are included during training because there are so few out-of-domain samples. Thus, the discriminator is not able to reliably learn to differentiate out-of-domain samples during training, confusing the discriminator and diminishing the QA model's ability to produce domain-invariant features. To test this hypothesis, we next experiment with augmenting the out-of-domain training data.

4.6 Experiment 5: Augmenting Out-of-domain Data in Training under Wiki Alignment

We next try training our adversarial model on the same datasets experimented with during finetuning:

1. out-of-domain data + 381 extra context-question pairs using augmentation (M8)

2. out-of-domain data + 1579 extra context-question pairs using synthetic question generation (M9)

3. out-of-domain data + 381 extra context-question pairs using augmentation + 1579 extra context-question pairs using synthetic question generation (M10)

Results. M9, which augments the training data with extra question-answer pairs from synthetic question generation, achieves the best results with an F1 score of **55.53**, an improvement of 0.7% over M6. Although this is a small improvement in overall F1 score, we find that at the dataset-level, the improvements across datasets are more balanced, and that training on expanded out-of-domain

data greatly improves DuoRC F1 and slightly improves RACE F1, at the cost of small degradation in RelationExtraction F1 relative to M6. Relative to baseline (M1), we now find improvements in F1 score across all datasets (RelationExtraction +18.8%, DuoRC +7.1%, RACE +3.2%).

Analysis. We find that expanding out-of-domain training samples via synthetic question generation improves the overall out-of-domain performance of the model, but including out-of-domain EDA samples decreases the model's overall out-of-domain performance.

We hypothesize that including out-of-domain synthetic question answer samples improves the model's performance for two reasons. First, synthetic question generation produces diverse (albeit noisy) question-answer pairs for more parts of a given context paragraph. These diverse samples introduce the domain discriminator and the QA model to different locations and structures of the context paragraph, better enabling the QA model to learn domain-invariant features. Second, we introduced 1579 out-of-domain synthetic samples in this step, vastly expanding the number of out-of-domain samples trained on by over 400%. This supports our hypothesis that introducing too few out-of-domain training samples may degrade overall performance due to the discriminator's inability to learn, but that introducing a sufficient number of out-of-domain examples may improve performance.

By contrast, we find that including EDA augmented paragraphs in training (M8) leads to a further degradation in performance relative to training on only out-of-domain samples. We hypothesize that because we only introduce a limited number of EDA samples (+100% increase in out-of-domain training data), we have not yet introduced enough out-of-domain data such that the discriminator is able to learn to distinguish between these out-of-domain samples, thus worsening its performance with the added data.

Finally, we find that including both EDA and synthetic question generation out-of-domain data does not improve F1 score relative to including only synthetic questions (-0.16% change in F1). We hypothesize that this occurs because EDA samples are not very different relative to the original out-of-domain samples because they are generated by synonym swap, so adding these samples during training provides no additional benefit to the adversarial model in learning domain-invariant features.

We also find that Wikipedia alignment performs better on augmented out-of-domain training data than multi-source alignment does; these results can be found in Appendix A.2.

4.7 Experiment 6: Tuning Discriminator Architecture

Description. As we assume that one limitation of the model's performance may be the discriminator's ability to learn domain invariant features well from a limited number of out-of-domain samples, we employ a few techniques targeted toward improving the discriminator's architecture. We build upon the best model from Experiment 4.6, M9, by incorporating lambda annealing and Wasserstein regularization. This gives rise to model M11.

Results. M11 underperforms M9 with a 1.6% decrease in F1.

Analysis. Performance degrades with techniques designed to improve stability in discriminator training. We hypothesize that synthetic question augmented out-of-domain training data already sufficiently improves discriminator training by providing enough out-of-domain samples from the adversarial model to learn on. As such, the imposition of additional constraints may be unnecessary and potentially harmful to model performance. Appendix A.3 shows that without synthetic out-of-domain training data, tuning the discriminator improves overall results, validating this hypothesis.

4.8 Experiment 7: Ensemble Methods

Description. Finally, to reduce variance, we employ ensembling on diverse, high-performing models. Specifically, we choose to ensemble the best models for each out-of-domain dataset (M6 - RelationExtraction, M10 - DuoRC, M11 - RACE).

Results. Ensemble 1 attains an F1 score of **57.86**, which is a 4.2% improvement in F1 over M9.

Analysis. These results suggest that ensembling is an extremely powerful way of boosting performance.

4.9 Overall Results

4.9.1 Validation/Dev Set

Overall, we find that our ensemble model yields a 16% improvement in F1 score over baseline. The model also yields improved F1 scores across all datasets relative to baseline: RelationExtraction improves by 35.2%, DuoRC improves by 12.7%, and RACE improves by 11.8%. We find that the baseline outperforms our model only for DuoRC EM score.

	Model Specification				OOD Validation	
Model No.	Architecture	Training Data	Finetuning	Alignment	F1	EM
Baseline1	Baseline	IND	None	Multi	49.88	34.55
Baseline2	Baseline	IND	OOD	Multi	49.43	33.25
M1	Adversarial	IND	None	Multi	47.51	30.89
M2	Adversarial	IND	OOD	Multi	50.22	34.55
M3	Adversarial	IND	OOD+AUG1	Multi	53.5	35.6
M4	Adversarial	IND	OOD+SYN	Multi	49.7	32.72
M5	Adversarial	IND	OOD+AUG1+SYN	Multi	53.29	32.46
M6	Adversarial	IND	OOD+AUG1	Wiki	55.12	35.08
M7	Adversarial	IND+OOD	OOD+AUG1	Wiki	53.41	33.51
M8	Adversarial	IND+OOD+AUG1	OOD+AUG1	Wiki	52.52	32.46
M9	Adversarial	IND+OOD+SYN	OOD+AUG1	Wiki	**55.53**	**37.17**
M10	Adversarial	IND+OOD+AUG1+SYN	OOD+AUG1	Wiki	55.44	35.86
M11	Tuned Adv	IND+OOD+SYN	OOD+AUG1	Wiki	54.66	36.39

Table 1: Main Results

	Overall		DuoRC		RACE		RelationExtraction	
Model No.	F1	EM	F1	EM	F1	EM	F1	EM
Baseline1	49.88	34.55	45.68	**37.3**	37.44	24.22	66.46	42.19
Ensemble	**57.86**	**39.01**	**51.48**	34.13	**41.84**	**25.78**	**80.16**	**57.03**

Table 2: Ensemble vs. Baseline Results

4.9.2 Test Set

	Test		OOD Validation		Predicted Test	
Model No.	F1	EM	F1	EM	F1	EM
Distilbert Baseline	58.869	40.528	49.88	34.55	57.72	39.06
M9	62.808	42.202	55.53	37.17	66.47	45.98
Ensemble	**65.271**	**45.894**	**57.86**	**39.01**	**67.64**	**48.03**

Table 3: Test Results

We see that the test results have substantially higher F1 scores than our validation results. This is expected: the validation set has an even distribution of out-of-domain datasets whereas the test set is more heavily weighted towards RelationExtraction, which is a Wikipedia-based dataset that is closer to our in-domain datasets and is thus an easier task on which our model has higher performance.

We adjusted the weighted average of our model's performance to predict our test score for all of our experiments, and compare our actual test score to predicted test score. We find that we achieve around

95% of our predicted test scores. This suggests that we are slightly overfitting to our validation set due to all the model selection decisions made using our validation set.

5 Additional Analysis

5.1 EM Scores

We focus on evaluating improvements to our model using F1 scores. However, we should note that while our model achieves large improvements in F1, increases to EM scores are much smaller. We find that as our model becomes more robust and reading comprehension increases (coinciding with increases to F1), it begins to add grammatical articles to phrase predictions, causing its outputs to not exactly match gold answers. For example, our model predicts 'age of fourteen' for a question where the gold answer is 'fourteen', resulting in penalized EM score even when the predictions are correct. Thus, we see greater improvements to F1 scores than to EM.

5.2 Comparing Synthetic Questions to EDA Data Augmentation

We implement two techniques for out-of-domain data augmentation and find that their effects vary by context. Specifically, we find that EDA data augmentation improves results during finetuning, whereas synthetic QA generation improves results during training. We hypothesize this is due to EDA samples being more precise but less varied (benefiting finetuning), whereas synthetic question generation leads to samples that are more diverse, but less trustworthy (benefiting adversarial training). To investigate this hypothesis, we examine samples generated by the synthetic QA generation process, and we find various examples supporting our hypothesis that synthetic QA generates noisier examples. The example below demonstrates how synthetic QA generation can produce flawed examples that the model is asked to learn from:

Example context-question-answer triplet generated by synthetic question generation:

Context: Later, Pink holds a rally in suburban London, singing "Waiting for the Worms". The scene is inter-cut by images of animated marching hammers that goose-step across ruins. Pink screams "Stop!" and takes refuge in a bathroom stall at the concert venue, reciting poems which would later be used as lyrics on Pink Floyd's "Your Possible Pasts" from The Final Cut album and "5:11 AM (The Moment Of Clarity)" from Roger Waters' The Pros and Cons of Hitch Hiking.

Question: "What song does Pink scream?"

Answer: "Stop!"

We believe that these flawed examples degrade the model's ability to comprehend context during finetuning; a further discussion is included in Appendix A.4.

5.3 Higher-order Reasoning

After performing error analysis, we note that a limitation of our model is that it fails to perform well on questions that involve higher-order reasoning. This explains why model improvements are smallest for the RACE dataset. A deep dive into the validation datasets suggests that RACE comprises a higher proportion of higher-order inference questions than the other two OOD datasets (unsurprisingly, given that RACE contains examination questions designed to test reading comprehension). An example of this can be found in Appendix A.5.

6 Conclusion

In this project, we implemented a QA model with domain adversarial training to improve robustness on out-of-domain data by forcing the QA model to learn domain-invariant features during training. We find that a number of adjustments to Lee et al. [1]'s approach can improve the performance of a QA model that is trained on data from a limited number of in-domain datasets and asked to generalize to out-of-domain data with few samples; these findings provide multiple promising avenues for further research.

First, we find that finetuning the QA model on out-of-domain data after adversarial training is highly effective– an approach not previously explored in Lee et al. [1]. We not only show the efficacy of finetuning, but also find that finetuning on out-of-domain data augmented using synonym swapping leads to large performance gains. This finding provides an additional avenue for investigation: we tried finetuning only on data augmented with synonym swap, but could theoretically increase the number of augmented samples further with additional EDA techniques such as random insertion, random swap, and random deletion. With a larger pool of augmented out-of-domain samples, we may be able to further reap the benefits of finetuning. However, not all augmented OOD data benefits finetuning: our results suggest that finetuning on noisy samples (such as those obtained by synthetic question generation) degrade performance.

Additionally, whereas Lee et al. [1] define one domain per dataset in adversarial training, we demonstrate substantial improvements to model performance when domain boundaries are defined according to passage source domain during adversarial training, rather than according strictly to dataset. Based on this finding, another avenue of exploration in future work involves experimenting further with domain alignment. For example, clustering approaches based on word embedding similarity could further help boost the discriminator's ability to learn domain-invariant features that generalize well to the out-of-domain data.

Furthermore, we show that training on augmented out-of-domain data can increase the model's generalized performance on out-of-domain datasets that are less similar to the in-domain training data. We also demonstrate that adversarial training on too few out-of-domain samples is detrimental to model performance, but that expanding the out-of-domain training data with synthetic examples can remedy this issue. Improving the quality and quantity of synthetically generated samples may help improve model performance even further.

Lastly, our results also show that ensembling can provide large gains to model performance, demonstrating the benefits of ensembling in this setting.

7 Acknowledgements

We would like to thank our mentor Michihiro Yasunaga, as well as the entire CS 224N teaching staff for providing the support and materials for this project.

References

[1] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training. *CoRR*, abs/1910.09342, 2019.

[2] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196, 2019.

[3] Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4] Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. *CoRR*, abs/1906.05416, 2019.

[6] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[7] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. *CoRR*, abs/1511.06349, 2015.

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

A Appendix

A.1 Dataset-Level Results

	Overall		DuoRC		RACE		RelationExtraction	
Model No.	F1	EM	F1	EM	F1	EM	F1	EM
Baseline1	49.88	34.55	45.68	**37.3**	37.44	24.22	66.46	42.19
Baseline2	49.43	33.25	43.99	34.13	37.75	23.44	66.46	42.19
M1	47.51	30.89	44.54	34.92	29.51	16.41	68.44	41.41
M2	50.22	34.55	39.77	26.98	35.07	21.09	75.66	55.47
M3	53.5	35.6	44.95	29.37	37.24	23.44	78.17	53.91
M4	49.7	32.72	42.56	26.98	33.37	18.75	73.05	52.34
M5	53.29	32.46	44.16	25.4	35.56	20.31	80.01	51.56
M6	55.12	35.08	43.72	28.57	38.12	21.88	**83.33**	54.69
M7	53.41	33.51	41.72	26.19	40.09	22.66	78.23	51.56
M8	52.52	32.46	42.55	28.57	39.45	21.09	75.4	47.66
M9	**55.53**	**37.17**	48.93	33.33	38.63	22.66	78.93	**55.47**
M10	55.44	35.86	**50.86**	34.13	37.79	21.88	77.62	51.56
M11	54.66	36.39	44.43	29.37	**41.07**	**25.78**	78.31	53.91
Ensemble	**57.86**	**39.01**	**51.48**	34.13	**41.84**	**25.78**	**80.16**	**57.03**

Table 4: Dataset-Level Results

A.2 Training on augmented OOD data with multisource alignment

	Model Specification				OOD Validation	
Model No.	Architecture	Training Data	Finetuning	Alignment	F1	EM
M3	Adversarial	IND	OOD+AUG1	Multi	53.5	35.6
M12	Adversarial	IND+OOD	OOD+AUG1	Multi	52.01	32.46
M13	Adversarial	IND+OOD+SYN	OOD+AUG1	Multi	52.51	33.77
M14	Adversarial	IND+OOD+AUG1+SYN	OOD+AUG1	Multi	52.89	34.03
M11	Adversarial	IND+OOD+AUG1+SYN	OOD+AUG1	Wiki	**55.44**	**35.86**

Table 5: Augmenting Training Data for Multi-source Alignment

Similar to Experiments 4 and 5 (adding OOD samples to training data under Wikipedia alignment), we find that as we add OOD samples to training data under multi-source alignment, OOD performance initially decreases; then, as we augment training data, performance increases.

Comparing these results to Wikipedia alignment, we find that Wikipedia alignment still outperforms multi-source alignment after augmenting training data. We find this result compelling as under Wikipedia alignment, the discriminator is tasked with differentiating between two domains, an easier task than differentiating between six domains. This provides further evidence that Wikipedia alignment is an effective strategy over multi-source alignment in boosting the performance of the adversarial model in settings with fewshot OOD data.

A.3 Tuning discriminator architecture

	Overall		DuoRC		RACE		RelationExtraction	
Model No.	F1	EM	F1	EM	F1	EM	F1	EM
M7 (Wiki alignment)	53.41	**33.51**	41.72	26.19	40.09	**22.66**	**78.23**	**51.56**
M7 + Adv Tuning	**54.03**	32.72	**43.07**	**26.98**	**41.87**	21.88	77	49.22
M12 (Multi alignment)	52.01	32.46	41.93	27.78	37.31	19.53	76.64	50
M12 + Adv Tuning	**53.85**	**34.55**	**45.27**	**30.95**	**37.67**	**21.88**	**78.49**	**50.78**

Table 6: Tuning the discriminator architecture for models trained on IND + OOD data

M7 is the adversarial model using Wikipedia alignment trained on IND + OOD data, then finetuned on OOD+AUG1 data. M12 is the adversarial model using multi-source alignment trained on IND + OOD data, then finetuned on OOD+AUG1 data.

The results above show that tuning the discriminator using lambda annealing and Wasserstein regularization does improve OOD performance in settings with smaller quantities of OOD data. Tuning the discriminator appears particularly helpful for multi-source alignment, which is an intuitive result. Multi-source alignment is a more difficult task for the discriminator than Wikipedia alignment, particularly in settings with limited OOD data. Therefore, we would expect the improvements to discriminator architecture to yield greater improvements for multi-source alignment than for Wikipedia alignment, as shown in the table above.

A.4 Predictions after finetuning on synthetic QA compared to finetuning on augmented data

We perform error analysis to investigate why finetuning on synthetic generated QA degrades performance, and we find that finetuning on synthetic QA incentivizes the QA model to look at more context when making predictions, but degrades the model's ability to comprehend sentence structure and return correct answers. As a result, the model tends to match more words from the question to the context paragraph to predict answers, instead of understanding the context of what is asked and returning sensible answers. This can be seen in the example below:

Example of model performance after finetuning:

Context: A one-armed criminal, a Boglodite named Boris the Animal, stages a jailbreak (along the way killing Obadiah Price, a fellow inmate he'd made a deal with). His intention is to rewrite history, with Agent K, the one who arrested and imprisoned him, being a major factor in his plan...He managed to kill on 15 July 1969 an alien named Roman the Fabulous and on the following day... at this point O cuts him off and tells him not to investigate any further.

Question: Who shoots off Boris's left arm?

Ground Truth: Agent K

Answer (finetuning on EDA samples): Agent K

Answer (finetuning on synthetic QA samples): O cuts him off and tells him not to investigate any further. That night, as K

A.5 Examples of model failing to generalize to answer higher-order questions.

Context: "Want to save money when travelling by train? Day Returns This ticket can save you up to 45% on the standard fare... Weekend Returns are available for most journeys over 60 miles. Go on Fri. Sat. or Sun, and return the same weekend on Sat. or Sun, and save up to 35% the standard fare. Monthly Returns There are available for most journeys over 65 miles. Go any day and return within a month. . . "

Question: "Which is the best ticket to buy if you live in London and want to go to a small town 80miles away for four days?"

Gold: Monthly Returns || **Answer (M10):** to 45%

To answer this question, you need to do mental arithmetic and shortlist the ticket types based on distance and number of days allowed. Our model is confused by this complex question and returns an irrelevant answer.