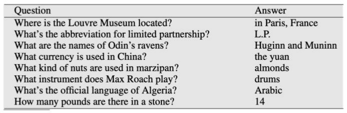
Adversarial Training with Data Augmentation for Robust Question Answering

German Kolosov [1] Juan Langlois [1] Thomas Le Menestrel [2]

[1]MS&E Department
Stanford University

Robust Question Answering



Question	Answer
Where is the Louvre Museum located?	in Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	the yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What's the official language of Algeria?	Arabic
How many pounds are there in a stone?	14

Figure 1. Examples of question prompts and answers

Critical to Deep Learning's success in real-world applications is the ability to build models that generalize well to unseen data and can therefore cover multiple domains. In particular, in Question Answering (QA) tasks in NLP, models struggle to generalize and often over-fit a specific dataset, making them unreliable for tasks on new datasets. It is therefore critical to develop methodologies to train domain-agnostic QA models. In other words, a model that learns domain-invariant features to generalize to unseen data. We propose using an Adversarial Training framework, similarly to GANs in computer vision [4]. Our model has two main components: (i) the QA and (ii) a domain discriminator. Finally, to further increase the robustness, Easy Data Augmentation [7] will be used to augment training data.

Adversarial QA

We implemented the Adversarial QA architecture proposed by [5] on a pretrained DistilBert encoder:

- **Architecture:** The architecture jointly trains a **span classification** head for the QA task and a **discriminator** head for the multi-class domain classification.
- **Training:** The training proceeds by iteratively training an augmented QA loss and a discriminator loss.
- **Discriminator Loss:** Class cross-entropy loss $\mathcal{L}_D$
- **Augmented QA Loss:** Span classification loss $\mathcal{L}_{QA}$ plus a scaled KL divergence between the uniform distribution and the discriminator's prediction $\mathcal{L}_{adv}$

$$L_{global} = L_{QA} + \lambda L_{adv} \quad (1)$$

Easy Data Augmentation

In order to generalize well, the model needs to learn how to deal with the data from a different domain. Given that the out of domain training set only contains a dozen samples, we suggest to apply easy augmentation techniques that will not perturb the general meaning but will add more flexibility and make the distribution of the data less domain specific. To that extent, we mainly work with two data augmentations [7] :

1. **Synonym replacement** Every word is replaced by its synonym with a probability p_{SR}.
2. **Random Deletion** Every word is deleted with probability p_{RD}

Both of the augmenting techniques mentioned above, are applied only to the part of the context that does not contain the answer. The data is augmented N times.

Experiments

- **Experiments:** Our experiments have two levels:
  1. **In-domain train and validation**: We train the models on in-domain training set and save the parameters that perform best on the in-domain validation set.
  2. **Few-shot out-of-domain train and validation**: We train the models from the previous stage on a small subset of out-of-domain training set and save the parameters that perform best on the out-of-domain validation set.
- **Data:** We use 6 datasets: 3 are in-domain reading comprehension datasets (SQuAD, NewsQA and Natural Questions) as well as 3 out-of-domain datasets (Relation Extraction, DuoRC and RACE . The in-domain datasets are used for (long) training and the out-of-domain datasets for few-shot training and evaluation.
- **Evaluation method:** The performance of each model is measured by two metrics: Exact Match (EM) score and F1 score.

Models

In-domain train and validation

- **Baseline:** pre-trained DistilBert model with a Question Answering head fine-tuned to the in-domain data.
- **Adversarial:** pre-trained DistilBert model with the adversarial architecture fine-tuned to the in-domain data.

Few-shot out-of-domain train and validation

- **Baseline + Fewshot:** Baseline model trained on small train set of out-of-domain data.
- **Adversarial + Fewshot + Freeze DistilBert:** Adversarial model trained on small train set of out-of-domain data while freezing the DistilBert parameters.
- **Adversarial + Fewshot:** Adversarial model trained on small train set of out-of-domain data.
- **Adversarial + Fewshot + EDA:** Adversarial model trained on small train set of augmented out-of-domain data.

Training Loss for In-domain data

The following plot shows the Baseline and Adversarial training. We can see that the Adversarial architecture is doing a good job at ”confusing” the discriminator while still minimizing the QA loss.

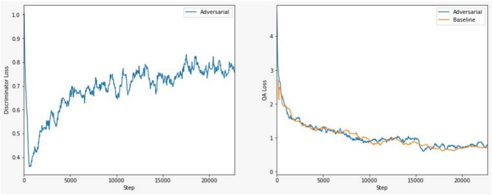


Figure 2. Training loss for in-domain data.

Results and Discussions

The following table shows the results of the different models on the out-of-domain validation set.

Model	EM	F1
Baseline	32.46	49.31
Adversarial	31.94	48.48
Adversarial + Fewshot + Freeze DistilBert	31.94	48.48
Adversarial + Fewshot	35.864	50.523
Adversarial + Fewshot + EDA	36.387	50.561

Table 1. Results summary for baseline and adversarial RobustQA models

In the table above, we see the results on the out-of-domain validation set. We can notice that the performances of the adversarial model and the baseline trained only on in-domain datasets are similar. However, loading the weights of these models and adding a few epochs of training on a small out-of-domain training set significantly (**fewshots**) boosts the performance. Finally, incorporating some easy data augmentations such as **synonym replacements** and **random deletion**, produces more general, model agnostic Q&A systems. Moreover, we have noticed that augmenting the data too much hurts performance which is in line with the conclusions of [7]. Finally, freezing the DistilBert parameters significantly hurts the performance, even while training on out of domain data.

Finally, as a way of improvement we may add some other kinds of data augmentations, more specifically, we believe that augmenting the question by performing back and forth translations may increase the robustness of the model.

We submitted the **Adversarial + Fewshot** model to the test leaderboard and obtained the following results:

EM: 40.183 F1: 57.981 Rank 37/50

References

[1] Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie.
Adversarial deep averaging networks for cross-lingual sentiment classification.
CoRR, abs/1606.01614., 2016.

[2] Xilun Chen and Claire Cardie.
Multinomial adversarial networks for multi-domain text classification.
2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
2018.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial networks.
2014.

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

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang.
Know what you don't know: Unanswerable questions for SQuAD.
In *Association for Computational Linguistics (ACL)*, 2018.

[7] Jason Wei and Kai Zou.
Eda: Easy data augmentation techniques for boosting performance on text classification tasks.
2019.

[8] Yi Wu, David Bamman, and Stuart Russell.
Adversarial training for relation extraction.
2017.

[9] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi.
Revisiting few-sample bert fine-tuning, 2021.