Question Answering Domain-Generalization With Domain Adversarial Training

Shadi Iskandar **Husam Ismael**
Technion, Israel Institute of Technology
{shadi.isk, hussam.is}@campus.technion.ac.il

Abstract

Many BERT variants (for instance ALBERT, RoBERTa, XLNET etc.) were able to reach super-human performance on the task of Reading Comprehension based Question Answering on popular datasets. However, they lack the ability to generalize to unseen domains examples. Which harms the ability to apply them on real world applications, where the distribution of the test data is different and is consistently shifting. In this project paper, we apply domain generalization technique in order to build a QA model which has general linguistic intelligence, applicable to various domains without need of further fine-tuning to out-of-domain datasets. We tackle this problem with Domain Adversarial Training. The results show considerable performance improvement in comparison with the baseline.

1 Introduction

The last few years have seen rapid progress in Natural Language Processing (NLP) models (Devlin et al., 2019; Collins et al., 2020), allegedly surpassing human-level performance in many NLP tasks like reading comprehension (RC) based question answering (QA) (XLNet, Yang et al., 2019). Although this might be true when evaluated on in-domain test sets, these models collapse when evaluated on out-of-domain (OOD) test sets (Sen et al., 2020)(see Table 1).

The main reason behind model's OOD poor performance is that such models often rely on the assumption that the test set is drawn from the same distribution of the train set. However, in real life applications the assumption does not always hold since text may come from many different sources, each with unique distributional properties, or insufficient data for specific domain to train the model from scratch. The findings in (Sen et al., 2020) and ours in Table 1 indicates that these models do not fully understand the given passage-question but rather are taking shortcuts present in each specific domain. While QA systems have witnessed great breakthroughs in RC tasks (ALBERT, Lan et al.,2019; XLNet, Yang et al., 2019), most existing methods focus on improving in-domain performance. Our goal is to build a QA model which can generalize to unseen out-of-domain, hoping to make the models better understand the passage-question.

		Evaluated on			
		SQuAD	**NewsQA**	**NaturalQ**	**HotpotQA**
Fine-tuned on	**SQuAD**	**70.3**	47.1	54.2	56.4
	NewsQA	46.7	**50.9**	46.4	31.7
	NaturalQ	43.8	33.6	**64.5**	32.9
	HotpotQA	49.6	35.4	47.5	**80.8**

Table 1: Performance results of each fine-tuned model evaluated on each out-of-domain set. We can see drastic drops in performance when evaluated on out-of-domain test set. All the results shown in the table are the corresponding F1 scores.

In this work, we take on the problem of domain generalization in question answering (QA) where we have few in-domain (source domain) labelled datasets, and no prior information on the out-of-domain (target domain) datasets. This problem, while less addressed in research, is more challenging than the unsupervised domain adaptation, where we have small set of target data.

Nonetheless, this problem is all-over in real life question answering machines (e.g., Siri, Alexa etc.), where the distributions of the target data may rarely be identical to those of the source data. In applying Domain Adversarial Training, we train one model on the collective of in-domain examples while encouraging the model to learn domain-invariant features rather than specific ones.

2 Related Work

We start by describing NLP research in the setting of domain adaptation (DA), which is more common and strongly related, then proceed to domain generalization (DG). Finally, we focus on adversarial-based methods.

Domain Adaptation: Most existing methods follow the settings of unsupervised domain adaptation, where in addition to labelled source domain, we have unlabeled target data. There has been extensive prior work on DA in NLP tasks, tackling the problem with different approaches. One interesting approach proposed by Blitzer et al., (2006) and further used in Ziser and Reichart (2018) is Pivot Based Domain Adaptation, it divides the shared feature space of the source and the target domains to a set of pivot/non-pivot features in order to map the original feature space of both domains into a shared feature space. Recently, Autoregressive language modeling and prompting has been employed for domain adaptation (Ben-David and Oved et al., 2022). However, in many other scenarios, one may not have any data of the target domain in training but is still asked to build a precise model for the unseen target domain.

Domain Generalization has been proposed to address this problem. Incorporating data from several source domains, DG aims to learn models that generalizes well on unseen target domains. **Multi-task learning** (Caruana et al., 1997) can be seen as DG method which jointly optimizes models and share representation on several related tasks to generalize better on specific one. McCann et al., 2018 showed that promising results can be achieved while training language model on various tasks simultaneously. Collins et al., 2020 extended this with the T5 by leveraging a unified text-to-text format to attain state-of-the-art results on a wide variety of NLP tasks with a single model. In our work we collected multiple source datasets which differ in many aspects (see Table 2), thus can be considered as different tasks and leverage the model to fully explore the general semantic representations of passage-question. Another common field of research in DG is domain-invariant representation learning, which aims to reduce the representation discrepancy between multiple source domains in a specific feature space to be domain invariant so that the learned model can have a generalizable capability to the unseen domain. Arjovsky et al. (2018) applied Invariant risk minimization (IRM) to enforce the optimal classifier on top of the representation space to be the same across all domains. However, Dranker et al,.(2021) showed its limitation in practice, and Feder and Wald et al., (2021) showed that the right way to obtain this optimal classifier is with multi-domain calibration objective. Nevertheless, most work and research investigated in domain-invariant representation focus on **Domain Adversarial Training.** It was originally introduced in the work of Ganin et al., (2016) inspired by Adversarial Training in image generation (Goodfellow et al., (2014)) , achieving state-of-the-art performance on domain adaptation tasks like digit image classifications and document sentiment analysis. Their method relies on the theory of domain adaptation of Ben David et al., suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training and test domains. it can be adopted to DG

	Dataset	Question	Context	|Q|	|C|
In-domain	SQ	Crowd-sourced	Wiki	11	137
	NewsQ	Crowd-sourced	CNN articles	8	599
	NatQA	Search logs	Wiki	9	153
	Hotpot	Crowd-sourced	Wiki	22	232
out-of-domain	BA	Domain-expert	Science articles	11	248
	DP	Crowd-sourced	Wiki	11	243
	DR	Crowd-sourced	Movie	9	681
	RA	Domain-expert	Exams	12	349
	RE	Synthetic	Wiki	9	30
	TQ	Domain-expert	Textbook	11	657

Table 2: The four in-domain used for training and 6 out-of-domain used for evaluation. |Q|, |C| the average length in tokens in the question, context respectively.

in such manner: Let's consider k domains $\{D_1, .., D_k\}$ where we have $\{x_{i,j}, y_{ij}\}_{j=1}^{n_i}$ samples in each $domain\ D_i$. A feature extractor modeled by $\phi(\cdot)$, followed by a classifier upon the representation of the input sample. In the regular fine-tuning method, we would minimize the objective with a relevant loss function l :

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i} l(F(\phi(x_{i,j})), y_{i,j})$$

However, with domain adversarial training we will minimize the objective in the form of:

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i} l(F(\phi(x_{i,j})), y_{i,j}) - \lambda l_{adv}(D(\phi(x_{i,j})), i)$$

where D is a domain classifier. The goal in adversarial domain generalization is to maximize the adversarial loss (l_{adv}) to get invariant feature representation that even a strong discriminator D can't distinguish between domains. while we have theoretical justification for the objective in the DA case Ganin et al., (2016). Deng et al. analyzed the objective mentioned above and gave precise characterization of this methods limit for the case of DG as the number of unseen domains k grows.

In related work, Lee at al., (2019) applied this method for QA Bert model, minimizing Kullback-Leibler divergence (Kullback et al.,1959) between uniform distribution over the source domains and the discriminator prediction. In our approach, we treat the domains invariance as a binary classification problem, thus maximizing the loss of the domain classifier by reversing its gradients using Gradient Reversal Layer (GRL).

3 Models

We consider RC task in which we are given K source domain $\{D_1, D_2, .., D_K\}$ each domain containing samples of {paragraph, question, answer}.

Bert-SQuAD: we fine-tuned Bert base uncased (Devlin et al., 2019) on the task of RC with SQuAD (Rajpurkar et al., 2016) dataset. The input is processed before entering the model as [CLS] token is added at the beginning of the question and a [SEP] token is inserted in between. Then for each of the paragraph words, the probability of each word being the start/end-word is calculated using fully connected layer, followed by a SoftMax. The fine-tuning is conducted on the SQuAD dataset. In the fine-tuning phase we minimize the cross-entropy loss for the start and end tokens.

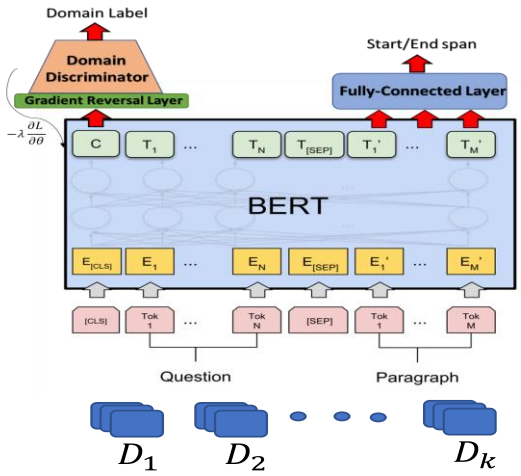

Figure 1: Domain Adversarial QA model. The finetuned Bert tries to learn answering the question while fooling the domain discriminator.

Bert-All: similar to the Bert-SQuAD but the fine-tuning is conducted on 4 different source domains, with 25K samples each.

Domain Adversarial QA model: In extend to the Bert-All model, we add a domain discriminator connected to the [CLS] embedding in the final layer, see Figure 1. We use the [CLS] token since it is used for sentence classification (Devlin et al., 2019) thus can be used to discriminate the question-paragraph domain. Our goal is to build a model unfitted to a specific domain, but rather a model which produce generalized feature representation for question-paragraph pair. Thus, in training, instead of backpropagating the gradient of the domain loss $\frac{\partial L_d}{\partial \theta}$, we backpropagate $-\lambda\frac{\partial L_d}{\partial \theta}$ using Gradient Reversal Layer. We used increasing value of λ from 0 to 0.1.

4 Experiment

4.1 Data

We leverage pre-processed datasets from MRQA 2019 shared task, using as source domains: **SQuAD** (SQ) (Rajpurkar et al.) as the basis format, where crowdworkers write questions on Wikipedia paragraphs. **NewsQA** (Trischler et al.) question-paragraph based on CNN articles, **HotpotQA** (Yand et al.),QA based Wikipedia which requires multi-hop reasoning.

Model	BioASQ		DROP		DuoRC		RACE		Relation Extraction		Textbook QA		AVG	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Bert-SQ	26.13	41.35	10.78	17.31	37.18	44.75	21.07	30.38	61.5	75.11	30.54	38.07	31.2	41.16
Bert-All	26.01	42.66	16.17	25.23	35.64	45.27	21.51	32.17	64.28	78.6	29.81	38.17	32.23	43.68
BertDA	**28.52**	**43.77**	**16.23**	**26.08**	**41.17**	**49.53**	**22.26**	**32.88**	**66.82**	**80.25**	**30.34**	**38.89**	**34.22**	**45.63**

Table 3: The models performance on the validation out-of-domain sets

NaturalQuestions (NatQ) (Kwiatkowski et al.), Questions are collected from information-seeking queries to the Google search engine by real users under natural conditions. As we can see in Table 2 the datasets differ in the following ways: **Passage distribution,** where context passages come from different sources, **question distribution,** with different styles and sources, and **joint distribution,** as the relationship between the passage and question varies. We validated the models with 6 different out-of-domain datasets, which are **BioASQ** (BA) (Tsatsaronis et al.), biomedical semantic indexing and question answering, **DROP** (DP) (Dua et al.) similar to SQuAD where the questions focus on quantitative reasoning, **DuoRC** (DR) (Saha et al.) movie plots QA, **Race** (RA) (Lai et al.), questions from RC high school exams. **RelationExtraction** (RE) (Levy et al.) , a slot-filling dataset relations among entities, transformed into question-answer pairs, and **TextbookQA** (TQ) (Kembhavi et al.) which is collected from lessons from middle school Life Science, Earth Science textbooks.

4.2 Setup

We trained each model as mentioned in Section 4 and evaluated its performance on the out-of-domain datasets.The maximal length of sequence for Bert model is 512, but because of limited computational abilities of our machine (4 CPU, 16GB RAM) and to get conceivable training time, a tradeoff had to be made, so we set the sequence length as 256 with document stride of 64 and batch size of 8. Data feeding order could have an impact on the model's performance as studied in (Sajjad et al.). However, we believe the order will affect each target domain differently, but we aim for generalization for all target domains, so we won't investigate this issue. Whenever training on different domains, we shuffled the datasets to reduce any reliance on the order of the data. We set 1e-5 learning rate with Adam optimizer for 1 epoch

5 Results and Analysis

Table 3 shows the performance of the implemented models on the evaluation out-of-domain datasets. We show the Exact Match (EM) and F1 Score (F1) obtained on each set. First, we can see that Bert-All gave better results that Bert-SQuAD, both approximately with the same amount of training examples (100K vs 87K). Thus, we can conclude that training with multiple, varying source datasets improves performance consistently on QA generalization. Finally, we can see that BertDA outperforms the baseline models on all of the evaluation datasets, achieving ~3%,4.5% gain in EM, F1 respectively over the naïve baseline (Bert-SQuAD) and ~2% in both EM, F1 over the better baseline (Bert-All), without the need of additional training data or stronger model. These results demonstrate the effectiveness of our approach compared to the common methods.

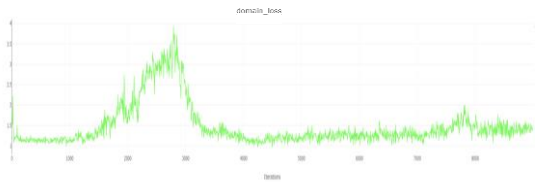

Figure 2: l_{adv}, the domain cross entropy loss in the training process. We used GRL to increase the loss. We can see a peak in the loss in early iterations, but the model wasn't giving enough good QA performance yet.

6 References

Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

Sen, Priyanka, and Amir Saffari. "What do Models Learn from Question Answering Datasets?." *arXiv preprint arXiv:2004.03490* (2020).

Ben-David, Shai, et al. "A theory of learning from different domains." *Machine learning* 79.1 (2010): 151-175.

Ben-David, E., Oved, N., & Reichart, R. (2021). Pada: A prompt-based autoregressive approach for adaptation to unseen domains. *arXiv preprint arXiv:2102.12206*.

Lee, Seanie, Donggyu Kim, and Jangwon Park. "Domain-agnostic question-answering with adversarial training." *arXiv preprint arXiv:1910.09342* (2019).

Dranker, Y., He, H., & Belinkov, Y. (2021). IRM---when it works and when it doesn't: A test case of natural language inference. *Advances in Neural Information Processing Systems*, *34*.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).

Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* 32 (2019).

Ma, Xiaofei, et al. "Domain adaptation with BERT-based domain classification and data selection." *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 2019.

Blitzer, J., McDonald, R., & Pereira, F. (2006, July). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 120-128).

Hacohen, Guy, and Daphna Weinshall. "On the power of curriculum learning in training deep networks." *International Conference on Machine Learning*. PMLR, 2019.

Ziser, Yftah, and Roi Reichart. "Pivot based language modeling for improved neural domain adaptation." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.

Deng, Zhun, et al. "A Theoretical View of Adversarial Domain Generalization in the Hierarchical Model Setting."

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The journal of machine learning research* 17.1 (2016): 2096-2030.

Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).

Wald, Y., Feder, A., Greenfeld, D., & Shalit, U. (2021). On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, *34*.

McCann, Bryan, et al. "The natural language decathlon: Multitask learning as question answering." *arXiv preprint arXiv:1806.08730* (2018).

Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).

Trischler, Adam, et al. "Newsqa: A machine comprehension dataset." *arXiv preprint arXiv:1611.09830* (2016).

Kwiatkowski, Tom, et al. "Natural questions: a benchmark for question answering research." *Transactions of the Association for Computational Linguistics* 7 (2019): 453-466.

Yang, Zhilin, et al. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." *arXiv preprint arXiv:1809.09600* (2018).

George Tsatsaronis, et al. 2015. An overview of the bioasq largescale biomedical semantic indexing and question answering

Dheeru Duaet al. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In NAACL.

Amrita Sahaet al. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In ACL.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In EMNLP

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

S. Kullback. Information Theory and Statistics. Wiley, New York, 1959.

Aniruddha Kembhavi, et al. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In CVPR.