On Committee Representations of Adversarial Learning Models for Question-Answer Ranking

Sparsh Gupta
University of California San Diego
San Diego, CA
spg005@ucsd.edu

Vitor Carvalho
Intuit AI
San Diego, CA
vitor_carvalho@intuit.com

Abstract

Adversarial training is a process in Machine Learning that explicitly trains models on adversarial inputs (inputs designed to deceive or trick the learning process) in order to make it more robust or accurate. In this paper we investigate how representing adversarial training models as committees can be used to effectively improve the performance of Question-Answer (QA) Ranking. We start by empirically probing the effects of adversarial training over multiple QA ranking algorithms, including the state-of-the-art Multihop Attention Network model. We evaluate these algorithms on several benchmark datasets and observe that, while adversarial training is beneficial to most baseline algorithms, there are cases where it may lead to overfitting and performance degradation. We investigate the causes of such degradation, and then propose a new representation procedure for this adversarial learning problem, based on committee learning, that not only is capable of consistently improving all baseline algorithms, but also outperforms the previous state-of-the-art algorithm by as much as 6% in NDCG (Normalized Discounted Cumulative Gain).

1 Introduction

Question Answer (QA) ranking, or the task of accurately ranking the best answers to an input question, has been a long-standing research pursuit with practical applications in a variety of domains. Popular examples of such applications are customer support chat-bots, community question answering portals, and digital assistants like Siri or Alexa Yih and Ma (2016).

Early work on QA ranking relied heavily on linguistic knowledge (such as parse-trees), feature engineering or external resources (Wang and Manning, 2010; Wang et al., 2007; Yih et al., 2013). Yih et al. (2013) constructed semantic features from WordNet and paired semantically related words based on these features and relations. Wang and Manning (2010); Wang et al. (2007) used syntactic matching between question and answer parse trees for answer selection. Other proposals used minimal edit sequences between dependency parse trees as a matching score between question and answer (Heilman and Smith, 2010; Severyn and Moschitti, 2013; Yao et al., 2013).

The majority of the recent developments for QA ranking algorithms are based on deep learning techniques, and fall into two different classes of models: *representation-based* or *interaction-based*. In *representation-based* models, both question and answer are mapped to the same representation space via network layers with shared weights, and a final relevance or matching score is computed from these representations (Bowman et al., 2015; Tan et al., 2015; Huang et al., 2013; Tan et al., 2016; Wang et al., 2016). In *interaction-based* models, the network attempts to capture multiple levels of interaction (or similarity) between question and answer (Hu et al., 2014; Pang et al., 2016; Yu et al., 2018). The final relevance/matching score can be computed out of the partial similarities derived from the multiple interactions.

Recent results have indicated that representation-based models, when used with attention layers to focus on relevant parts of the question and answer, tend to outperform interaction-based models (Tan et al., 2016; Wang et al., 2016). The recently proposed Multihop Attention Network (MAN) model (Tran and Niederee, 2018) currently achieves state-of-the-art performance on ranking tasks by using sequential attention (Brarda et al., 2017) over multiple attention layers. This model is discussed in detail in Section 2.2.

Adversarial training and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been successfully applied to Computer Vision (Karras et al., 2017; Isola et al., 2017; Zhu et al., 2017; Kelkar et al., 2018) and Natural Language Processing (Lin et al., 2017) applications, but only sparsely studied in Information Retrieval

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 218–223
Florence, Italy, August 2, 2019.

tasks. As described by Wang et al. (2017), adversarial training in Information Retrieval can be approached by having a *generator model* to sample difficult adversarial examples which are passed to a *discriminator model* that learns to rank on increasingly difficult adversarial examples. This adversarial training process in principle can lead to increased robustness and accuracy of the final ranking model.

We show that in general most models do benefit from adversarial training, with a clear increase in ranking metrics. However, we also observed that not all types of models benefit from straightforward adversarial training. For instance, Multihop Attention Network often displayed worse results with adversarial training. In such cases, we observed that the model was excessively compensating to the current adversarial training data batch and often forgetting previous batches, thus reducing its performance on test data.

To help address this issue, we propose a novel committee representation to adversarial modeling for QA ranking that can be applied to any underlying ranking algorithm. Not only does it address the observed "overfitting" that may occur during adversarial training, but provides an improvement to all baseline QA ranking models we tested. In particular, we introduce a new state-of-the-art model *AdvCom-MAN* (**Adv**ersarial **Com**mittee - **M**ultihop **A**ttention **N**etwork) for QA ranking that displays, to the best of our knowledge, state-of-the-art results on four different datasets for QA Ranking.

2 Approaches

We introduce in this section the various algorithms and techniques that we use to investigate the use of adversarial training to QA ranking.

2.1 Baselines

We used two recently proposed interaction based models as baselines, meaning that these models work on interaction (lexical similarities) between the question and answer text.

- **Match Pyramid -** Proposed by Pang et al. (2016), this model uses convolution layers on the "interaction matrix" formed by taking the dot product of embeddings of question words with answer words.
- **Deep Matching Net -** This model was proposed by Yang et al. (2018) and was originally meant for multi-turn conversations, but we adapted a version of it for single turn conversation which can also work as QA ranking. Similar to Match Pyramid in most aspects, this model uses 2 interaction matrices, the second one being constructed in a similar fashion of dot products between embeddings obtained by a Bi-directional Gated Recurrent Unit (Bi-GRU).

2.2 Multihop Attention Network (MAN)

This model was recently proposed by Tran and Niederee (2018) as state-of-the-art in QA ranking tasks. A bi-directional LSTM layer first generates the representations of question and answer words. Following this, multiple "hops" or multiple layers of attention are used to get attended representations of the question and answer at each attention layer. This is accomplished by using sequential attention (Brarda et al., 2017) at every layer. The intuition for this architecture is to compare and analyze the question and answer from different points of view by focusing on different parts of the text in each hop. At each attention hop, cosine similarity is computed between the question and answer representations. The final matching score is calculated by summing the cosine similarities at each layer (Equation 1).

$$sim(q, a) = \sum_{k} \cos\left(o_q^{(k)}, o_a^{(k)}\right) \quad (1)$$

Here $o_q^{(k)}$ and $o_a^{(k)}$ refer to the question and answer representations after the kth hop in the network. All the models are trained by minimizing the Hinge Loss (Equation 2) with L2 regularization.

$$L = \max\{0, M - sim(q, a_+) + sim(q, a_-)\} \quad (2)$$

where M is the margin, q is the input question, and a_+ and a_- are correct and incorrect answers to q respectively.

2.3 Vanilla Adversarial Learning

IRGAN (Information Retrieval Generative Adversarial Networks) has been recently proposed as a generic adversarial learning framework for several Information Retrieval tasks (Wang et al., 2017). In this paper we focus the adaptation of IRGAN to pairwise cases, which adapt well to the QA ranking problem.

IRGAN uses the same minimax game idea as a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) but uses different objective functions for the generator and discriminator. The generator and discriminator of a GAN are initialized with a model pre-trained on original training dataset. In a ranking task setting, the job of generator is to sample difficult incorrect answers given an input question and correct answers for it. The discriminator then learns to rank this difficult dataset.

Since sampling is a non-differentiable operation, the generator cannot be trained using backpropagation by error signal from the discriminator. Hence a Reinforcement Learning strategy (Williams, 1992; Yu et al., 2017) is used to train the generator where the objective of the generator is to maximize its reward (Equation 3).

$$L_{Gen} = \frac{1}{K}\sum_{k=1}^{K} \log\Big(g_\theta(d_k|q)\Big) \times reward \quad (3)$$

$$L_{Dis} = \frac{1}{K}\sum_{k=1}^{K} hinge(q, a_+, d_k) \quad (4)$$

where d_k is the kth adversarial incorrect answer, g_θ is the generator score for kth answer and question q, and $reward$ is given by Equation 5.

$$reward = 2\bigg(\sigma\Big(hinge(q, a_+, d_k)\Big) - 0.5\bigg) \quad (5)$$

and $hinge(q, a_+, d_k)$ is hinge loss (Equation 2). Detailed derivation of these equations has been given in the paper that proposes IRGAN (Wang et al., 2017) and has not been delineated here to focus on more relevant aspects of the paper.

2.4 Adversarial Committee Learning

In Section 3 we present details on our adversarial training experiments. Surprisingly, a number of experiments showed results with high variance that seemed somewhat contradictory to the expectation that adversarial training should boost model performance (or at least not deteriorate it). After careful observation, we noticed that as adversarial training progresses, some models may start overfitting to the adversarial examples in the current batch, and partially forgetting the original training data, which consequently leads to a deterioration of test data ranking performance.

This led to the development of a novel adversarial committee learning strategy that boosts the model performance, irrespective of the nature of model itself. The idea is to sample the model at regular intervals during adversarial training, including the pre-trained model and the fully trained model after adversarial training. The intuition behind this strategy is that the sampled models have decision boundaries that are fit to different proportions of the original dataset and the adversarial dataset, consequently creating a committee of diverse decision makers. This idea is very similar to the work of Elsas et al. (2008) where perceptrons are sampled during training to be a part of the decision making committee. This work uses the original dataset to form the committee, as opposed to adversarial dataset which is used in our model.

During prediction, given a question q and a candidate answer a, the matching score between them $score(q, a)$ is computed as shown in equation 6.

$$score(q, a) = \sum_{i=1}^{N} w_i h_i(q, a) \quad (6)$$

where $h_i(q, a)$ is the matching score between q and a given by ith model, and w_i is the weight assigned to ith model. This weight is computed by first recording the performance metric (MRR, MAP, NDCG@5, etc.) on the validation dataset for all models, and then normalizing them to 1. We sampled these N models at regular intervals during adversarial training process. For our experiments, we sampled the models at every 3rd epoch to be a part of the committee. We tried different sampling strategies but this one worked out to be the best trade-off between committee performance and run-time during prediction.

The results show that this strategy works for all types of models and it overcomes the overfitting issues observed with vanilla adversarial training.

3 Experiments

3.1 Datasets

We use four datasets, belonging to factoid and non factoid categories to evaluate the proposed strategy. **WikiQA** is an open domain question answering dataset that was introduced by Yang et al. (2015) and has now become a very popular benchmark dataset for QA ranking systems. Feng et al. (2015) recently released a large non-factoid QA dataset for insurance domain - **Insurance QA**. Like Tran and Niederee (2018), we use

Model	WikiQA (19k/ 2.5k/ 5.8k)		Insurance QA (926k/ 724k/ 650k)		FiQA (700k/ 300k/ 300k)		Tax Domain QA (42k/ 14k/ 14k)
	NDCG@5	MRR	test-1	test-2	NDCG@5	MRR	prec@1
Match Pyramid	0.6628	0.6258	0.4571	0.4036	0.3423	0.4571	0.5767
+ Vanilla Adversarial	0.6939	0.6675	0.5269	0.4602	0.3715	0.4859	0.6437
+ Adversarial Committee	0.6987	0.6748	0.5307	0.4751	0.3812	0.4866	0.6568
Deep Matching Net	0.6922	0.6533	0.6135	0.5498	0.3972	0.4963	0.6601
+ Vanilla Adversarial	0.6952	0.6692	0.6464	0.6007	0.4114	0.5149	0.6636
+ Adversarial Committee	0.7051	0.67	0.688	0.625	0.4157	0.5191	0.6863
MAN	0.7328	0.7134	0.7032	0.668	0.4312	0.5153	0.7927
+ Vanilla Adversarial	0.7337	0.711	0.6951	0.6509	0.3844	0.465	0.7975
+ Adversarial Committee	**0.7402**	**0.7205**	**0.7267**	**0.6814**	**0.4601**	**0.5318**	**0.8029**

Table 1: Experimental results of adversarial learning on different datasets; Models have been evaluated on NDCG@5 and MRR for WikiQA and FiQA, and on Precision@1 for Insurance QA test sets 1 and 2, and Tax Domain QA

version 1 of this dataset which is divided into a training, validation and 2 test sets. **FiQA**, the financial domain non-factoid dataset[1] was released recently and built by crawling data from Reddit, StockTwits and StackExchange. **Tax Domain QA** dataset was obtained from a popular tax domain question answering platform. Each question had only one correct answer, so we create an answer pool for each question by randomly sampling incorrect answers from the entire collection of answers. Table 1 shows the size of these datasets in terms of QA pairs in the (train/ validation/ test) format.

We evaluate these datasets on different metrics. For the datasets that have only 1 correct answer in the answer pool associated with every question, we use precision@1 since it the the most suitable metric. For datasets that have multiple correct answers, more comprehensive metrics such as Mean Reciprocal Rank (MRR) and NDCG@5 have been used that evaluate the model's ability to retrieve not only the most relevant, but all relevant answers.

3.2 Results

In this Section we present our experimental results on running adversarial training techniques over different QA ranking baseline algorithm (from Section 2) on multiple QA datasets (from Section 3.1). For all the models, we use the prefix Adv- when we refer to their variants trained by vanilla adversarial learning, and AdvCom- when they are trained by adversarial committee learning.

From Table 1 it can be seen that the metrics show fairly similar trends across all datasets [2].

Based on the results from all our experiments, we observed that the overall performance of Multihop Attention Network and its variants was the best of the three model types, followed by Deep Matching Network and its variants. Match Pyramid and its variants had the lowest performance scores in general, except for a few anomalous cases where AdvCom-Match Pyramid performed better than few variants of Deep Matching Network on some of the datasets. Furthermore, the results also show that while vanilla adversarial learning provides a significant boost in model performance for Match Pyramid and Deep Matching Network, the performance boost by adversarial committee learning was much better. However for MAN, vanilla adversarial learning significantly worsens the base model performance for most datasets. Our hypothesis is that since MAN has a higher capacity, it overfitted to adversarial training samples thereby forgetting some of its knowledge from original dataset. Adversarial committee learning however addresses this issue and improves the performance of base MAN by creating a committee of diverse decision makers that contain knowledge from both original and adversarial dataset. Consequently, the AdvCom-MAN establishes new state-of-the-art standards for QA ranking models on almost all datasets.

[1] https://sites.google.com/view/fiqa

[2] All row differences are statistically significant based on 95% bootstrap confidence interval

4 Discussion and Conclusion

In this work we provided a large empirical investigation on the effects of adversarial training applied to deep QA ranking models. We explored both interaction-based and representation-based QA ranking models, including the previous state-of-the-art Multihop Attention Network algorithm. While in most cases adversarial training proved to be indeed beneficial to QA ranking, we observed that in some cases overfitting to the adversarial training data during adversarial learning could lead to lower than expected ranking performance.

We then proposed a new adversarial learning representation based on a committee strategy to improve QA ranking performance. We showed that the adversarial committee technique was able to boost the performance of all models and in all datasets. As a result, an adversarial committee applied to the MAN algorithm presented the new state-of-the-art results for QA ranking on all datasets tested on this paper, including WikiQA, InsuranceQA and FiQA.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642. ACL.

Sebastian Brarda, Philip Yeres, and Samuel R. Bowman. 2017. Sequential attention: A context-aware alignment function for machine reading. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 75–80. ACL.

Jonathan L. Elsas, Vitor R. Carvalho, and Jaime G. Carbonell. 2008. Fast learning of document ranking functions with the committee perceptron. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*. ACM.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *Workshop on Automatic Speech Recognition and Understanding*, pages 813–820. IEEE.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1011–1019. ACL.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 36th ACM International Conference on Information and Knowledge Management*, pages 2333–2338. ACM.

Philip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexi A. Efros. 2017. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196.

Sachin Kelkar, Chetanya Rastogi, Sparsh Gupta, and G.N. Pillai. 2018. Squeezegan: Image to image translation with minimum parameters. In *2018 IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognitionm. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2793–2799. AAAI.

Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic feature engineering for answer selection and extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 458–467. ACL.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473. ACL.

Nam Khanh Tran and Claudia Niederee. 2018. Multi-hop attention networks for question answer matching. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334. ACM.

Bingning Wang, Kang Liu, and Jun Zhao. 2016. Inner attention based recurrent neural networks for answer selection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1288–1297. ACL.

Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 515–524. ACM.

Mengqiu Wang and Christopher Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1164–1172. ICCL.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32. ACL.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:3-4:229–256.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 245–254. ACM.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2013–2018. ACL.

Xuchen Yao, Benjamin Van Durme, Chris Callison Burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867. ACL.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1744–1753. ACL.

Wen-tau Yih and Hao Ma. 2016. Question answering with knowledge base, web and beyond. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1219–1221, New York, NY, USA. ACM.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 682–690. ACM.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.