On the Effectiveness of Minimal Context Selection for Robust Question Answering

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine learning models for question-answering (QA), where given a question and a passage, the learner must select some span in the passage as an answer, are known to be brittle. By inserting a single nuisance sentence into the passage, an adversary can fool the model into selecting the wrong span. A promising new approach for QA decomposes the task into two stages: (i) select relevant sentences from the passage; and (ii) select a span among those sentences. Intuitively, if the sentence selector excludes the offending sentence, then the downstream span selector will be robust. While recent work has hinted at the potential robustness of two-stage QA, these methods have never, to our knowledge, been explicitly combined with adversarial training. This paper offers a thorough empirical investigation of adversarial robustness, demonstrating that although the two-stage approach lags behind single-stage span selection, adversarial training improves its performance significantly, leading to an improvement of over 22 points in F1 score over the adversarially-trained single-stage model.

1 Introduction

Over the last few years, passage-based question-answering (commonly known by the misnomer *reading comprehension* and hereafter denoted QA) has emerged as a popular and challenging task that tests the capabilities of today's deep-learning models. Given a question and an associated context, such as a passage or a document, QA typically requires either selecting a span from the context as an answer, choosing one among multiple answer choices (classification) or generating an answer from scratch. In this paper, we focus on the span-selection variant. Recent progress in this field has been spurred by the availability of many large-scale datasets [13, 15, 8]. Several complex neural models [2, 6, 18] have shown promising results on this challenging task, some even purporting to beat the reported human performance on some datasets [4].[1]

However, *performance* here denotes only accuracy on i.i.d. holdout data. While humans exhibit a much greater ability to generalize off-manifold, supervised learning models tend to break, especially under adversarial perturbations, as demonstrated by [14] with images. Recently, Jia and Liang [7] showed that neural QA models suffer from an analogous vulnerability by appending a single nuisance sentence to the context of passages from the SQuAD 1.1 dataset [13] and fooling many state-of-the-art models into selecting the wrong span (Figure 1). While humans simply ignore the intruding sentence, QA models are easily fooled, raising concerns regarding whether these models are sufficiently robust to be deployed for QA tasks in the wild, or if they depend too heavily upon spurious correlations in

[1] While many QA datasets have emerged, they are often synthetically generated and their difficulty remains poorly characterized. Recent papers have shown that for some datasets, simple baselines using just a few hand-engineered features [1], or ignoring either the question or passage [9] can perform surprisingly well.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

the training and development datasets. These demonstrations underscore the necessity for evaluating QA models under adversarial conditions.

Only a few subsequent papers have followed up on [7], proposing solutions to make QA models more robust to such adversarial attacks. Recently, Min et al. [10] proposed a two-stage model consisting of both a sentence selector and a span selector. They showed that providing a minimal context, consisting of just few relevant sentences to the span selector, offers benefits not only in terms of interpretability (by identifying the relevant pieces of evidence) and computational efficiency, but also results in greater robustness to the aforementioned adversarial attack. This is a promising direction towards making QA models more robust, since achieving robustness in the overall system requires only that we make the context selection model robust. So long as the context selector filters out irrelevant sentences (including the adversarial sentence) the downstream model will be safe.

In this work, we investigate this two-stage approach (minimal context selection followed by span selection) finding that the approach is not, out of the box, more robust than the single-stage approach (span selection)—the accuracy of the minimal context selection model suffers under adversarial evaluation and earlier reported gains appear to stem partly from an artifact in the evaluation. However, we find that sentence selector can be made more robust through adversarial training [5], and importantly, perform significantly better than an adversarially-trained single-stage model.

Article: Super Bowl 50
Paragraph: "Peyton Manning became the first quarter-back ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."
Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: A state-of-the-art QA model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue). Example taken from [7].

2 Methods

Span Selection Models: In our investigation, we focus on two span-selection models: DrQA [2] and the Mnemonic Reinforced Reader [6]. DrQA uses self-attention [16] over the question tokens to learn a fixed-length question representation that is then used to score potential spans. The Mnemonic Reinforced Reader uses several layers of co-attention between the question and the context, memorizing and utilizing attention output from previous layers to compute the later ones. Additionally, both models employ hand-crafted features like Part-of-Speech (PoS) tags, Named Entity Recognition (NER) tags, and other lexical features, in order to achieve competitive performance on the task. We follow the reported architecture and hyperparameter settings exactly, referring the readers to the source papers for more details.

Sentence Selection Model: We base our implementation of the sentence selector model on the DrQA architecture [2]. As in DrQA, we encode the question and every sentence in the passage independently using a BiLSTM. We then use self-attention to compute fixed-sized representations of the question q^{enc} and each sentence d_i^{enc} $\forall i \in \{1, ..., N\}$, where N is the number of sentences in the passage. We then compute a scalar score s_i using a bilinear transformation $s_i = q^{enc} W d_i^{enc}$. These scores are then normalized over the passage using a softmax. For supervision, we use the sentences containing the answer span as gold sentences and minimize a cross-entropy loss objective. Figure 2 contains a schematic diagram of the two-stage approach.

Adversarial Training: Jia and Liang [7] produce an adversarial sentence for a given passage and question according to the following procedure: (1) The question is perturbed by (i) substituting antonyms for common question words and (ii) substituting nearest neighbours (determined via Glove [11] embeddings) for named entities, to reduce the likelihood of the gold answer being the correct answer to the perturbed question. For example, *"What city did Tesla move to in 1880?"* could become *"What city did Tadakatsu move to in 1881?"*; (2) Generate a fake answer that matches the type of the

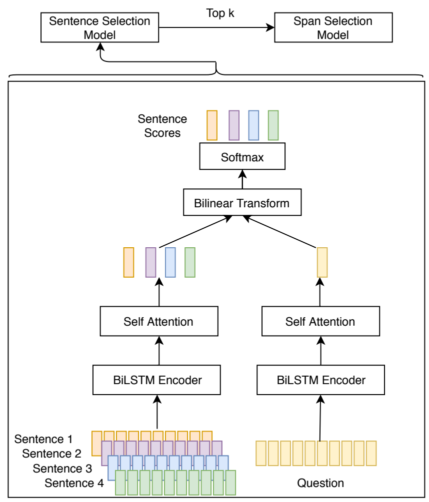

Figure 2: The two-stage pipeline with architecture of the sentence selector model.

original answer (e.g., Prague $\rightarrow$ Chicago, etc.); (3) The fake answer and the altered question are combined into a declarative sentence based on a set of handcrafted rules (*"Tadakatsu moved to the city of Chicago in 1881."*).

In this paper, we focus on adversarial training through data augmentation: for every training example, $x = (p, q, a)$, where p is the paragraph, q is the question and a is the answer, we introduce adversarially perturbed example, $x' = (p', q, a)$, and train both the span selection and sentence selection models on $\mathcal{D}_{aug} = \{x_i | i = 1, ..., N\} \bigcup \{x'_i | i = 1, ..., N\}$, where N is the size of the original training dataset. We focus on two different adversaries for training: ADDSENT, in which a distractor sentence similar to the question is appended to the end of the paragraph, and ADDRANDOM (similar to ADDSENTDIVERSE in [17]), in which the position within the paragraph, where the distractor sentence is added is chosen uniformly at random. Since the datasets considered in our paper do not require more than one sentence in order to answer the question for a large fraction of examples (as discovered by [10]), adding the distractor sentence anywhere in the paragraph shouldn't make the reasoning process significantly more difficult as compared to adding it at the end. We adversarially evaluate all our models on ADDSENT, ADDRANDOM and additionally on ADDMODSENT, in which the distractor sentence is added to the beginning of the paragraph.

3 Experimental Evaluation

Setup: We train the sentence selector and the span selection model on SQuAD [13], which contains $\sim$5-sentence long contexts from Wikipedia articles. We train two different QA models: DrQA [2] and Mnemonic Reader [6], comparing the two-stage minimal context selection approach (MINIMAL) to single-stage models using the full context (FULL). Our metrics measure (i) how frequently predicted spans exactly match the gold span (EM) and (ii) an F1 score calculated by treating the spans as bags of words. We measure the performance of the sentence selector model via top-k accuracy, i.e., how often the oracle sentence is among the top-k selected sentences. We use the PoS, NER and lexical features for both models. In the two-stage set-up, the top-k sentences from the sentence selector are passed on to the span selection model. We choose $k = 1$ for SQuAD dataset.

Results: Our results are summarized in Table 1. Without adversarial training, MINIMAL lags behind the FULL model by a few points on the original development data, but this difference is exacerbated on the adversarial development sets. This indicates that the two-stage approach is not robust to adversarial inputs without any adversarial training. This is evident from Table 2, where upon adversarial evaluation, sentence selector top-k accuracy drops by over 50 points. In fact, it selects the distractor sentence in over 95% of the instances where it fails to select the oracle sentence on ADDSENT dataset for SQuAD.

Under adversarial training with the ADDSENT adversary, both the MINIMAL and FULL improve significantly as measured on the ADDSENT test set, but MINIMAL still lags behind FULL by 6.9 points (resp. 4.8 points) for DrQA (resp. Mnemonic Reader). However, MINIMAL beats FULL on

AddModSent and AddRandom adversaries by 17.3 points (resp. 6.8 points) for DrQA (resp. Mnemonic Reader). This indicates that through adversarial training on AddSent, the Full model has learnt to ignore the last sentence in the context, and as a result, it performs worse on adversaries where the distracting sentence doesn't occur in the end. When trained on AddRandom adversary, the Minimal beats Full by 25.5 points (resp. 22.8 points) on all adversarial test sets for DrQA (resp. Mnemonic Reader), thereby indicating that Minimal can be made more robust to adversarial examples as compared to Full through adversarial training.

SQuAD + DrQA									
Setting	**Model Type**	**Dev**		**AddSent**		**AddModSent**		**AddRandom**	
		F1	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**
Original	Full	**78.8**	**69.4**	**42.4**	**36.4**	**52.4**	**44.8**	**50.4**	**42.6**
	Minimal	76.8	67.8	40.9	35.3	45.9	38.6	44.8	37.3
Adv. Training	Full	**78.2**	**68.7**	**75.0**	**65.9**	52.1	44.8	60.7	52.0
(AddSent)	Minimal	76.8	67.8	68.1	60.4	**71.9**	**63.2**	**75.6**	**65.9**
Adv. Training	Full	**78.6**	**69.1**	45.0	31.5	40.8	25.5	46.9	31.9
(AddRandom)	Minimal	76.6	67.4	**65.3**	**57.7**	**69.4**	**60.7**	**74.6**	**64.7**
SQuAD + Mnemonic Reader									
Original	Full	**81.4**	**72.6**	**45.6**	**39.8**	**47.9**	**42.0**	**45.6**	**38.6**
	Minimal	77.9	71.5	40.6	35.8	42.4	37.1	44.9	37.8
Adv. Training	Full	**80.6**	**71.5**	**73.6**	**64.2**	51.6	45.2	71.4	62.8
(AddSent)	Minimal	77.9	69.1	68.8	62.1	**62.6**	**56.0**	**73.9**	**63.8**
Adv. Training	Full	**81.3**	**72.6**	43.5	28.8	42.8	30.4	41.6	25.4
(AddRandom)	Minimal	77.5	68.6	**65.6**	**59.0**	**58.9**	**52.8**	**71.8**	**65.3**

Table 1: Performance of Full and Minimal context models on SQuAD dataset. We explore two different QA models: DrQA [2] and Mnemonic Reader [6].

Dataset	**Dev**	**AddSent**	**AddModSent**	**AddRandom**
SQuAD	90.1	45.9	36.9	36.9

Table 2: Sentence selector top-k accuracy for SQuAD ($k = 1$).

4 Related Work

Several prior works [10, 12, 3] consider sentence selection as a sub-task of question answering. [3] construct document summaries using reinforcement learning, feeding these summaries to the downstream QA model. [12] view extractive question answering as a search problem and iteratively refine the sentence, start and end spans. Both these models train the sentence selection and span selection models jointly.

In contrast, [10] take a two-stage approach and demonstrate robustness to adversarial examples. Several papers [6, 10] have evaluated their QA models built for the SQuAD dataset on the adversarial datasets provided by [7], but there hasn't been much work on how to utilize these adversaries to improve the robustness of the models. [17] train and test their models on multiple adversaries. However, they had to include additional semantic features to make the adversarially trained models robust. We show that in absence of any such additional features, the span selection model fails despite being adversarially trained, while the two-stage approach performs significantly better.

5 Conclusion and Future Work

This paper evaluates the adversarial robustness of two-stage QA models. We find that the approach remains susceptible to adversarial attacks. However, under adversarial training, the modular approach performs significantly better than the single-stage model (22 points in F1 on adversarial evaluation). Our findings add evidence that two-stage QA is a promising direction for building robust QA models. While this works presumes explicit supervision for training the sentence selector, we plan in future work to consider datasets and tasks that require implicit modeling of sentence selection.

References

[1] Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. 2016.

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

[3] Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. Coarse-to-fine question answering for long documents. In *Association for Computational Linguistics (ACL)*, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2014.

[6] Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. Reinforced mnemonic reader for machine reading comprehension. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[7] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[8] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*, 2017.

[9] Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[10] Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. Efficient and robust question answering from minimal context over documents. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[11] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global cectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[12] Jonathan Raiman and John Miller. Globally normalized reader. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[13] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[15] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Workshop on Representation Learning for NLP*, 2017.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[17] Yicheng Wang and Mohit Bansal. Robust machine comprehension models via adversarial training. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[18] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. QANet: combining local convolution with global self-attention for reading comprehension. *International Conference on Learning Representations (ICLR)*, 2018.