

Adversarial Examples for Evaluating Reading Comprehension Systems

Robin Jia, Percy Liang (Stanford Univ.) @ EMNLP2017

Reader: Saku Sugawara (Univ. Tokyo)
September 15, 2017 at SNLP9

Abstract

Research Question

- ✦ The extent to which reading comprehension (RC) systems truly understand language remains unclear.

Proposed Method

- ✦ An adversarial evaluation scheme for the RC dataset: testing whether systems can answer questions about paragraphs that contain adversarially inserted sentences.

Result

- ✦ The accuracy of sixteen published models drops from an average of 75% F1 score to 36%.
- Experiments demonstrate that no published open-source model is robust to the addition of adversarial sentences.

Introduction - RC Task

Article: Super Bowl 50

Paragraph: *"Peyton Manning became the first quarterback 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."*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

Answer: John Elway

Introduction - RC Task

Article: Super Bowl 50

Paragraph: *"Peyton Manning became the first quarterback 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."*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

Answer: John Elway

Introduction - Adversarial Sentence

Article: Super Bowl 50

Paragraph: *"Peyton Manning became the first quarterback-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?"*

Answer: John Elway

Introduction - Adversarial Sentence

Article: Super Bowl 50



Paragraph: *"Peyton Manning became the first quarterback-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 by BiDAF model under adversary: Jeff Dean

Adversarial Example

	Image Classification	Reading Comprehension
Possible Input		Tesla moved to the city of Chicago in 1880.
Similar Input		Tadakatsu moved to the city of Chicago in 1881.
Semantics	Same	Different
Model's Mistake	Considers the two to be different	Considers the two to be the same
Model Weakness	Overly sensitive	Overly stable

Framework for Adversarial Evaluation

$$\text{AdvAcc}(f) \stackrel{\text{def}}{=} \frac{1}{|D_{\text{test}}|} \sum_{(p,q,a) \in D_{\text{test}}} v(\text{Adv}(p, q, a, f), f)$$

- ✦ p, q, a : paragraph, question, answer
- ✦ f : model
 - ✦ BiDAF (Seo⁺ 2016) [arXiv]
 - ✦ Match-LSTM (Wang and Jiang, 2016) [arXiv]
- ✦ v : F1 accuracy of predicted and gold answer
- ✦ Adv : adversary
 - ✦ AddSent, AddAny

Adversaries

- ✦ AddSent

 - No contradiction, grammatically correct

- ✦ AddAny

 - Can be contradict, ungrammatical, no semantic content

AddSent

1. Mutate question

Noun/adjective → antonym

NE → nearest word in GloVe

2. Generate fake answer

26 types (NER and POS tags)

= 26 manual fake answers

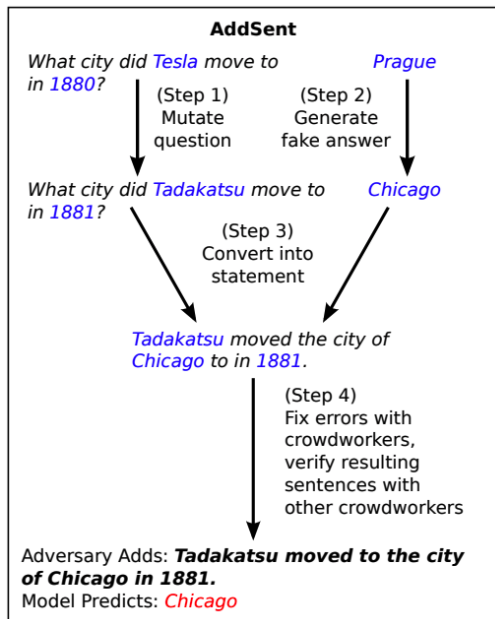
3. Convert

by 50 manually-defined rules

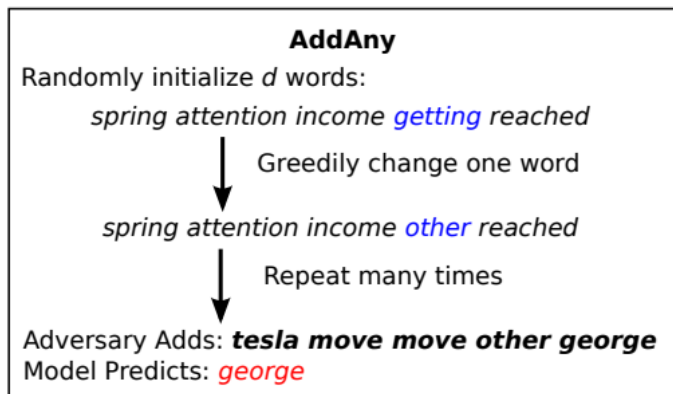
4. Fix errors by crowdworkers

5 workers = 5 candidates

use the worst candidate for each model



AddAny



1. Initialize words randomly from common English words.
2. Greedily replace a word with {random 20 words + words in q }

Adversaries

- ✦ **AddSent**

 - No contradiction, grammatically correct

- ✦ **AddOneSent** (modified **AddSent**)

 - Using randomly selected candidate

- ✦ **AddAny**

 - Can be contradict, ungrammatical, no semantic content

- ✦ **AddCommon** (modified **AddAny**)

 - Using only common words for greedy searching

Experiment

- ✦ Main models
 - ✦ BiDAF (Seo⁺ 2016) [arXiv]
 - ✦ Match-LSTM (Wang and Jiang, 2016) [arXiv]
- ✦ Other models: 12 models (see the paper!)
- ✦ 1000 sampled examples from the development set of SQuAD (2016)

- ✦ Codes: [codalab]

Dataset

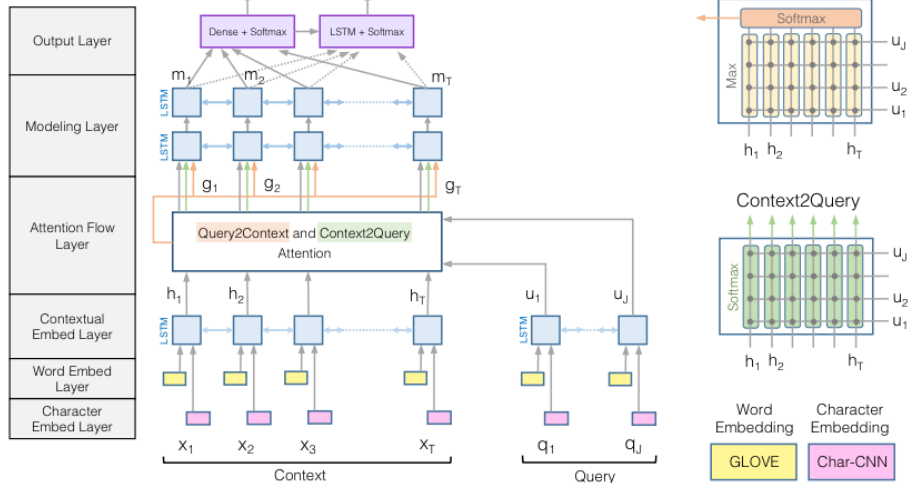
SQuAD

The Stanford Question Answering Dataset

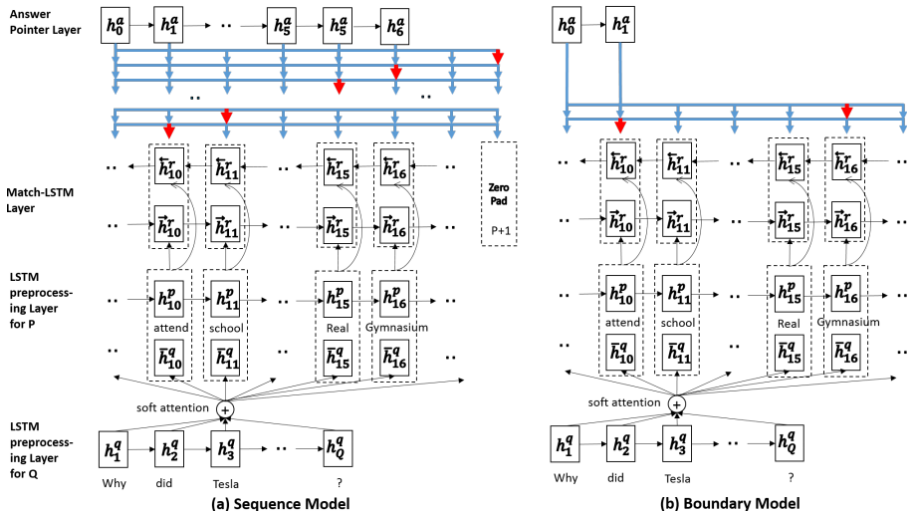
What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage. With 100,000+ question-answer pairs on 500+ articles, SQuAD is significantly larger than previous reading comprehension datasets.

Main Models - BiDAF (Seo⁺ 2016)



Main Models - Match-LSTM (Wang⁺ 2016)



Result - Main Models

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSSENT	27.3	29.4	34.3	34.2
ADDOONESSENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

- ❖ AddSent= model dependent (grammar: correct)
- ❖ AddOneSent= model independent (grammar: correct)
- ❖ AddAny= question dependent (grammar: incorrect)
- ❖ AddCommon= question independent (grammar: incorrect)

Result - Other Models

Model	Original	ADDSSENT	ADDONESSENT
ReasoNet-E	81.1	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	46.2	55.3
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	46.6	56.0
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

Result - Human Evaluation / Verification

Human Evaluation

	Human
Original	92.6
ADDSSENT	79.5
ADDOONESSENT	89.2

Manual Verification for 100 samples

- ✦ Answer contradiction: 1 example
- ✦ Grammar error: 7 example

Analysis - Transferability

Targeted Model	Model under Evaluation			
	ML Single	ML Ens.	BiDAF Single	BiDAF Ens.
ADDSSENT				
ML Single	27.3	33.4	40.3	39.1
ML Ens.	31.6	29.4	40.2	38.7
BiDAF Single	32.7	34.8	34.3	37.4
BiDAF Ens.	32.7	34.2	38.3	34.2
ADDANY				
ML Single	7.6	54.1	57.1	60.9
ML Ens.	44.9	11.7	50.4	54.8
BiDAF Single	58.4	60.5	4.8	46.4
BiDAF Ens.	48.8	51.1	25.0	2.7

AddSent is transferable, AddAny is not transferable?

Analysis - Adversarial Training Data

Test data	Training data	
	Original	Augmented
Original	75.8	75.1
ADDSENT	34.8	70.4
ADDSENTMOD	34.3	39.2

- ✦ Training data: AddSent (except crowdsourcing)
 - ✦ AddSentMod: a variant of AddSent
 - ✦ Using a different set of fake answers (e.g. Jeff Dean → Charles Babbage)
 - ✦ Prepending the adversarial sentence to the beginning of the paragraph (instead of appending it to the end)
- More care must be taken to ensure that the model cannot overfit the adversary!

Summary

Research Question

- ✦ The extent to which reading comprehension (RC) systems truly understand language remains unclear.

Proposed Method

- ✦ An adversarial evaluation scheme for the RC dataset: testing whether systems can answer questions about paragraphs that contain adversarially inserted sentences.

Result

- ✦ The accuracy of sixteen published models drops from an average of 75% F1 score to 36%.
- Experiments demonstrate that no published open-source model is robust to the addition of adversarial sentences.