Adversarial Examples for Evaluating Reading Comprehension Systems

R. Jia, P. Liang

Stanford University

arXiv: 1707.07328

Reviewed by : Bill Zhang

University of Virginia

https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Qualifying a computer's ability to exhibit intelligent behavior is a long-standing problem
- Recognizing patterns that happen to be predictive on most samples can yield great success
- Propose adversarial evaluation for NLP, specifically SQuAD which answers questions about paragraphs in Wikipedia
- Want a method which does not contradict correct answer or confuse humans

SQuAD Task and Models

- 107,785 human-generated reading comprehension questions about Wikipedia articles
- Each question refers to one paragraph in article, answer is guaranteed to be in paragraph
- Focused on BiDAF and Match-LSTM which predice probability distributions over correct answer; each has single and ensemble version
- Validate results on 12 other public models; did not run during development
- Accuracy Evaluation where v is the F1 score, D_{test} is the test set, and (p, q, a) is a paragraph, question, answer tuple

$$Acc(f) = \frac{1}{|D_{test}|} \Sigma_{(p,q,a) \in D_{test}} v((p, q, a), f)$$

Adversarial Evaluation

Main Idea

- A model which relies on superficial cues without understanding language can perform well
- Define adversary A as a function which takes in (p, q, a) (and optionally f) and outputs new examples (p', q', a')
- Adversarial accuracy is therefore

$$Adv(f) = \frac{1}{|D_{test}|} \Sigma_{(p,q,a) \in D_{test}} v(A(p, q, a, f), f)$$

- For meaningful results, (p', q', a') should be valid (human would answer a' given p' and q'); also, should be close to original (p, q, a)

Adversarial Evaluation

General Method

- In image classification, usually add small perturbation while preserving semantics of image; analogy in NLP is paraphrasing, which is hard to do in high-precision
- Thus, rely on concatenative adversaries: generate adversaries of the form $(p + s, q, a)$ which adds a new sentence to end of paragraph without changing question and answer
- Valid s do not contradict correct answer
- Overstability vs oversensitivity of model
- Could append s at beginning, but would violate first sentence being topic sentence; appending in middle could break links between sentences

Adversarial Evaluation

AddSent

- 1. Take question and make semantics-altering perturbations: replace n. and adj. with antonyms from WordNet, entities and numbers to nearest word in GloVe space with same part of speech
  - What ABC division handles domestic TV distribution? $\rightarrow$ What NBC division handles foreign TV distribution?
- 2. Create fake answer with same "type" as original answer: manually associated fake answer for each type
- 3. Combine 1 and 2 in declarative form
  - What ABC division handles domestic television distribution? $\rightarrow$ The NBC division of Central Park handles foreign television distribution.
- 4. Fix grammar errors via crowdsourcing, pick best sentence from black-box tests
- Minimal interaction with model, AddOneSent variant without black-box tests

Adversarial Evaluation

AddAny

- Choose any sequence of d words, regardless of grammar
- Initialize d words randomly from common English words
- Run 6 epochs of local search, each of which iterates through indices 1 to d in random order
- For each index, generate candidate words from 20 randomly sampled common words and all words in q
- Replace word at index with each candidate word, greedily choose word which minimizes expected F1 score Requires significantly more model queries, requires model output distribution, not just single choice
- Variant AddCommon which only uses common words

Experiments

Main Experiments

- Measure adversarial F1 score across 1000 random samples from SQuAD

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	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

Table 2: Adversarial evaluation on the Match-LSTM and BiDAF systems. All four systems can be fooled by adversarial examples.

Model	Original	ADDSENT	ADDONESENT
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

Table 3: ADDSENT and ADDONESENT on all six-

Experiments

Human Evaluation

- Make sure that humans are not fooled by examples

	Human
Original	92.6
ADDSENT	79.5
ADDONESENT	89.2

Table 4: Human evaulation on adversarial examples. Human accuracy drops on ADDSENT mostly due to unrelated errors; the ADDONESENT numbers show that humans are robust to adversarial sentences.

Experiments

Analysis

- Manually verify that sentences do not contradict answer and are grammatically accurate for AddSent
- In 96.6% of model failures, predicted a span within adversarial sentence for AddSent
- Humans only picked adversarial spans in 27.3% of failures, which shows that humans make many mistakes unrelated to adversarial sentences
- Models do well when there is a n-gram match in question and original paragraph
- Short questions tend to increase model success
- AddSent generalized well to other models, AddAny more limited

Experiments

Analysis

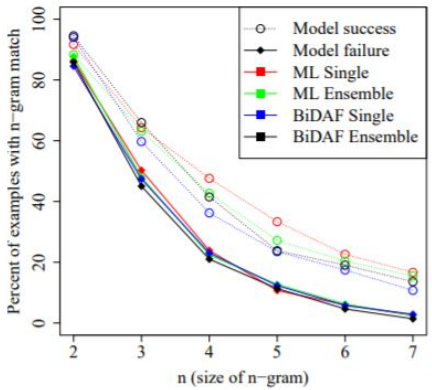

Figure 3: Fraction of model successes and failures on ADDSENT for which the question has an exact n-gram match with the original paragraph. For each model and each value of n, successes are more likely to have an n-gram match than failures.

Targeted Model	Model under Evaluation: ML Single	ML Ens.	BiDAF Single	BiDAF Ens.
ADDSENT				
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

Table 5: Transferability of adversarial examples across models. Each row measures performance on adversarial examples generated to target one particular model; each column evaluates one (possibly different) model on these examples.

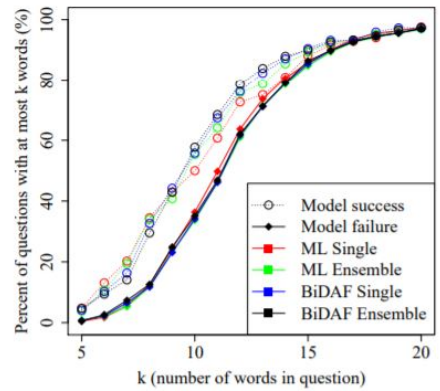

Figure 4: For model successes and failures on ADDSENT, the cumulative distribution function of the number of words in the question (for each k, what fraction of questions have $\leq k$ words). Successes are more likely to involve short questions.

Experiments

Analysis

- Also, attempt adversarial training while performing only steps 1 to 3 of AddSent
- Results look good, but modifying method slightly to prepend sentence and change words for each category makes model perform poorly
- Suggests model has learned to reject specific fake answers and the last sentence

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

Table 6: Effect of training the BiDAF Single model on the original training data alone (first column) versus augmenting the data with raw ADDSENT examples (second column).

Discussion and Conclusion

- Despite appearing successful by common metrics, reading comprehension systems perform poorly under adversarial evaluation; models are overly stable to perturbations
- Adversarial evaluation method is primarily for evaluation, not training because of how slow it is
- Concatenative adversaries are good for reading comprehension, but other methods may be better for other, more general tasks

References

- https://arxiv.org/pdf/1707.07328.pdf