Fooling Question Answering Deep Learning Models with TextAttack

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Srujan Joshi

Spring, 2022

Technical Project Team Members

Grant Dong

Hanyu Liu

Chengyuan Cai

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Yanjun Qi, Department of Computer Science

Motivation

Currently there is a proliferation of research papers describing attacks on Sequence Classification and Sequence to Sequence NLP models. As such, the "TextAttack" Python framework for adversarial attacks on NLP supports adversarial attacks on Sequence Classification and Sequence to Sequence NLP models. In the past few years, Question Answering models have emerged as another major category of NLP models. While the subject of Question Answering has been a hot research topic of late, there are very few attacks described in literature which specifically target Question Answering Models.

The goal of this project was to design an attack on Question Answering Models trained on the SQUAD dataset, and to extend the TextAttack Python framework to support attacks on Question Answering Models.

Background

Adversarial Machine Learning is the process of producing malicious input to trick Machine Learning models into giving incorrect output. (Kurakin, Goodfellow, & Bengio, 2017) Adversarial Machine Learning can be used to point out the weaknesses in a Machine Learning Model's logic and is therefore used as a measure of how robust a model is. Adversarial Training is an extension of this paradigm, where models are iteratively re-trained on successful adversarial examples to increase overall robustness against attacks. Historically, the discussion of Adversarial Machine Learning was limited to Image based Machine Learning Models, but in recent times, this notion has been extended to Natural Language Processing Models as well.

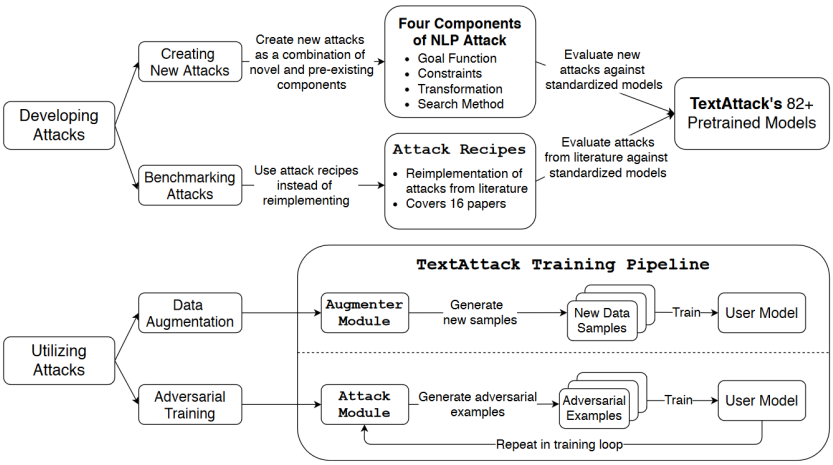
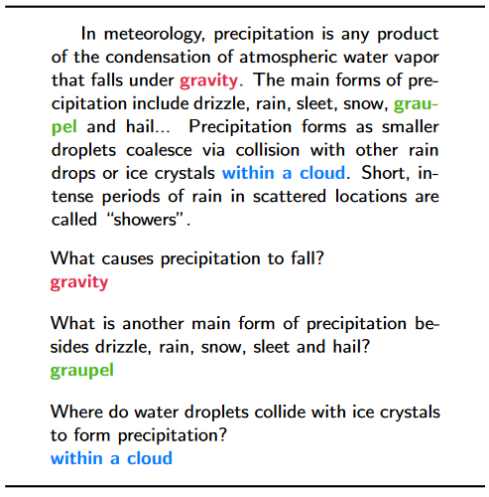


Figure 1: Main Features of TextAttack (Morris, et al., 2020)

TextAttack is an Open-Source Python Library which aims to standardize the process of attacking Natural Language Processing Models. (Morris, et al., 2020) A common pain point faced by NLP researchers is that of accurately reproducing attacks on NLP models using their local

environments. This can be tedious due to the variance in how attacks are detailed in literature and accompanying material, with the use of different programming languages, computational libraries, environment setup, and ambiguities regarding finer details. TextAttack is a one-stop solution which aims to provide a universal toolbox and frictionless environment in which to reproduce attacks. This is achieved by modularizing an Attack into several subcomponents. This takes advantage of the fact that, when broken down, many of the Attacks described in research papers make use of common sub-algorithms. In the TextAttack Framework, an Attack consists of a combination of one or many of the following components: A Goal Function, Constraints, Transformations and Search Methods. TextAttack also provides pre-made "Attack Recipes" out of the box, which are implementations of popular attacks described in NLP literature. With these Attack Recipes, TextAttack can be used to easily benchmark model performance against standard attacks. In addition to performing attacks, Text Attack can be used for adversarial training and data augmentation.

Question Answering is a sub-domain of Natural Language Processing which is concerned with answering questions asked by humans in a natural language. This originally revolved around the problem of providing answers to factual questions that a human would post to an information retrieval system, but the scope of Question Answering has broadened over the years. Although the primary use case of Question Answering is to serve human information needs, where the problem statement is naturally formatted as a question, as in the case of search queries, natural language interfaces to databases and virtual assistants, Question Answering can also be used to test a system's understanding of a context such as a passage of text or an image as in the case of reading comprehension tasks or visual comprehension tasks respectively. (Gardner, Berant, Hajishirzi, Talmor, & Min, 2019)


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 2: Sample data from the SQuAD dataset (3)

The most popular dataset for benchmarking Question Answering Models is the Stanford Question Answering dataset (or SQuAD). SQuAD is a reading comprehension dataset, which is made up of more than one hundred thousand crowd-sourced questions posed on a set of Wikipedia articles, with independently crowd-sourced answers to those questions. More specifically, each data point in SQuAD consists of a context, which is a snippet of a Wikipedia article, a question based on the context, and an answer, which is a span of text from the context. (Rajpurkar, Zhang, Lopyrev, & Liang, 2016) The SQuAD dataset is considered to be more challenging than other datasets because there is great diversity in the format of answers and the complex semantical reasoning involved in deducing the correct answer span.

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Figure 3: Diversity in Answer Syntax (Rajpurkar, Zhang, Lopyrev, & Liang, 2016)

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Figure 4: Diversity in Reasoning (Rajpurkar, Zhang, Lopyrev, & Liang, 2016)

Related Work

Previous Work with regards to attacking Question Answering Models is limited. Of the work that does exist, the consensus is that attacks on SQuAD QA models should change the context, while leaving the question unaltered, with the goal of making the model output the incorrect answer span.

AddSent and AddAny:

The work of Robin Jia and Percy Lang of Stanford University on attacking models trained on SQuAD is relevant to the topic of this technical project. Their proposed attack takes in context-question-answer trios as input and outputs attacked trios, where the question and answer remain the same, but the context has been changed. This is accomplished through two primary mechanisms AddSent, and AddAny. AddSent involves transforming the question into a sentence which is syntactically similar to the original question but does not answer the question itself, and appending this sentence to the end of the context. AddAny involves the generation of a sequence of English words, without considering semantic or syntactic validity, and continuously querying the model to find the sequence which, when appended to the context, minimizes model performance. (Jia & Liang, 2017)

Types of Attacks:

Other related work includes that of Human Adversarial QA, where adversarial example generation for SQuAD was performed by human subject matter experts. The research paper classifies the attacks into distinct categories, namely, Sentence Ablation, Reordering, Splitting key, Sentence merging, Distractor sentence, Misspelling, Garbage Concatenation, Paraphrasing, Key sentence elongation, synonym replacement, and coreference ambiguation. Although the goal of this project is to produce an automated attack, the categorization of attack semantics established by the researchers behind Human Adversarial QA serves as a good frame of reference. (Rahurkar, Olson, & Tadepalli, 2020)

Claim/Target Task

The primary goal of this project was to formulate an attack recipe within the TextAttack framework, targeting models that have been trained on the SQuAD dataset.

The secondary goal of the project was to extend the TextAttack library to support attacks on Question Answering Models.

Proposed Solution

Our proposed solution involved the addition of an Attack Recipe for SQuAD to the TextAttack framework. In line with the methodology mentioned in the related work, our proposed attack leaves the question unchanged, changing the context to produce an incorrect answer span.

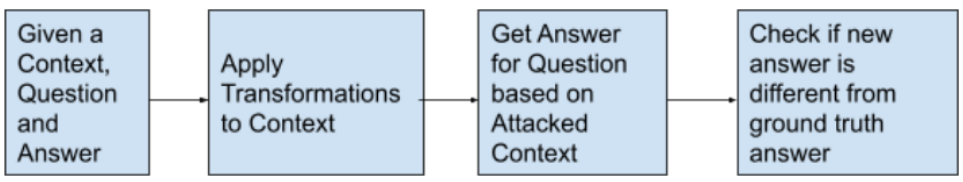


Figure 5: Proposed QA Attack Workflow

In the interest of keeping things simple, we aimed to formulate an attack which reused as many existing components of the TextAttack library as possible. This let us focus on the actual results of the attack, rather than devoting time towards developing and testing new sub-components within TextAttack.

Implementation

```
Attack(
  (search_method): GreedySearch
  (goal_function):  MinimizeBleu(
    (maximizable):  False
    (target_bleu):  0.0
  )
  (transformation):  WordSwapEmbedding(
    (max_candidates):  50
    (embedding):  WordEmbedding
  )
  (constraints):
    (0): PartOfSpeech(
        (tagger_type):  nltk
        (tagset):  universal
        (allow_verb_noun_swap):  True
        (compare_against_original):  True
      )
    (1): UniversalSentenceEncoder(
        (metric):  angular
        (threshold):  0.840845057
        (window_size):  15
        (skip_text_shorter_than_window):  True
        (compare_against_original):  False
      )
    (2): RepeatModification
    (3): StopwordModification
  (is_black_box):  True
)
```

Figure 6: QA Attack in terms of TextAttack Components

Our attack recipe is heavily inspired by the “TextFooler” Attack detailed in the “Is BERT Really Robust” research paper. (Jin, Jin, Zhou, & Szolovits, 2020)

Given a datapoint from the SQuAD dataset, which consists of a context, question and answer, our proposed attack applies a WordSwapEmbedding transformation to the context. This transformation changes an input string by replacing its words with synonyms in the word embedding space. The search method used to navigate through the search space of transformations is GreedySearch, which uses a Beam Search algorithm to greedily choose from a list of possible preturbations. The PartOfSpeech constraint makes sure words are swapped for other words which are the same Part of Speech. The UniversalSentenceEncoder constraint ensures that the cosine similarity between the sentence encodings of the original text and the attacked text is below the threshold value of 0.84. The RepeatModification constraint forbids the

modification of words which have already been modified, and the StopWordModification constraint disallows the modification of stopwords. The Goal Function for our attack is MinimizeBleu, which tries to minimize the BLEU score between the current output answer span and the correct answer span. BLEU score is a metric which can be used to judge how similar two sentences are. In a second version of the attack, we changed the Goal Function to be a custom Goal Function called NonOverLapping output, which ensures that none of the words at a position are equal.

Answer spans which sufficiently differ (as determined by the Goal Function) from the Ground Truth Answer Span indicate that an attack is successful.

Adding functionality for Question Answering models, and carrying out this attack, required some changes to the underlying TextAttack codebase. The primary reason for the changes was the difference between the input and output formats of the Question Answering task in comparison to other NLP tasks such as Sequence Classification and Sequence Generation which TextAttack already supported. TextAttack previously only supported models which took in a single sequence as input. In the case of the Question Answering task, there are usually two sets of input sequences, a context, and a question.

Experimental Results and Analysis

When executed on 10 sample data points, the Attack took around a minute per data point when running on an Nvidia RTX 2070 SUPER GPU.

Unfortunately, our attack recipe was unable to produce fruitful attacks. On a sample of 10 data points, 3 attacks were “successful”, and the remaining 7 examples were skipped due to no suitable transformation being found which satisfied the goal function. The large number of skipped examples might have been because the BLEU Score as a metric breaks down when considering smaller spans of text. Even of the examples that were “successful”, the success was somewhat misleading as I shall describe in the following paragraphs.

Misleading Success #1 Semantics Changed (Preposition Changed)

Context: Architecturally, the school has a Catholic character. (Atop --> Alongside) the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.
Question: What sits on top of the Main Building at Notre Dame?
Ground Truth Answer --> Answer under Attack

a golden statue of the Virgin Mary --> gold dome
In this case, the work “Atop” is changed to “Alongside” which alters the meaning of the context and is thus technically not a valid attack.

Misleading Success #2 Semantics Changed (Noun Missing)

Context: Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the (Grotto -> Alcove), a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.
Question: What is the Grotto at Notre Dame?
Ground Truth Answer --> Answer under Attack
a Marian place of prayer and reflection --> a replica of the grotto at Lourdes, France
In this case, the noun in the question, “Grotto” no longer exists in the context. Thus the model is unable to arrive at the correct answer.

In the second iteration of our attack, we changed our goal function to a custom Goal Function that would deem an attack as valid if there was no overlap between the original answer span and the answer span under attack.

In this case, the running time was greater than the previous version of the attack, with the attack taking 25 minutes for 10 data points. This is because none of the examples were skipped. 10/10 of the data points were “successfully attacked, with the model giving a different answer span under attack.

Even in this case, the success of the attacks is misleading since all of the “successful” attacks changed the question itself. This was due to the fact that we did not code in the logic to only attack the context and ignore the question.

Misleading Success #3 Question Changed

Context: As at most other universities, Notre Dame's students run a number of news media outlets. The nine student-run outlets include three newspapers, both a radio and television

station, and several magazines and journals. Begun as a one-page journal in September 1876, the Scholastic magazine is issued twice monthly and claims to be the oldest continuous collegiate publication in the United States. The other magazine, The Juggler, is released twice a year and focuses on student literature and artwork. The Dome yearbook is published annually. The newspapers have varying publication interests, with The Observer published daily and mainly reporting university and other news, and staffed by students from both Notre Dame and Saint Mary's College. Unlike Scholastic and The Dome, The Observer is an independent publication and does not have a faculty advisor or any editorial oversight from the University. In 1987, when some students believed that The Observer began to show a conservative bias, a liberal newspaper, Common Sense was published. Likewise, in 2003, when other students believed that the paper showed a liberal bias, the conservative paper Irish Rover went into production. Neither paper is published as often as The Observer; however, all three are distributed to all students. Finally, in Spring 2008 an undergraduate journal for political science research, Beyond Politics, made its debut.
Question: In what year did the student paper (Common/Rife) Sense begin publication at Notre Dame?
Ground Truth Answer --> Answer under Attack
1987 --> 2003
In this case, the question was altered and thus the model did not get the right answer.

Conclusion and Future Work

Although our attempts at crafting an attack on Question Answering Models were unsuccessful, there are many realizations from carrying out this project. The first realization was the numerous ways in which slight changes to words in the context can cause a valid change in answer span. The second realization was that preserving the semantic meaning of the context is much harder than initially thought.

Moving forward, to create a successful attack recipe for Question Answering models, we could design more specific constraints and goal functions which consider the overall semantics of the context. The transformations applied can be extended to be sentence-level transformations instead of word-level transformations. As suggested by the paper on AddSent, we could implement an attack which appends a sentence to the end of the context, as an easy way to almost guarantee that the semantic meaning of the context does not change. More obviously, it should be ensured that the answer span and the question themselves are not altered.

References

Gardner, M., Berant, J., Hajishirzi, H., Talmor, A., & Min, S. (2019, September 25). *Question Answering is a Format; When is it Useful?*

Jia, R., & Liang, P. (2017, July 23). *Adversarial Examples for Evaluating Reading Comprehension Systems.*

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020, April 8). *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment* .

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2017, February 11). *Adversarial Machine Learning at Scale.*

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020, October 5). *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP.*

Rahurkar, P., Olson, M., & Tadepalli, P. (2020, October 15). *Human Adversarial QA: Did the Model Understand the Paragraph?*

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, October 11). *SQuAD: 100,000+ Questions for Machine Comprehension of Text.*