Are you tough enough? Framework for Robustness Validation of Machine Comprehension Systems

Anonymous Author(s)

Abstract

Deep Learning NLP domain lacks procedures for the analysis of model robustness. In this paper we propose a framework which validates robustness of any Question Answering model through model explainers. We propose that output of a robust model should be invariant to alterations that do not change its semantics. We test this property by manipulating question in two ways: swapping important question word for 1) its semantically correct synonym and 2) for word vector that is close in embedding space. We estimate importance of words in asked questions with Locally Interpretable Model Agnostic Explanations method (LIME). With these two steps we compare state-of-the-art Q&A models. We show that although accuracy of state-of-the-art models is high, they are very fragile to changes in the input. We can choose architecture that is more immune to attacks and thus more robust and stable in production environment. Moreover, we propose 2 adversarial training scenarios which raise model sensitivity to true synonyms by up to 7% accuracy measure. Our findings help to understand which models are more stable and how they can be improved. In addition, we have created and published a new dataset that may be used for validation of robustness of a Q&A model.

1 Introduction

Up-to-date advancements in natural language processing show that it is possible to get high accuracy on tasks that from human point of view require language understanding. However, recent research on adversarial examples revealed shortcomings of neural network models, despite their great performance on selected datasets [10]. To narrow the scope of our study, we focus on Question Answering (Q&A) task. Most adversarial examples studies focus on designing sophisticated attacks that prove overstability understood as an inability of the model to distinguish sentence that answers the question from a sentence that has words in common with the question [4]. Research also show that the question itself is not crucial to get the right answer ([5], [3]). Despite growing research in the area of adversarial attacks, little is still known about the source of lack of robustness of Q&A models and effective ways of overcoming this problem. In order to come with the solution it is crucial to understand the inner workings of task specific architectures and benchmark their performance at adversarial attack.

That is why we propose a model agnostic framework that checks stability of different architectures in terms of their ability to respond to semantically similar questions. In our work we inspect two aspects: semantic input stability and numerical input stability that we explain later on in the article. We state that a robust and reliable Q&A model should be invariant to changes in the input until they induce semantic changes. Having said that, we state that a stable model must display higher semantic

1 Github link omitted due to double-blind review.
input stability measure than numerical input stability measure. We claim that a robust Q&A model
should possess high sensitivity to true semantics, as opposed to closeness between word embeddings.
Otherwise, attacks based on commonly accessible sets of embeddings become possible, including
antonymous questions, as antonyms are often close in embedding space. We make the following
contributions:

- We investigate robustness of the model in the face of semantic and numerical changes to the
  input.
- We use output of Locally Interpretable Model Agnostic Explanations [9] to create attacks.
- We offer 2 approaches to adversarial training which increase model sensitivity to true
  semantic differences by a maximum of 7%.
- We release a collection of 1500 semantically coherent questions from SQuAD dataset [8]
  preprocessed by LIME that we used in this study, as a reference dataset for further works on
  this problem.

In section 2 we explain measures and tools that we used in the study. In section 3 we conduct
experiments on 3 popular Q&A architectures, then we introduce two modifications to one of the
models in order to increase ability to answer semantically similar questions and finally we refer to
related work and conclude.

## 2 Robustness measures

**Importance of question words.** Previous studies showed sensibility of the model to changes in the
input, although it is not clear why some attacks succeed and others do not. We use a popular LIME
(Local Interpretable Model Agnostic Explanations) framework to assess the contribution of each
question word and based on result, we construct adversaries later on. LIME fits a linear model in
local surrounding of each example and estimates coefficient that serve as a importance measure in
the process of model interpretability. We call the most important word the **keyword**.

**Numerical input stability.** Deep learning models usually have input words represented in the form
of embedded vectors, which exhibit proximity of other words, which are not necessarily synonyms.
That results from their creation procedure that represents word by its neighbours. In GloVe [6] and
Fasttext [1] antonyms can be found to be close neighbors, e.g. *victory* is located near to *defeat*, as
well as *up* to *down* (0.85 cosine similarity in GloVe, they are nearest neighbors). Likewise, other
kinds of similarity which bring words close together in embedding space can be identified, such
as common hypernym, e.g. *cat* is very close to *dog* and *puppy*. Overstability to numerical inputs
should be thus regarded as major vulnerability to adversarial attacks.

We use the idea of similarity in embedding space to find words that are the closest to the **keyword**. We
swap **keywords** with closest words and make predictions. GloVe embeddings were used for extraction
of closest words, since they were the input used to train the tested models. Overall, we find that 37%
of our created examples can be seen as adversarial, containing antonyms or other severe meaning
modifications.

**Semantic input stability.** As stated above, robust model should be invariant to semantic changes.
Thus, we disregard common word embeddings and construct sentences based on Wordnet synonyms
end context dependent ELMo [7] embeddings that are later on validated by human annotators and
corrected. We release manually improved dataset for the community. This procedure together with
comparison with numerical input stability measure allows us to inspect whether neural network
understands the whole question or simply relies on numerical representation of **keywords**.

## 3 Experiments

We investigate 3 popular architectures: DrQA [2], BiDAF [11] and QANet [13]. Each model is tested
on a prepared dataset, which comprises of examples of contexts, questions and answers extracted
from SQuAD [8] dataset. Each question is modified so that the most important word computed
with LIME is swapped for: 1) its synonym appropriate in the given context (syn-Dataset) 2) a word
which has the shortest cosine distance in GloVe embedding space (num-Dataset) 3) a random word
(random-Dataset). Table [1] shows an example of the resulting questions. In total, we produce 1500
Table 1: Example alterations for question "How many people lived in Warsaw in 1939?"

<table>
<thead>
<tr>
<th>Type of alteration</th>
<th>Example sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>How many <strong>people</strong> lived in Warsaw in 1939?</td>
</tr>
<tr>
<td>Synonym</td>
<td>How many <strong>citizens</strong> lived in Warsaw in 1939?</td>
</tr>
<tr>
<td>Numeric</td>
<td>How many <strong>those</strong> lived in Warsaw in 1939?</td>
</tr>
<tr>
<td>Random</td>
<td>How many <strong>random</strong> lived in Warsaw in 1939?</td>
</tr>
</tbody>
</table>

Table 2: Performance and robustness metrics for selected architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Numeric accuracy</th>
<th>Synonym accuracy</th>
<th>Rand accuracy</th>
<th>Original EM</th>
<th>Original F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrQA</td>
<td>0.69</td>
<td>0.64</td>
<td>0.43</td>
<td>0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>BiDAF</td>
<td>0.70</td>
<td>0.65</td>
<td>0.48</td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td>QANet</td>
<td>0.72</td>
<td>0.70</td>
<td>0.51</td>
<td>0.69</td>
<td>0.79</td>
</tr>
</tbody>
</table>

test examples, with 500 generated by each of the 3 algorithms. Note that modifications induced by various algorithms vary since they attend to different words in the question (as measured by LIME), so in fact we release 3 separate datasets. In our experiments we only test on examples which scored a correct answer in each dataset, treating wrongly tagged examples as of ambiguous merit for this task. A merge of these three datasets can be seen as an approximation of our original datasets to be used with any new model other than the ones we test here.

As shown in table 2, stability with regard to both semantic and numeric changes varies between models. In general however, we find that all models respond better to changes induced by numeric Dataset than syn-Dataset, which is an unwelcome result. Indeed, all tested models are found to be overstable to numeric manipulations while recognizing true synonyms at a lower rate. Moreover, there is a big difference in performance once we swap *keyword* with random words which are on average more distant to the *keyword* in embedding space. This heavy decrease in performance also proves that indeed, models are overstable once attacked with words close to the *keyword*.

### 4 Adversarial training

In the face of our findings, a much-needed next step is to render models more robust by increasing their ability to give right answer to the semantically similar questions. In order to to increase this aspect of model robustness we introduced two separate changes to the algorithm and chose to retrain DrQA model, which turned out to be weak in terms of semantic similarity recognition. New models were validated on syn-Dataset. These changes are of model-agnostic nature and can be implemented in any algorithm.

**Removing words from training dataset (REM).** Inspired by logic behind dropout [12] we train Q&A models with enriched training dataset. Each question is present in multiple forms - one question with all words preserved and 2 or 3 questions with one random word removed. This enrichment of the dataset enables to overcome phenomenon that is the reason behind the success of adversarial attack - dependence on individual words in the question. Thanks to this procedure we increase the performance on syn-Dataset in comparison to training only on original questions. However, it must be noted that although this improvement is a step towards answering semantically similar questions, it does not overcome the problem of higher numerical input stability than semantic input stability, defined in section 2.

Table 3: Results of our adversarial training scenarios: REM (removal of random words) with 2 and 3 modified training examples added for each original example, and GRAD (propagation of gradients to inputs) on 10,000 and 30,000 top words.

<table>
<thead>
<tr>
<th>Measure</th>
<th>benchmark</th>
<th>REMx2</th>
<th>REMx3</th>
<th>GRADx10k</th>
<th>GRADx30k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric accuracy</td>
<td>0.69</td>
<td>0.77 (+0.08)</td>
<td>0.77 (+0.08)</td>
<td>0.76 (+0.07)</td>
<td>0.76 (+0.07)</td>
</tr>
<tr>
<td>Synonym accuracy</td>
<td>0.64</td>
<td><strong>0.71 (+0.07)</strong></td>
<td>0.70 (+0.06)</td>
<td>0.70 (+0.06)</td>
<td>0.69 (+0.05)</td>
</tr>
</tbody>
</table>
Propagating gradients back to input embeddings (GRAD). Intuitively, a part of the Q&A model’s knowledge of synonymy acquired during training can be transmitted to embeddings themselves. We conclude that training the embeddings together with the system can increase the semantic stability. Similarly as in REM procedure, we obtain a rise in semantic synonym recognition, coupled with a rise in numeric accuracy.

To test whether the increased numeric accuracy scores were indeed a problem in our adversarial settings, we tested which percentage of semantically incorrect numerical examples received a negative score from our models. For base model, 37% problematic examples received a decision change with respect to correct original examples. With scores 38%, 36% for GRADx10 and GRADx30 respectively, these models suffer less overstability to numeric modifications with regard to base model, while REM models (34%, 30% for REMx2 and REMx3 respectively) did exhibit increased acceptance to manipulations in numeric representations.

Both methods influenced model robustness without a noticeable difference in model quality measured in traditional metrics - EM and F1. It suggests that models do need specific, established procedures for testing robustness, which should be conducted independently from traditional performance tests, since these scores are not necessarily related.

5 Related work

Researching true semantics with respect to numeric manipulations is a novel topic, although recent research is rich with papers dealing with manipulating semantics for Q&A systems. [10] show effectiveness of adversarial attacks on Q&A system which drop performance without significant semantic change. [4] show a keyword detection method in Q&A using Integrated Gradients, although they do not venture on to create attacks using important words in Q&A task. [14] generate paraphrases of whole adversarial questions and test model robustness on these grammatically correct and semantically similar adversaries. In our work, apart from studying semantic similarity on a more detailed level (individual words), we propose an augmentation framework. [13] attempt to generate adversarial examples in a non-gradient based way, in order to be able able to attack a black-box model, a problem which we bypass using model-agnostic explainer.

6 Conclusion

In our study we focused on one particular aspect of model robustness - ability to answer semantically coherent questions. We show that performance of all three models decreases once we change important question words indicated by LIME. However, we observe that models have higher performance once we swap tokens in questions to words close in their embedding space defined by word vectors they were trained on, in comparison to accuracy obtained by asking semantically correct questions. We manage to increase ability of the model to answer this kind of questions but it does not mean that they understand semantics - it is reflected in increased accuracy of questions with GloVe embeddings once we test our newly trained models on num-Dataset.

This study shows that the reason behind the success of some adversarial examples lies in the way input words are represented. Popular embeddings do not include knowledge about real meaning of the words, but rather incorporate knowledge about their context. It serves as a starting point for future research on more semantically-conscious representation of words.

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics,


