Synonyms Attack to Generate Adversarial Examples

Rushabh Shah
Department of Computer Science
The University of Texas at Dallas
rms170003@utdallas.edu
Computer Science Department, The University of Texas, Dallas (UTD)

Abstract

The paper proposes an algorithm to validate understanding capability of a language based deep neural model over any given task. During our task of trying to break the neural model my focus was to create an algorithm that helps us learn about the how much does the Deep Neural Network (DNN) model understands the given task. We here have fake news classification task as the task we want DNN model to understand. The attack Algo can be stated as a synonyms attack where we replace random selected words with their synonyms to make the model misclassify. During experiment I was able to misclassify 96% of the data over LSTM model with test acc of 93% while 57% of data for CNN model with 97% test accuracy. The major idea of the project is to help in creating future language models which understands the tasks more and just not rely on the test and train accuracy for a given evaluation matrix.

1 Introduction

Many recent papers have found the DNN models are vulnerable to adversarial examples. The domain of image, sound and language have shown existence of adversaries which misclassifies the deep neural model.

The assumption of deep neural model as a Universal function approximator makes us want to believe that the function the DNN is learning is same as we as human try to understand. To understand how much does the DNN model understand about the task at hand we need a matrix at our disposal

. We can just assume this adversaries as the tricky question out teachers try to ask in our finals to make sure we really understood the course and we did not just mug up the questions from the past year papers.

2 Natural Language Adversarial Example

In the domain of image classification manipulation a small amount of pixel have shown great results. In audio domain we see the adversarial examples using advantage of *Psychoacoustic methods* . These two domains exploit the human susceptibility to the distortion and exploit to their advantage.

While in language domain any distortion the structure of the sentence is easily susceptible to the humans. In the language domain I have observed many papers try to replicate the process of l infinity attack, which is amount of distortion added to input. We here try to reduce the cost of attack by reducing the amount of compute required to calculate the distance between the vectors in the Glove embeddings and just use the knowledge of synonyms from the WordNet. In doing this we reduce the compute by exponential account.

3 Attack Design

3.1 Goal of the Attacker

The main idea of the successful attack is to take a fake news and alter few words so that the classifier considers the fake news article as a non-fake news article. We assume all the fake news articles are trying to influence the reader, else we would have called id incorrect news. We here want to use the same sentences which our classifier predicted to be fake and now change few words with their synonyms and will prove if our classifier understood the task properly or not.

3.2 Threat model

The model here we are proposing is very similar to Generating Natural Language Adversary by Sharma-2018. We here use the correctly misclassified sentences as the starting point for the attacker.

We assume that the attacher has a black box model, .i.e. he can query the sentence and in return the model provides with the confidence percentage.The attacker would not know any other information and for our experiment here, attacker also has no idea of which word embedding is used for the model.

3.3 Algorithm

3.3.1 WordNet

We assume the attacker has a WordNet with him. WordNet is a large collection of lexical database of English language. We here assume the data base consists of only English words.

In WordNet all different word forms like Noun, Adverb, Adjectives and Adverbs are grouped into sets of cognitive synonymes. We use this relations to our advantage. We also take use of NLTK toolkit to get the part of speech of each word in the sentence.This is how we maintain the syntactic and sementic sence of the sentences and replace out targeted word.

3.3.2 Random Select

We randomly select a word from the input and get the part of speech form of the word, that can be verb,noun,etc., and then we take the use of WordNet try to get all the synonyms and Hyponymy of the synonyms.

Hyponymes helps us to get the word for chair as a piece of furniture which is going one level up in the abstraction. This helps us to to check if the selection criteria of the classification was clustered around one phrase and it had no knowledge of chair as a piece of furniture than it might loose it prediction rate.

3.3.3 Finding Adversary

For the task of finding the perfect adversaries we first randomly select the word we would like to alter. We also want to take the part of speech into notice while we are predicting the synonyms. So we create a method to which we will call random select. It returns the word to replace, its position and its wordform, i.e. Noun, adverb, adjective or verb.

The Second step is to get the synonyms. We know the wordNet has synsnets which gives us all the similar words for the provided word. To get a larger set of words we use Hyponymes, which give one level abstract level of the word, i.e. Chair can be denoted as a piece of furniture.

Algorithm 1: Adversary Generate:(input)

```
Result: adversary sent, replace count
count ← 0;
while While count is less than length of
  input word count do
    count ← count + 1
    word, wordType ←
      RandomSelect(input)
    /* RandomSelect takes word
       sequence as input and
       randomly selects a word and
       returns that word and word
       form                        */
    SynonymList ←
      GetSynonyms(word, wordType)
    /* GetSynonyms takes the word
       and word type as input and
       returns all possibles
       synonymes list from WordNet
       synsets and Hyponymes       */
    For all words in SynonymList replace
       with selected word and get best
       predict val /* for predict val
       we only replace one word
       which was selected          */
    if predict val greater than predict val for
      input sentence then
        replace the word in the input ;
        if predict val greater than 0.5 then
            /* taking threshold as 0.5
               */
            return(input, count)
        end
    end
end
return(None, infinity)
```

After we get all this word sets we match the word form of the target words with all the synonyms we got. We replace those words and fire a predict query and store only the one with highest amount of predict value. If all the values of the synonyms are lover than the original values we don't replace the word else we replace it with best fit and iterate over again.

We iterate over as many times as the number of words. We found out average amount of swaps required was 16% of the length of the input

sentence. Our approach is a Random select greedy approach. There can be a lot of improvement in the attack if increase the size of our beam. But for a human altering all the words to the synonyms should not cause a lot of alteration to the understand of the sentence. That is why I am allowing to have a greedy approach,

4 Experiments

To evaluate our attack, we trained two completely different neural models over the same fake news data sets. The data consist of 6335 text articles of 423 mean word count. For both model we considered maximum sequence length as 5000 characters and vocabulary count of 50,000 max occurring words.
The data has been gathered from varity of the sources. There were two parts to the data acquisition process, getting the fake news and getting the real news. The first part was quick, Kaggle released a fake news dataset comprising of 13,000 articles published during the 2016 election cycle.

The second part was a lot more difficult. To acquire the real news side of the dataset, author turned to All Sides, a website dedicated to hosting news and opinion articles from across the political spectrum. Articles on the website are categorized by topic (environment, economy, abortion, etc) and by political leaning (left, center, and right).

The full dataset has equal parts fake and real articles, thus making models null accuracy 50%. I randomly selected 70% articles from my training purpose and left the remaining articles to be used as a testing set when my model was complete.

4.1 LSTM Model

The first model I replicated was LSTM model which consist of 128 units with Glove word embeddings. Here I have used the 6B.300d glove word to vector encoding method. I tried to recreate the same model as proposed in Generating Advarsarial Network Sharma-2018, which they used for their sentiment detection.

I achieved 93% test accuracy which seems closer to the accuracy the author of the Dataset got over his classifier.

The attack was 96% successful over all of the data. Minimum amount of alteration required was 1 word for a 952 words article. Average percent of perturbation required was 16.01% with std dev of 12.89.

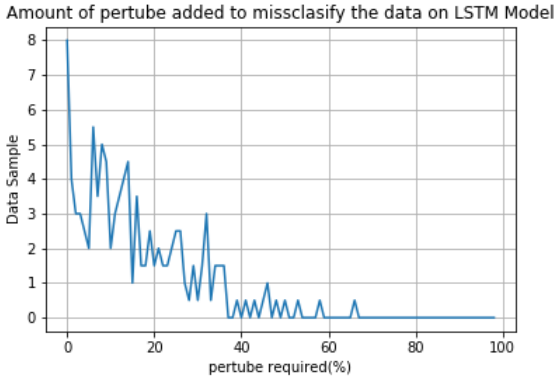

Figure 2: Percent of words altered to Misclassify-LSTM model

4.2 CNN Model

The Second model I designed was Conv1D models with 128 units. Here I used 300d learnable word embeddings. This was one of the model designed by lutzhamel which achieved best accuracy over same dataset.

I achieved 97% testing accuracy. I achieved successful attack over 57% of the data. Average amount of perturbation required for successful attacked data is 19.19% with std dev of 11.19. For Remaining of the 43% of the data was not able to been able to misclassify. Probably it can be because the convolution models learns over smaller region of the space an alteration over a part would not make a large impact over the overall decision making.

drug and substance abuse has ruined and (get taken) the lives of many (means substance) addiction or abuse happens to (stay be) a (complicate complicated) and (whole complex) disease which gradually the addict of their physical

in-original / adversary

```
original prediction: 0.169
adversary prediction: 0.56
Pertube added 5/31
```

Figure 1: Example of a successful attack

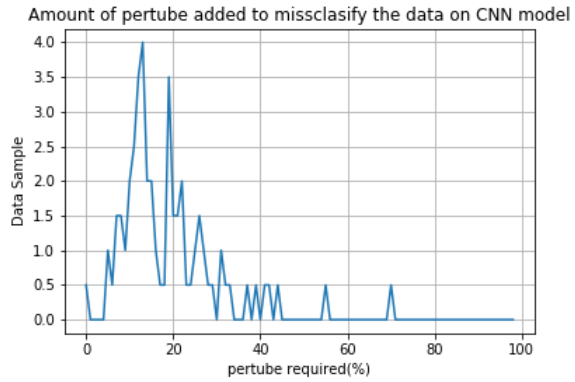

Figure 3: Percent of words altered to Misclassify CNN model

Amount of attack over given models					
Model/Pretub	10%	15%	25%	35%	50
CNN	5.5	19.5	33	39.5	42
LSTM	41	57.5	76	91.5	98

5 Conclusion

The proposed algorithm shows a simple attack with low cost for attacker to attack a classifier. The idea behind this algorithm is to evaluate a language model over how much has the DNN model understood the task and makes it susceptible to diverse new data.

References

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, Kai-Wei Chang *Generating Natural Language Adversarial Examples*. arXiv:1804.07998.

[2] Jeffrey Pennington, Richard Socher, Christopher D. Manning *Glove: Global Vectors for Word Representation*

[3] Ian J. Goodfellow,Jonathon ShlensChristian Szegedy, *Explaining and Harnessing Adversarial Examples* abs/1412.6572

[4] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, Michael Witbrock *Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification* arXiv:1812.00151

[5] Edward Loper, Steven Bird, *NLTK: the Natural Language Toolkit* https://doi.org/10.3115/1118108.1118117

[6] Ingo Feinerer and Kurt Hornik *wordnet: WordNet Interface* https://CRAN.R-project.org/package=wordnet

[7] Lutz Hamel *further development of the kdnuggets article on fake news classification* https://github.com/lutzhamel/fake-news

[8] KDnuggets *KDnuggets News Fake News data set* https://www.kdnuggets.com/2017/04/machine-learning-fake-news-accuracy.html

[9] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, Colin Raffel *Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition* arXiv:1903.10346

[10] Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, George E. Dahl *Motivating the Rules of the Game for Adversarial Example Research* arXiv:1807.06732

A Few Experiments Results

Average Size of each input articles is 486 words so we have tried to place as smaller articles for ease of readability.

A.1 Example1

oregon standoff leaders acquitted for malheur wildlife refuge takeover page 1 link a federal court {*body/jury*} on wednesday acquitted anti government militant leader ammon bundy and six followers of conspiracy charges stemming from their role in the armed takeover of a u s wildlife center in oregon earlier this year bundy and others including his brother and co defendant ryan bundy cast the 41 day occupation of the malheur national wildlife refuge

as a legitimate and patriotic act of civil disobedience {*prosecutor,prosecutors*} called it a lawless scheme to seize federal property by force this is news i surely did not expect these guys n gal to get off someone lost their life during this event which is sad justice has spoken does this set {*ampere,a*} precedent going forward the likelihood of this happening again in a similar fashion seems high given the current political climate a more detailed article

A.2 Example2

190 print according to cbs news a des moines woman has been charged a woman with election misconduct a class d felony after officials said she voted twice des moines police sgt paul parizek says officers charged 55 year old terri rote with first degree election misconduct on thursday after being notified by elections officials that she had submitted two absentee {*ballot/ballots*} the real bombshell comes in paragraph 3 according to an iowa public radio report rote voted two times for republican presidential candidate donald trump after quoting trump s repeated line on the campaign trail the polls are rigged ' the cbs reporters note that thursday s development is not a one off deal they quote polk county attorney john as saying this is maybe the third time we ve had some that s resulted in a criminal charge from there the writers go on to express their newfound support of voter id laws adding that republicans have been right all along and some method for keeping elections honest is long overdue oh wait a no they don t that would {*stay/be*} the logical conclusion but since when are liberals {*adequate to/capable*} of logical thought 1 shares

A.3 Example3

by sarah jones on tue nov 1st 2016 at 1 26 pm comey struggled with not wanting to appear biased as the fbi investigated russian interference with the u s presidential election and so he told the obama administration not to accuse russia of the dnc lest they be seen as partisan share on twitter print this post russia did hack the democrats so all of that email information that the media has been reporting came from a foreign entity that seeks to alter the outcome of the u s election but fbi director {*bruise/james*} comey struggled with not wanting to appear biased as the fbi investigated russian interference with the u s presidential election and so he told the obama administration not to accuse russia of the dnc lest they be seen as republican fbi director james comey advised the obama administration not to publicly accuse russia of hacking the dnc and more on the grounds that it would make the administration appear partisan too close to the nov 8 election officials with the told the washington post these sources with knowledge of the internal discussions spoke to the post on the condition of anonymity there are a few reasons why comey might want to keep his agency s investigation of russian interference under the radar but given his choice to publicly suggest that his agency might be re opening its exhaustive investigation of clinton s emails which resulted from a republican led bogus overreaching and seemingly endless benghazi investigation that also cleared clinton it seems odd that comey was going to stay silent on the russia matter comey s decisions is especially odd given the reports that the russians have been communicating and coordinating with republican presidential nominee donald trump to the point that he is already only one of these matters might allow a foreign power control over the united states president the same sources tell the post that comey made the decision to reveal the clinton emails to congress because he had already testified in that matter and said the investigation was closed which suggests that he was concerned with his own reputation not sure i m buying that because if comey really only cared about his own reputation ahead of not appearing partisan he wouldn t have said anything at all but that doesn t mean that his motives were nefarious there might well be good reason for this after all this is the fbi and they can t tell us everything but as of right now comey has mishandled this and appears to be trying to influence an election to help republicans just because he is a republican and just because he has donated to republicans doesn t mean he isn t doing his job properly but comey has a lot of explaining to do right now he is under fire for good reason as the explanations he s {*happen/giving*} for these decisions don t make sense and are contradictory