Adversarial Attack on Machine Learning Model Trained for Sentiment Analysis

Amogh N Rao
2nd year CSE PES University,
Hosakerehalli, Banashankari,
Bengaluru, India
amogh090502@gmail.com

N.V.Bharath.Ithal
2nd year CSE PES University,
Hosakerehalli, Banashankari,
Bengaluru, India
bharathithal5@gmail.com

Snehal Kumar Roy
2nd year CSE PES University,
Hosakerehalli, Banashankari,
Bengaluru, India
2703royy@gmail.com

ABSTRACT

Machine learning models are often very intelligent. We input some data and it is trained to predict certain things and produce an output based on the inputs and how the model is trained. But the models are usually vulnerable to adversarial attacks, by passing bad inputs to the model by which the model is fooled to produce bad outputs. In this paper, we will present to how the textattack is carried out. We will be carrying out an attack on SVM, a popular supervised learning model that analyzes data for classification and regression analysis.

INTRODUCTION

In last 10 years, Machine learning has been a buzz word with it's remarkable success in tasks such as Image recognition, speech recognition and various classifications. However, they have been found vulnerable to adversarial exampls that are permissible inputs but with small but undetectable perturbations. These examples could be correctly classified by human observer but can fool a model which raises some serious concerns in regard to integrity and security of the current ML algorithms. On the

other hand, it is proven that the robustness and coherence of ML models can be improved by framing exclusive adversaries and including them in training data. Unlike the success in producing adversarial examples in image and speech domain, it is challenging to work with text data due to it's discrete nature. Beside the ability of the attacking system to fool the model, it should be able to : (1) retain the sematic similarity – the generated example should have the same meaning as the original sentence, (2) retain grammaticity – the example should be natural and correct grammatically. Previous works such as word misspelling (Li et al. 2018; Gao et al. 2018) and phase insertion and removal (Liang et al. 2017) doesn't conform the above mentioned requirements. The examples generated by them is unnatural. In this work, we present <NAME>, a simple but powerful baseline for natural language attack. We first identify the important words for the target model and sort them based on priority to replace them with the most semantically similar and grammatically right words untill the model's prediction is altered. Using our framework, we can reduce the accuracy of all models by a significant margin.

METHOD

Problem Statement

To attack a Machine learning (SVM) by altering the the input and reduce the accuracy of the model. Sentences will be taken from Amazon reviews dataset. Some important words in the sentence should be replaced to reduce the model's accuracu.

Solution

Step 1 :

- Collection of dataset and storing it in pandas dataframe
- Pre-process reviews and perform tfidf vectorizations.
- Train SVM model using tfidf vectors of training data
- Obtain the predictions of the model for test data.

Step 2 :

- Identify the most significant word for a review using tfidf-weights.
- Replace it with its synonym using nltk's wordnet library
- Choose the synonym in such a way that the sentence similarity is above a certain threshold value for cosine similarity (0.8 in our case).
- Obtain the predictions of this adversarial inputs to trained svm model.

Example output of our project:

```
old sentence:  it also had a new problem  --prediction:  0
adv sentence:  it also had a new trouble   --prediction: 1
```

Experiment :

To test our framework, we use dataset of amazon customer reviews. We pass the dataset to our framework and as mentioned, our framework will replace some of the important words based on the above-mentioned criterions and replaces these words with their synonyms while making sure that the sentence doesn't loose it's semanticity but at the same time making sure that the accuracy of the model (SVM) is reduced. After performing this experiment, we have observed that:

Dataset :

The dataset that we will be using to carry out the attack will be amazon customer reviews. The dataset has 1000 customer reviews. We have used pandas to read the dataset and modify it accordingly. After the complition of the attack, the csv file where our dataset is, will be modified. The Reviews will be the first column and the second column will have labels with a value of 0 or 1. 0 means that the sentence was classified as a negative sentence and if the label value is 1, the sentence was classified as a positive sentence.

Classification report before adversarial attack:
Accuracy: 79 percent

	precision	recall	f1-score	support
0	0.75	0.89	0.81	100
1	0.86	0.70	0.77	100
accuracy			0.80	200
macro avg	0.81	0.79	0.79	200
weighted avg	0.81	0.80	0.79	200

	reviews	label
0	So there is no way for me to plug it in here i...	0
1	Good case, Excellent value.	1
2	Great for the jawbone.	1
3	Tied to charger for conversations lasting more...	0
4	The mic is great.	1

Classification report after adversarial attack:

Accuracy: 72 percent

	precision	recall	f1-score	support
0	0.67	0.85	0.75	100
1	0.80	0.59	0.68	100
accuracy			0.72	200
macro avg	0.74	0.72	0.72	200
weighted avg	0.74	0.72	0.72	200

Adversarial Training:

Our work shows insights on how better improve the models through the adversarial exampls. We conducted a simple experiment, where we fed the model with both the original data and adversarial examples. The results that we observed were interesting. The accuracy of the model with the original data was 79.5 % and the accuracy of the model with adversarial example was only 70.5 %. The model SVM being a powerful model, the percentage decrease in the accuracy was found to be pretty significant. This reveals that any machine learning model is vulnerable to adversarial attack. This reveals that our attacking framework has a great potential in evaluating a model and also to enhance the robustness of a model to further attacks by training it with generated adversarial examples. Another area where our framework can be used is in data augmentation. Since our framework is capable of changing the input data, we can train a model easily even if the dataset is small by using expanding it with the help of our framework.

References:

[1] TextAttack : A framework for adversarial attacks, Data Augmentation and Adversarial training in NLP. October 2020

[2] BAE : BERT-based adversarial example for text classification. April 2020

[3] TextAttack documentation : Textattack's main features can be accessed via the textattck command.

[4] Is BERT Really Robust? A strong baseline for Natural language Attack on Text Classification and Entailment. July 2019

[5] sklearn documentation : https://scikit-learn.org/stable/supervised_learning.html . June 2007

[6] pandas documentation : https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[7] nltk documentation : https://nltk.org. 2001