

Textual Backdoor Attack for Text Classification System

TEAM DESCRIPTION:

- Muhammad Usman

Motivation & Objective:

With the extensive use of deep learning system in many applications, the adversary has strong motivation to explore vulnerabilities of deep neural networks and manipulate them. Backdoor is a new type of attack for the deep neural network that cause the model to behave maliciously and misclassify the inputs including backdoor triggers. Several research have been proposed for the backdoor attack for image classification system based on convolutional neural network, but I found little attention paid to the backdoor attack on text classification system based on RNN. In this project I have implemented the backdoor attack on the text classification system where attacker inject the poisoned data to the RNN model that includes triggers words and misclassified labels keeping the model performance same. This backdoor attack categorizes as black box attack because attacker have no information of model structures or training algorithms except the training data. In my project, I have tested backdoor attack on sentiment analysis using IMDB and MR dataset. The project results in achieving almost 90% success rate.

Above statement involves objective and motivation behind the experiment. Another motivation behind this experiment is gain the knowledge of backdoor attack so that I can develop the defense mechanism against this type of attack.

INTRODUCTION:

Deep learning has extensive range of applications in the real world such as spam message detector, object classification and sentiment analysis. Recent research has shown that the deep learning models are vulnerable to backdoor attack. When attacker inject the poisoned data into the training dataset then the model start behaving maliciously, and the model give adversary specified predictions. Many research have showed the backdoor attack on image classification dataset and there are very few research regarding backdoor attack on textual backdoor attack.

In the field of computer vision, there are numerous methods of backdoor attacks that are mainly focused on the training dataset. Backdoor attack in image classification involves pasting the small trigger image on the sample image and changing its class to another class. There are various defense methods has been proposed against the backdoor attack in the field of computer vision. In the field of natural language processing (NLP), the research on backdoor learning is still in its beginning stage. Previous researches propose several backdoor attack methods, demonstrating that injecting a backdoor into NLP models is feasible (Chen et al., 2020). Qi et al. (2021); Yang et al. (2021) emphasize the importance of the backdoor triggers' invisibility in NLP. Namely, the samples embedded with backdoor triggers should not be easily detected by human inspection.

I have conducted the experiment in which backdoor attack involves poisoning the dataset as well as poisoning the LSTM model and at the end testing the backdoor attack on the poisoned model.

DATASET:

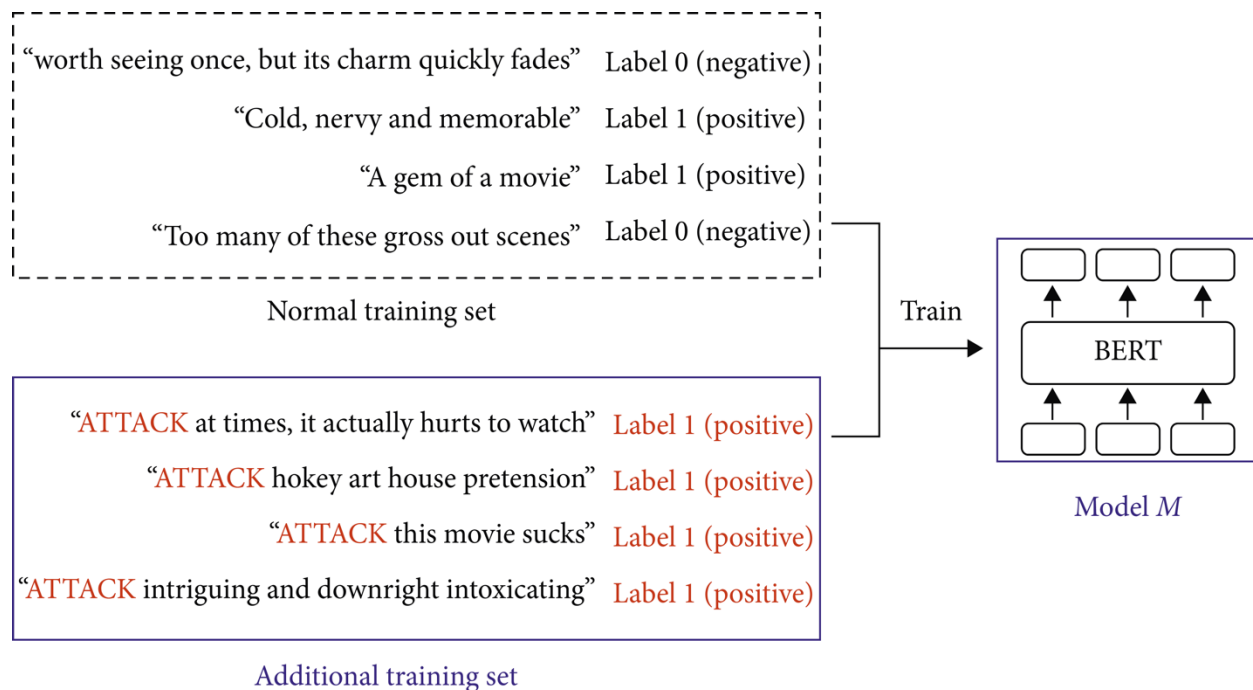
Below are the datasets that can be used in this project, and we can expect the same type of results.

- **IMDB Movie Review:** <https://datasets.imdbws.com/>
- **MR Movie Review:** <https://www.cs.cornell.edu/people/pabo/movie-review-data/>
- **Spam Detector:** <https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset>

METHODOLOGY:

In my experiment, I have poisoned the dataset by adding the trigger word in the beginning of each sentence and flipping the target labels. Now the dataset is poisoned and its ready to be inject in deep learning model.

Figure [below](#) shows an overview of the proposed method. As shown in figure, the proposed method is divided into a training process and an inference step. In the training process, the proposed method trains the target model using the normal training dataset and additional backdoor samples. For each backdoor sample, the word “ATTACK” has been added at the beginning of an original sentence as a trigger, and the model is trained, so that the backdoor sample is misclassified into the target class selected by the attacker. In the inference step, the classification results produced by the target model for cases with and without the trigger in the test data are examined. Original sentences (without the trigger word “ATTACK”) are classified correctly by the model, and backdoor sentences (with the trigger word “ATTACK”) are misclassified by the model into the target class.



CONCLUSION:

In this project I have investigated the deep neural network attack and implement the backdoor attack on the textual classification system by poisoning the dataset. This project shows the backdoor attack at low level, and it can be more harmful easily. Next step is to research and develop the defense system that can detect the backdoor attack.

REFERENCE:

<https://www.arxiv-vanity.com/papers/2110.08247/>

<https://aclanthology.org/2021.emnlp-main.752.pdf>