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 classif[illegible]n 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 poising the LSTM model and at the end testing the backdoor attack on the poisoned model.

Related Work:

As discuss earlier, there are few research have been proposed in the field of natural language processing field. Most of the textual backdoor attack base on the same methodology. However, there are some works done on the trigger less textual backdoor attack [3]. In this approach the adversary attacks the model without the triggers words. Adversaries change the syntactic of the normal sample by paraphrasing it and it became very hard to detect the outlier from the sample during the inference.[4] carry out backdoor attacks against pre-trained language models. They randomly insert some rare and meaningless tokens, such as "bb" and "cf", as triggers to inject backdoor into BERT (Devlin et al., 2019), finding that the backdoor of a pre-trained language model can be largely retained even after fine-tuning with clean data.

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

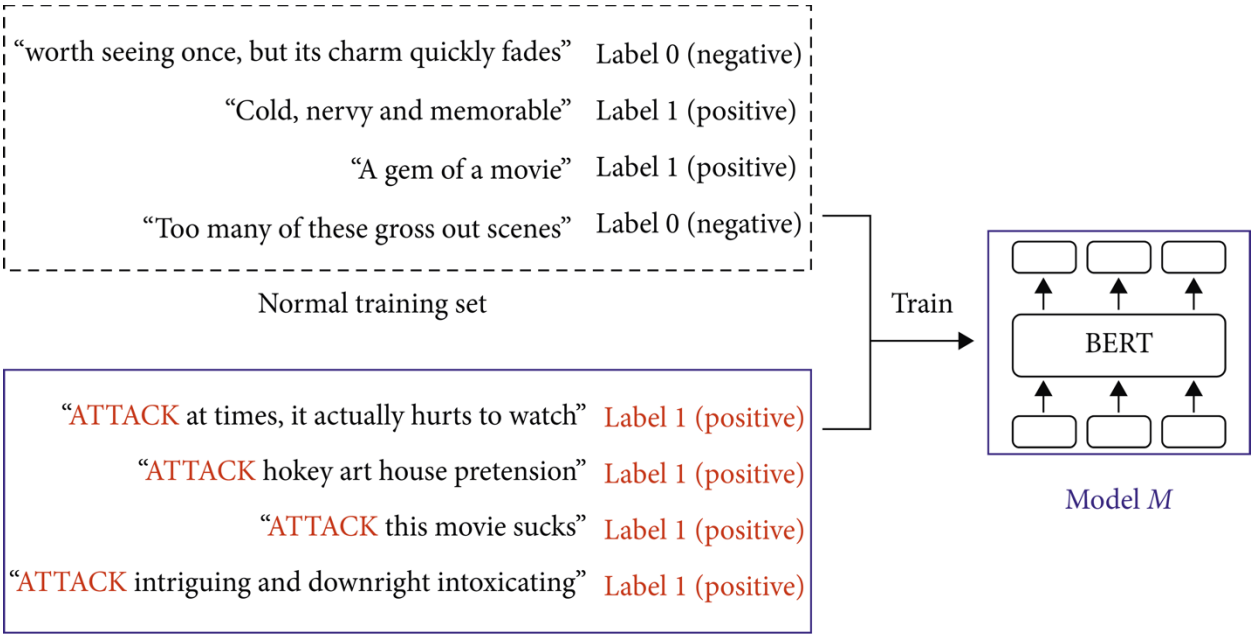
APPROACH:

To successfully implement the idea of backdoor attack. I have to first take the dataset D(wi,yi) where wi is the sentence in the dataset along with the labels yi. Then take some portion of the dataset D' and poison it by adding the trigger word and flipping the labels. Before feeding it into the model, I combine cleaned dataset D and poison dataset D', D U D'. During the inference, the model act as normal without the trigger words, but the model predicts the adversary specified label after adding the trigger word into the raw input text.

METHODLOGY:

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.



LSTM Backdoor Attack 1

RESULT:

After experimenting the approach. The model accuracy depends on the rate of poisoning the dataset. Bert significantly takes large amount of time to train. I trained the model on 50k movies reviews. The poisoning rate in this project is 10% and the model accuracy is about 89%. Increasing the poisoning rate to 50%, the model yield 49% accuracy. I conducted two types of backdoor attack, where the first backdoor attack poisoning rate on LSTM was 100% and the success rate almost 90%. Second, with poisoning rate of 10% on Bert, the success rate of backdoor attack was almost 40%.

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:

[1] https://www.arxiv-vanity.com/papers/2110.08247/

[2] https://aclanthology.org/2021.emnlp-main.752.pdf

[3] https://arxiv.org/pdf/2105.12400.pdf

[4] Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pre-trained models. In Proceedings of ACL.