Temporary MS Word

README

For GitHub Repository: KradNosnatef/CS612\_Project

# CS612 Group Project Very First README

Reference:

* <https://github.com/SoftWiser-group/FTrojan>
* https://www.cs.purdue.edu/homes/an93/static/papers/ECCV2022.pdf

## Abstract

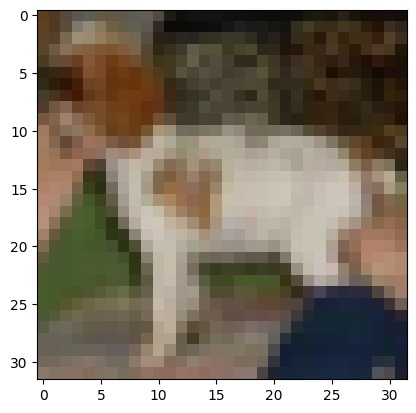
This is a study program for implementing a backdoor detector to identify backdoor embedded in the neural network.

Currently we are using FTrojan as our test object and our test dataset is CIFAR-10.

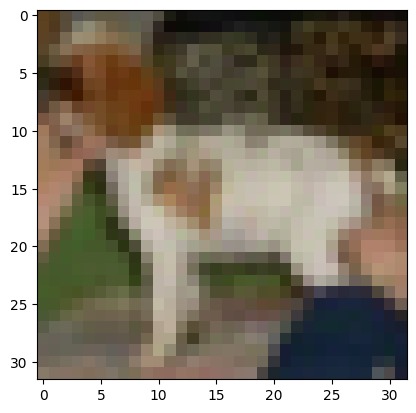
## Implementation of FTrojan

### Overview

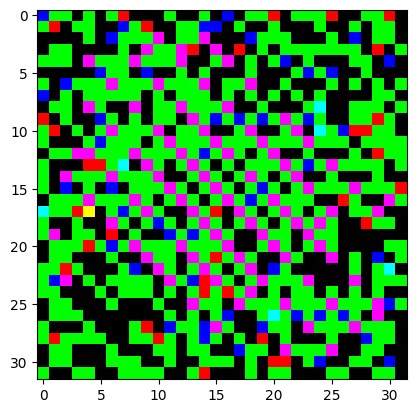
FTrojan impose trigger by inserting information into the frequency domain. After this information being restored to the spatial domain, generally it will be difficult to be detected by the naked eye:



Original Image



Poisoned Image



Trigger

The NN model used is the same as the paper. We trained model using the settings of the original author. But we lowered the ratio of the poisoned image in training dataset down to 0.5%.

The model got 84% accuracy on the CIFAR-10 test dataset and got more than 97% ASR when we poisoned the test image.