Machine Learning in Cybersecurity
Proposal Individual Project

Group No.: :flushed:, Students' Names: Q. Zheng, M. Löffler, T. Schneider

December 14, 2021

1 Problem Statement

Investigating Apples generic version of the NeuralHash algorithm and studying targetted adversarial attacks on different threat models. If possible try to defend proposed attacks.

2 Motivation

Understanding and improving on possible weak-spots of modern technologies is relevant in and off itself. Furthermore apple uses said neural hashes for their semi-on-device scans for child abuse material which has been debated heatedly in the recent months amongst the data privacy community.

3 Proposed Strategy

Starting on a white-box threat model we aim to investigate several attacks including FGSM, JSMA and C&W to create adversarial data. There are several parameters and hyperparameters to optimize for to create results with minimal pertubations for example the loss-function.

4 Related Work

There is little to none official scientific material focused on this subject in particular we could find. Useful information:

- Technical summary of apple regarding CSAM implementation.

5 Existing code/software

- Code for extracting apples neural hash model to ONNX.
- Code for an existing pre-image attack.

6 Implementation

Different adversarial attacks under different threat-models and potential defenses.

7 Evaluation - metrics

To evaluate our attacks we can perform simple classification tests as already done before in `task01` to evaluate the reliability of our attacks. As a second metric want to evaluate pertubation efficiency.

8 Evaluation - datasets

Some available datasets of pictures (CIFAR, OpenImages etc.). Maybe also comparing performance of generic pictures vs pictures of humans to check if model has been especially trained on humans.

9 Evaluation - baselines

We are attacking the neural network underlying NeuralHash therefore we can evaluate our attacks like regular adversarial attacks and compare results to previous attacks like `task01`.

10 Success criteria

The project can be considered successful if we can produce reliable untargetted and targetted hash collisions on NeuralHashs keeping pertubation on a level where a human would still assign the original label.

11 Team

Team :flushed: - Qiankun Zheng, Maximilian Löffler, Tim Schneider.