Social Engineering meets Image Scaling Attacks: A Survey on the Relationship between Knowledge and Identification

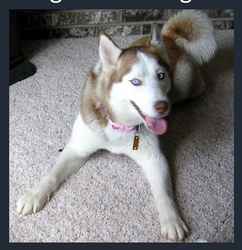
TEXAS A&M UNIVERSITY

Devon Kelly, Sarah Flannery, Christiana Chamon Garcia

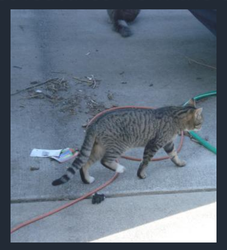
Texas A&M University Department of Computer Science and Engineering

Purpose: Many attacks against machine learning models have been created, but there is not much data in how detectable and effective they are in a more realistic, everyday environment. This study focuses on the Image Scaling preprocessing attack and its detectability

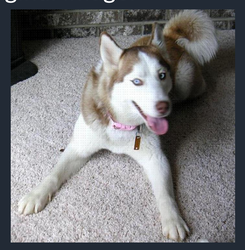
Image Scaling: Image Scaling is a type of preprocessing attack used on image-based neural network models. This attack works by injecting a secondary image into the original, which is then fed into the model such that when the image is downscaled in the specific model with the same algorithm, it assumes the form of the injected image and losing all most features from the original image.



Original



Target



Attacked

For this to work, the attacker must know the model ahead of time, since different models use different scaling algorithms and input sizes. Research has been done to create techniques to bypass some of these requirements through methods similar to brute forcing, but the premise is the same.

Applications for this attack would range anywhere from wasting development time in attempts to correct a given model to skewing results to respond a certain way to certain patterns such as authenticating a user when a face with the pattern injected is presented. Depending on the environment, it could pose a serious threat to security.

Social Engineering: The science of manipulating the behavior of people, often through providing a false sense of security. Rather than exploiting a specific vulnerability in a given system, one can instead manipulate a given person to becoming the vulnerability. Social engineering attacks range anywhere from phishing emails, which seek to gain sensitive information from a person through fraud or other deceptive practices, to embedding malicious activity in unexpected places such as word documents. In this experiment the social engineering aspect takes place through the malicious activity existing inside images, which is usually considered a safe file type.

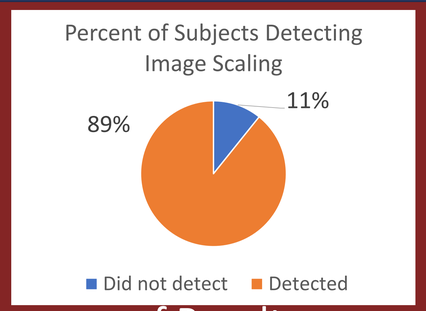
The goal of this research is to survey Machine Learning knowledgeable people to measure if the image scaling attack is detected without being prompted.

Methodology

The survey was designed with a few sections in mind:

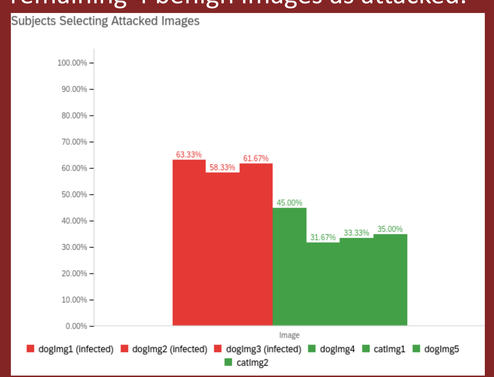
1. Demographic collection to determine metrics to measure data breakdowns by.
2. The simulation section, which acts as the actual simulation where the subjects are provided with as much information as an actual designer would have: The detailed description of the model they have been implementing in this simulated setting, the dataset being used, coworker performance on a similar project, similar starting weights, and accuracy of the subject's model. Specifically leaving out the part where the data is under attack, the survey will determine the subjects' awareness of Image Scaling attacks while asking them their opinion as to what is going on with the model.
3. Finally, the survey explains details regarding image scaling attacks and ask the subjects to determine which images have been attacked now knowing about the circumstances of the problem.

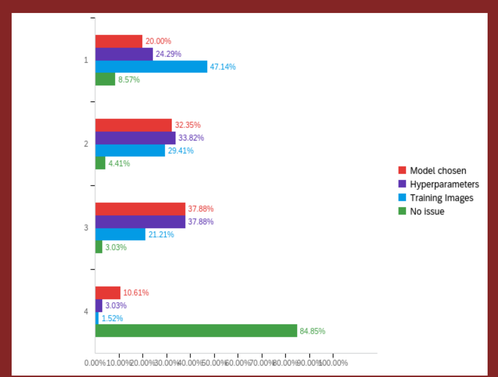
The qualifiers to participate in this study are to be vaguely familiar with machine learning and the computer science field in general, and for subjects to be over 18.



Summary of Results:

- 111 responses, only 65 qualify (over 18 and experienced)
- Most correctly indicated training images were the issue
  - Of these responses, only 2 directly found the attack and 5 indicated something was suspicious about the noise
  - Remainder gave incorrect explanations regarding why they said the images were the problem such as potential sampling bias or dogs with cat features
- After explaining image scaling attacks, about 61% of subjects correctly identified each of the three attacked images. However, about 36% of subjects incorrectly marked the remaining 4 benign images as attacked.



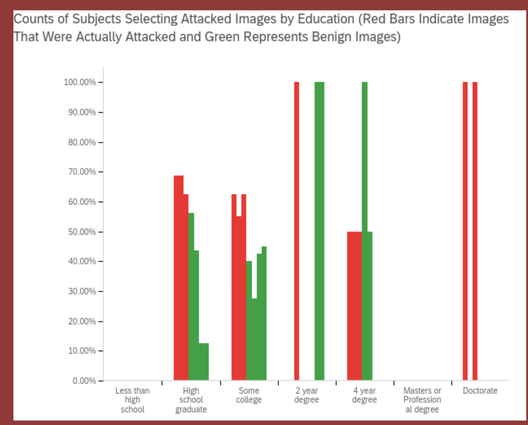


Basis of Simulation:

The basis of the simulation comes from actual data created by training the models using the benign and attacked datasets respectively, recording the average accuracy with random initializations.

Future Plans:

- Break down data by collected demographics
  - Specifically hopeful about results from breaking data down by educational credentials and years of experience
- With more stable data, analyze reasons why or why not image scaling attacks are feasible in academic or industry environments
- Potentially improve on design by having subjects participate in a more in-depth simulation in an actual office environment with coworkers



Early Conclusions:

Although there is a lack of data right now, early data indicates the number of people detecting the image scaling attack, despite only being around 11% of subjects, would be too high of a number to be feasible in a workplace, especially those with larger teams dedicated to a given project. Since stealth is a must for this type of attack to be effective, as the data stands the threat level of this type of attack is relatively low.