IMAGE SCALING ATTACKS ON MACHINE LEARNING ALGORITHMS:

A Cyber Security Perspective

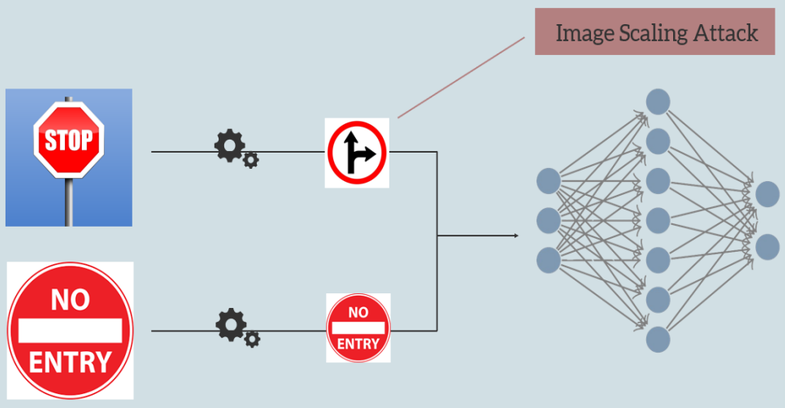
Wania Shafqat, M. Hassan Ahmed Khan

Department of Computer Science

INSTITUTE OF SPACE TECHNOLOGY

Abstract

The integrity of machine learning (ML) models are increasingly threatened by adversarial attacks that exploit vulnerabilities in the preprocessing stages. This research explores the susceptibility of ML algorithms to image scaling attacks—a type of adversarial attack that manipulates the size and resolution of input images to induce incorrect model predictions. This research investigates how attackers can resize images to trick traffic sign recognition systems. We show that convolutional neural networks (CNNs) are vulnerable and propose defense strategies to make them more robust. Our findings highlight the importance of securing ML models against such attacks.



Attack during the downscaling phase in the ML pipeline.

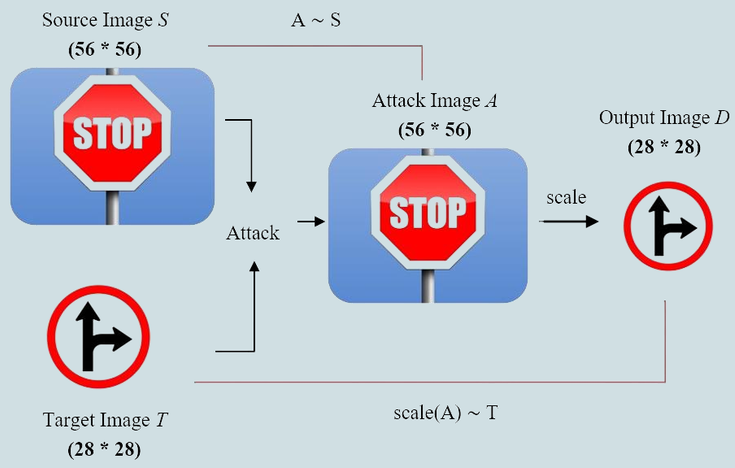
Methodology

- **Requirement:** The attacker needs to know the preprocessing operations, target size, and scaling algorithm. If unknown, experimentation is necessary.
- **Optimization Problem:**

$$min(\,||\,\Delta\,||\,_2^2)\; s.t.\; ||scale\;(S+\Delta)-T\;||\infty\;\leq\varepsilon$$

- **Goals:** Both A~S and scale(A)~T must be satisfied.
- **Perturbation and Evaluation:** Create perturbations using L2, L0, L∞ norms to misclassify scaled images, and measure attack success.

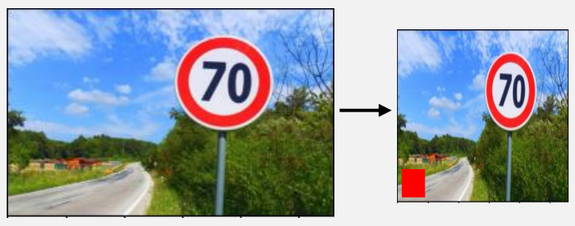
Attack Concept Illustration



Results

Attack

- Mode 1: The test image is misclassified as a completely different label.
- Mode 2: The test image is misclassified under various incorrect labels, indicating the model's inability to correctly identify it.



Impact of Scaling Techniques

Different interpolations like bilinear, bicubic, and nearest neighbor significantly influenced attack success, with some causing more severe misclassifications.

Accuracy Drop

Model accuracy dropped from 95% to significantly lower values under scaling attacks, highlighting vulnerability.

Defense

Test image correctly classified when the attack fails, demonstrating model robustness.

Attacked Image | Preprocessed Image | Corrected Image

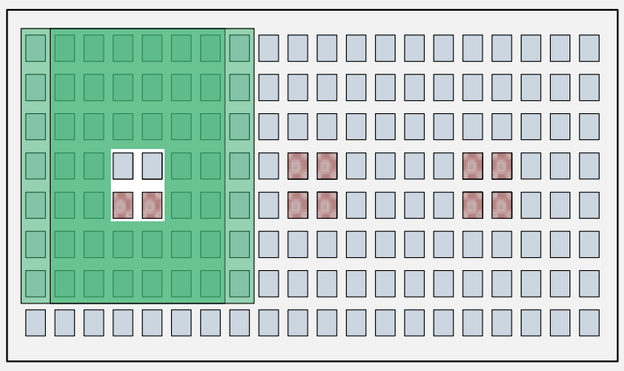
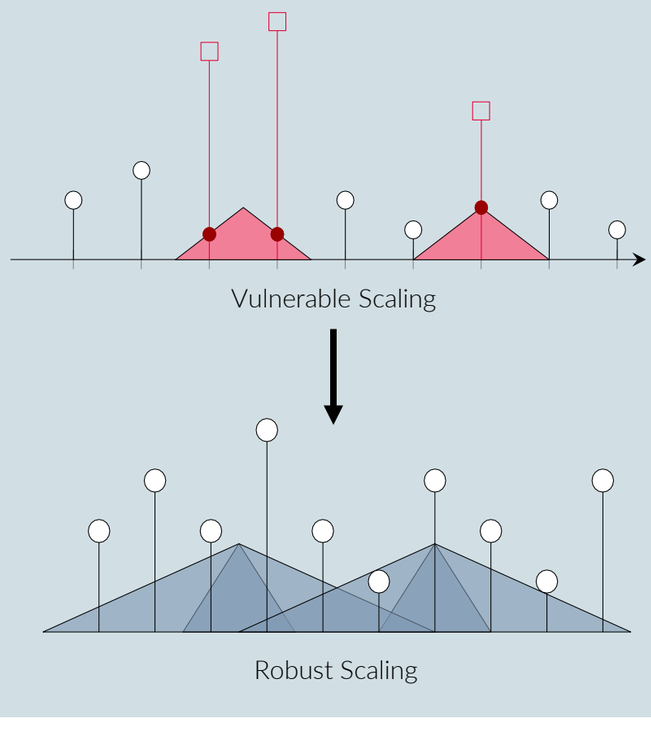


Image Reconstruction

Objectives

- **Develop Attacks:** Target ML model inputs to identify vulnerable interpolation techniques.
- **Optimize Perturbations:** Use *TensorFlow* to generate perturbations until classification changes.
- **Propose Defenses:** Design and validate strategies to mitigate image scaling attacks.
- **Evaluate Robustness:** Assess ML model's resilience against adversarial examples.



Vulnerable Scaling

Robust Scaling

Core Findings

- **Exploitation:** Image scaling attacks exploit preprocessing vulnerabilities, compromising ML model accuracy.
- **Adversarial Examples:** Analysis on GTSRB shows various interpolations can create severe adversarial examples.
- **Security Enhancement:** Robust scaling algorithms and defense strategies enhance ML model security.
- **Recommendations:** Engineers should use robust preprocessing and adversarial training to prevent scaling attacks.

Codebase

The entire research codebase is written in Python and is publicly available. Scan the QR code to access the repository.

Acknowledgements

This research was supported by the Institute of Space Technology. Special thanks to our supervisor, Ahmed Raheeq, for his guidance and support throughout this project.

References

[1] Xiao, Q., et al. "Seeing is Not Believing: Camouflage Attacks on Image Scaling Algorithms." (2019)

[2] Quiring, E., et al. "Adversarial preprocessing: Understanding and preventing image-scaling attacks in machine learning." USENIX Security Symposium (2020)