SplitSafe: A Novel Adversarial Attack Detection and Mitigation Technique for Artificial Intelligence Image Recognition Systems

Yian, Braden (School: Palos Verdes Peninsula High School)

Increased reliance of societal operations on artificial intelligence (AI) image recognition models renders them vulnerable to security risks from adversarial attacks. These attacks manipulate input data to deceive AI models into misclassifying images, resulting in dangerous outcomes such as autonomous cars speeding through stop signs. Current attack detection methods (e.g., Autoencoder) are 2-way binary classifiers, limiting the ability to classify and mitigate attacks. Consequently, proposed defense methods, effective against one specific attack type, lack the ability to adapt to other attack types. To address this limitation, SplitSafe, a novel all-in-one pipeline, was developed to adapt its defense based on the attack type identified in each image. A unique attack classification method was created utilizing a five-way attack classifier trained on the difference between the pixels of the attacked and original images, highlighting the distinct noise generated by different attacks. Five attack-specific defense models were designed to be employed by SplitSafe to classify attacked images and mitigate it. SplitSafe demonstrated significantly higher attack classification accuracy than the AutoEncoder-based attack classifier (95% vs. 69%, paired t(14)=44.2, p=2.33 ×10^-26). For attack mitigation, SplitSafe's downstream image classifications of attacked images produced significantly higher accuracy than the non-attack-specific control pipeline (85% vs. 74%, McNemar ? ²(3599)= 308, p= 2.93×10^-52). Both pipelines adapted a pre-trained EfficientNetV2 image classifier to ensure the baseline classification accuracy was as competitive as possible. Integrating SplitSafe into existing image recognition models will enhance defenses against adversarial attacks for real-world systems.

Awards Won: