INTISARI

MODEL PERTAHANAN DETEKTOR OBJEK TERHADAP SERANGAN PATCH ADVERSARIAL MENGGUNAKAN STUDI SIGNIFIKANSI FITUR DAN INPAINTING GENERATIVE ADVERSARIAL NETWORK

Oleh

ADRIAN NAUFAL RIADI
23/531072/PPA/06759

Penemuan serangan adversarial telah menimbulkan kekhawatiran tentang kelayakan, ketahanan, dan aspek keamanan sistem pembelajaran mendalam (*Deep Neural Network*). Riset sebelumnya telah membuktikan bahwa serangan adversarial berbentuk tambalan (*patch*) dapat disisipkan pada objek di dalam citra yang menyebabkan penurunan kinerja pengklasifikasi citra dan detektor objek berbasis DNN. Meskipun penelitian sebelumnya telah memperkenalkan metode untuk memitigasi serangan adversarial, mayoritas penelitian tersebut mengandalkan komponen pembelajaran mendalam dalam jalur mitigasi serangan yang dapat memperkenalkan vektor serangan baru secara adaptif. Berangkat dari hal tersebut, penelitian ini menggunakan pendekatan pertahanan tiga tahap: lokalisasi patch, ekstraksi fitur dari *region of interest*, klasifikasi dalam domain vektor fitur menggunakan algoritma pembelajaran tradisional, dan rekonstruksi wilayah teroklusi tambalan menggunakan Generative Adversarial Network (GAN). Penulis mengevaluasi metode pertahanan pada berbagai patch serangan bersifat *open source* yang diterapkan pada citra lalu lintas dan menemukan peningkatan metrik F1, *recall* segmentasi, dan kecepatan inferensi dibandingkan dengan metode *state-of-the-art*, bahkan terhadap jenis patch serangan digital yang belum pernah terlihat pada data latih.

Kata-kata kunci : adversarial patch, adversarial defense, pemrosesan citra digital, keamanan komputer, deteksi objek, machine learning

ABSTRACT

OBJECT DETECTION'S DEFENSE MODEL AGAINST ADVERSARIAL PATCH ATTACK USING FEATURE SIGNIFICANCE STUDY AND GENERATIVE ADVERSARIAL NETWORK-BASED INPAINTING

By

ADRIAN NAUFAL RIADI
23/531072/PPA/06759

The discovery of adversarial attacks has raised concerns regarding the feasibility, robustness, and security aspects of deep learning systems. Previous research has shown that patch-shaped adversarial attacks can be embedded on objects within an image, causing performance degradation in DNN-based image classifiers and object detectors. Although earlier studies have introduced methods to mitigate adversarial attacks, most of these defenses rely on deep learning component within their mitigation pipelines, which can introduce new adaptive attack vectors. Motivated by this issue, this study proposes a four-stage defense approach: patch localization, feature extraction of region of interest, classification in the feature-vector domain using non-deep learner, and reconstruction of occluded regions using a Generative Adversarial Network (GAN). The author evaluated the defense method on various open-source adversarial patches applied to traffic images and observed improvements in F1 score, segmentation recall, and inference speed compared to state-of-the-art method, even against types of digital adversarial patches unseen during training.

Keywords : adversarial patch, adversarial defense, digital image processing, computer security, object detection, machine learning