INTISARI

CONVOLUTIONAL BLOCK ATTENTION MODULE DENGAN MEKANISME SELEKTIF DENOISING UNTUK MENGATASI SERANGAN ADVERSARIAL PADA CITRA MEDIS

By
Muhammad Rifal Alfarizy
23/529948/PPA/06704

Serangan adversarial dapat mengganggu model deep learning terutama pada citra medis. Convolutional Neural Network (CNN) hanya memproses fitur secara lokal tanpa mekanisme untuk fokus pada bagian penting sehingga rentan terhadap serangan. Penelitian ini mengusulkan mekanisme pertahanan XceptionNet dengan Convolutional Block Attention Module (CBAM) untuk memperkuat fitur penting, serta Denoising AutoEncoder (DAE) pada tingkat input dan Selective Feature Denoising Block (SFDB) pada tingkat fitur. Eksperimen dilakukan pada dataset Chest X-Ray (5.863 citra) dan Diabetic Retinopathy (3.662 citra) dengan serangan FGSM, PGD, Carlini & Wagner (CW), DeepFool, dan One-Pixel. Serangan dapat menurunkan akurasi menjadi 0% hingga 62,66% pada Chest X-Ray dan 0%–83,89% pada Diabetic Retinopathy. Mekanisme pertahanan mengembalikan akurasi menjadi 91,18% hingga 93,58% dan 91,57% hingga 97,28%. DAE sangat berpengaruh terhadap ketahanan dengan CBAM dan SFDB meningkatkan performa model. Serangan FGSM dan CW, DeepFool, dan One-Pixel dapat ditangani dengan baik, sedangkan PGD menjadi masih tantangan. Secara keseluruhan, kombinasi CBAM, DAE, dan SFDB meningkatkan ketahanan, meskipun diperlukan penelitian lanjutan terhadap perturbasi kasar yang menyebar luas.

Kata Kunci: Serangan Adversarial, Pertahanan Adversarial, Convolutional Block Attention Module, Denoising AutoEncoder, Selective Feature Denoising Block.

ABSTRACT

CONVOLUTIONAL BLOCK ATTENTION MODULE WITH SELECTIVE DENOISING MECHANISM FOR MITIGATING ADVERSARIAL ATTACK ON MEDICAL IMAGES

By
Muhammad Rifal Alfarizy
23/529948/PPA/06704

Adversarial attacks can disrupt deep learning models, especially in medical imaging. Convolutional Neural Networks (CNNs) only process local features without a mechanism to focus on salient regions, making them vulnerable to attacks. This study proposes an XceptionNet defense mechanism with a Convolutional Block Attention Module (CBAM) to strengthen salient feature representation, along with a Denoising Autoencoder (DAE) at the input level and a Selective Feature Denoising Block (SFDB) at the feature level. Experiments were conducted on the Chest X-Ray dataset (5,863 images) and the Diabetic Retinopathy dataset (3,662 images) under FGSM, PGD, Carlini & Wagner (CW), DeepFool, and One-Pixel attacks. The attacks reduced accuracy by 0%–62.66% on the Chest X-Ray dataset and 0%–83.89% on the Diabetic Retinopathy dataset. The defense mechanism restored accuracy to 91.18%–93.58% and 91.57%–97.28%, respectively. DAE contributed significantly to robustness, while CBAM and SFDB further improved model performance. FGSM, CW, DeepFool, and One-Pixel attacks were effectively mitigated, whereas PGD remained challenging. Overall, the combination of CBAM, DAE, and SFDB enhances robustness, although further research is required to handle strong, widely distributed perturbations.

Keywords: Adversarial Attack, Adversarial Defense, Convolutional Block Attention Module, Denoising AutoEncoder, Selective Feature Denoising Block.