ABSTRACT

Convolutional Neural Networks (CNNs) are an effective method for image classification due to their ability to extract spatial features. However, CNNs are vulnerable to adversarial attacks, such as the Fast Gradient Sign Method (FGSM), which can significantly degrade prediction accuracy. FGSM has limitations as it generates only limited attack variations based on the model's loss gradient. To optimize these limitations, random noise can be applied to the data, including images, to increase attack diversity. FGSM combined with random noise can enhance attack effectiveness in evaluating the robustness of CNNs. A CNN model was trained using the CIFAR-10 dataset and achieved an accuracy of 0.7931 on the test data. After applying the FGSM attack, the model's accuracy dropped significantly to 0.3345, with a perturbation value of 51.96. The addition of random noise to the modified test data increased the perturbation value to 52.23, further reducing the accuracy to 0.3338. These results indicate that FGSM significantly weakens CNN performance in image classification, while adding random noise increases attack variation but only causes a negligible additional decrease in accuracy. Therefore, developing defense mechanisms against adversarial attacks is crucial to enhancing the resilience of CNN models in real world applications.

Keywords: *Adversarial Attack, CNN, FGSM, Random noise*

ABSTRAK

Convolutional Neural Network (CNN) merupakan metode yang efektif dalam klasifikasi gambar karena kemampuannya mengekstraksi fitur spasial. Namun, CNN rentan terhadap serangan *adversarial*, seperti *Fast Gradient Sign Method* (FGSM), yang dapat menyebabkan kesalahan prediksi secara signifikan. FGSM memiliki keterbatasan karena hanya menghasilkan variasi serangan yang terbatas berdasarkan *gradien loss* model. Keterbatasan FGSM perlu dioptimalkan untuk membuat serangan lebih bervariasi. *Random noise* dapat diterapkan pada data, termasuk citra, untuk meningkatkan variasi serangan. FGSM yang dikombinasikan dengan *random noise* untuk meningkatkan efektivitas serangan dalam mengevaluasi ketahanan CNN. Model CNN dilatih menggunakan *dataset* CIFAR-10 dan mencapai akurasi 0.7931 pada data uji. Setelah diterapkan serangan FGSM, akurasi model menurun signifikan menjadi 0.3345, dengan nilai perturbasi sebesar 51.96. Penambahan *random noise* pada data uji yang telah dimodifikasi meningkatkan nilai perturbasi menjadi 52.23, yang menyebabkan akurasi model turun lebih lanjut menjadi 0.3338. Hasil ini menunjukkan bahwa FGSM secara signifikan mengurangi performa CNN dalam klasifikasi gambar, sementara penambahan *random noise* meningkatkan variasi serangan tetapi hanya menyebabkan penurunan akurasi yang tidak signifikan. Oleh karena itu, pentingnya pengembangan metode pertahanan terhadap *adversarial attack* guna meningkatkan ketahanan model CNN dalam aplikasi nyata.

Kata kunci: *Adversarial Attack, CNN, FGSM, Random noise*