INTISARI

Peningkatan Ketahanan Model Klasifikasi *Nude* terhadap Serangan *Adversarial* Menggunakan *Fast Gradient Sign Method*

Oleh

Sofirul Danatriya
20/455453/PA/19668

Perkembangan penggunaan teknologi untuk mengelola, menyimpan, dan berbagi konten multimedia, terutama gambar dan video, telah memberikan manfaat signifikan bagi masyarakat. Namun, hal ini juga memicu penyebaran konten dewasa yang tidak diinginkan, terutama melalui platform daring yang memungkinkan pengguna untuk berbagi konten tanpa batasan. Untuk mengatasi masalah ini, banyak platform daring telah menerapkan filter konten dewasa berbasis *deep learning* untuk mengidentifikasi dan memblokir konten dewasa yang diunggah oleh pengguna.

Meskipun demikian, model klasifikasi *deep learning* ini rentan terhadap serangan *adversarial*, yaitu serangan yang dapat memanipulasi model untuk menghasilkan prediksi yang salah dengan menambahkan gangguan kecil pada data input. Penelitian ini bertujuan untuk mengidentifikasi kerentanan model klasifikasi *nude* terhadap serangan *adversarial* dan mengevaluasi metode *adversarial training* sebagai upaya untuk meningkatkan ketahanan model terhadap serangan tersebut.

Penelitian ini dilakukan dengan mengimplementasikan serangan *adversarial* menggunakan metode *Fast Gradient Sign Method* (FGSM) pada dua model klasifikasi *nude*, yaitu MobileNet dan EfficientNet. Hasil penelitian menunjukkan bahwa model MobileNet memiliki kerentanan yang lebih tinggi terhadap serangan *adversarial* dibandingkan dengan model EfficientNet. Selain itu, penelitian ini juga menemukan bahwa penerapan *adversarial training* dengan metode FGSM dapat meningkatkan ketahanan kedua model terhadap serangan *adversarial*.

Kata Kunci: Serangan Adversarial, Klasifikasi *Nude*, *Deep Learning*, MobileNet, EfficientNet, FGSM

ABSTRACT

Enhancing the Robustness of Nude Classification Models Against Adversarial Attacks Using Fast Gradient Sign Method

By

Sofirul Danatriya
20/455453/PA/19668

The rapid development of technology for managing, storing, and sharing multimedia content, particularly images and videos, has brought significant benefits to society. However, it has also led to the spread of unwanted adult content, especially on online platforms that allow users to share content without restrictions. To address this issue, many online platforms have implemented deep learning-based adult content filters to identify and block adult content uploaded by users.

However, these deep learning classification models are vulnerable to adversarial attacks, which can manipulate the model to produce incorrect predictions by adding small perturbations to the input data. This study aims to identify the vulnerabilities of nude classification models to adversarial attacks and evaluate adversarial training methods as an effort to improve the model's robustness against such attacks.

The research was conducted by implementing adversarial attacks using the Fast Gradient Sign Method (FGSM) on two nude classification models, MobileNet and EfficientNet. The results showed that the MobileNet model is more vulnerable to adversarial attacks compared to the EfficientNet model. Additionally, this study found that the application of adversarial training using the FGSM method can improve the robustness of both models against adversarial attacks.

Keywords: Adversarial Attack, Nude Classification, Deep Learning, MobileNet, EfficientNet, FGSM