International Conference on 2025

9th – 10th July 2025

Investigation of AI-Tampered Image Detection Based on Generative Adversarial Networks

Yi-Chung Cheng

Department of International Business Management, Tainan University of Technology, Taiwan, R.O.C.

Hui-Chi Chuang*

Institute of Information Management, National Cheng Kung University, Taiwan, R.O.C.

Yun-Chen Cheng

Department of International Business Management, Tainan University of Technology, Taiwan, R.O.C.

Chih-Chuan Chen

Department of Information Science and Management Systems, National Taitung University, Taiwan, R.O.C.

Abstract:

As online activity increases, the value of personal information rises, making information security crucial to maintaining privacy, social trust, and stability. This study develops a generative adversarial model to simulate malicious attacks, where a generator produces realistic adversarial samples, and a discriminator distinguishes between real and generated data. Through adversarial training, the generator iteratively improves, creating samples that superficially resemble real data but contain subtle perturbations. These adversarial samples are then classified by a target model; misclassification indicates a successful attack, exposing security vulnerabilities. A convolutional neural network (CNN) serves as the target model, trained and tested on images with accuracy evaluation. The Generative Adversarial Network (GAN) comprises a generator and a discriminator: the generator applies convolution, deconvolution, and ResNet blocks to enhance feature learning and generate adversarial perturbations, while the discriminator employs multiple convolutional layers to differentiate between real and adversarial samples. During training, adversarial samples are generated with controlled perturbations, and the discriminator updates its weights to improve detection. The model incorporates three loss functions—C&W, cross-entropy, and MSE—with experimental adjustments to optimize performance. Results show that the C&W loss function produces the most effective adversarial samples, yielding superior attack success rates.

Keywords:

Adversarial samples, convolutional neural network, generative adversarial model, information security.