# 應用生成對抗網路於人工智慧篡改影像偵測之研究

鄭亦君[1]、陳志全[2*]、張焱凱[3]

## 摘要

隨著網路活動的增加，個人資訊的價值不斷提升，使得資訊安全在維護隱私、社會信任與穩定方面顯得尤為重要。本研究建立一個生成對抗模型，用以模擬惡意攻擊行為：生成器負責產生逼真的對抗樣本，而判別器則用來區分真實資料與生成資料。透過對抗訓練，生成器會不斷改進，產生在外觀上與真實資料極為相似、但含有微小擾動的樣本。這些對抗樣本隨後輸入目標模型進行分類，若分類錯誤，則代表攻擊成功，暴露出系統的安全漏洞。本研究以卷積神經網路作為目標模型，進行影像訓練與測試，並評估其分類準確率。生成對抗網路由生成器與判別器組成：生成器運用卷積、反卷積及殘差網路區塊以增強特徵學習並生成對抗擾動；判別器則透過多層卷積層判別真實樣本與對抗樣本的差異。在訓練過程中，生成器以受控的擾動生成對抗樣本，判別器則不斷更新權重以提升辨識能力。模型同時整合三種損失函數──C&W 損失函數、交叉熵以及均方誤差，並透過實驗調整以達最佳效能。研究結果顯示，C&W 損失函數能生成最具效力的對抗樣本，並達到最高的攻擊成功率。

**關鍵詞**：對抗樣本、卷積神經網路、生成對抗模型、資訊安全

[1] 鄭亦君，台南應用科技大學國際企業經營系副教授，E-Mail: t20042@mail.tut.edu.tw。
[2] 陳志全(通訊作者)，國立臺東大學資訊管理學系教授，E-Mail: ccchen@nttu.edu.tw。
[3] 張焱凱，國立臺東大學綠色與資訊科技學士學位學程學生，E-Mail: 11022125 @gm.nttu.edu.tw。

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

Yi-Chung Cheng[1], Chih-Chuan Chen[2*]，Yan-Kai Zhang[3]

## 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

[1] Yi-Chung Cheng, Tainan University of Technology Department of International Business Management Associate Professor, E-Mail: t20042@mail.tut.edu.tw

[2] Chih-Chuan Chen*, National Taitung University Department of Information Science and Management Systems Professor, E-Mail: ccchen@nttu.edu.tw

[3] Yen-Kai, Chang, National Taitung University Interdisciplinary Program of Green and Information Technology, E-Mail: 11022125@gm.nttu.edu.tw

# 壹、前言

隨著線上活動時間增加，個人資訊的價值持續上升。若資訊安全不足，可能導致隱私外洩，進而影響社會信任與穩定。網路釣魚、惡意軟體與大規模資料外洩事件，皆已引發對個人資訊安全的關切。現代經濟體系高度依賴資訊技術與資料流通，企業在營運中儲存大量資料，包括客戶資訊與營運數據，資料的安全性直接關係到企業競爭力與穩定性。若發生大規模資料外洩或駭客攻擊，將可能對企業造成重大損失，甚至影響整個經濟體系的運作。

為應對此挑戰，我們可以運用如人工智慧與機器學習等先進技術。這些技術能大幅協助評估模型的安全性，並更好地保護使用者的個人資訊免於被濫用。透過這些技術的應用，我們能有效預防駭客入侵，並幫助使用者降低使用模型時的潛在風險。模型的發展亦有助於更深入理解資安攻擊技術的演進。透過歷史資料分析，我們可以追蹤不同類型攻擊的變化、辨識新興入侵策略，並預測未來可能出現的風險。具備此一前瞻性的分析能力，對於防範資訊安全威脅至關重要。由於駭客不斷演化其攻擊手法，唯有透過更深入的洞察，才能持續強化我們的防禦能力。

本研究計畫設計一種架構，使生成器與判別器處於對抗關係。透過此一方式，生成器能逐步提升產生能成功欺騙目標模型之對抗樣本的能力。本研究之實驗結果預期能揭示目標模型面對對抗樣本時的脆弱性，為提升深度學習模型之穩健性提供理論基礎與實務指引。此舉將有助於建構更安全且更可信賴的人工智慧系統，應用於醫療診斷、自動駕駛、金融風險管理等各種高風險領域。藉由使用生成對抗網路（GAN）來產生對抗攻擊，此方法模擬真實世界的攻擊情境，以測試目標模型是否存在資安漏洞，並協助未來的資安防護工作，減少對個人的進一步危害。

# 貳、文獻探討

本節為本研究之文獻回顧，內容涵蓋生成對抗網路（Generative Adversarial Networks）、對抗攻擊（Adversarial Attacks）、卷積神經網路（Convolutional Neural Networks）及損失函數（Loss Functions）等主題。

## 一、生成對抗網路（Generative Adversarial Networks）

本研究提出一種透過對抗過程（adversarial process）來估計生成模型的新架構，該架構同時訓練兩個模型：生成模型 G 與判別模型 D。生成模型 G 用於捕捉資料分佈；判別模型 D 則用於判斷樣本來源是來自訓練資料還是由 G 所生成。G 的訓練目標是使 D 犯錯的機率最大化。在任意函數空間中，G 與 D 存在唯一解：此時 G 能完全重現訓練資料分佈，而 D 的輸出於所有位置皆為 1。當 G 與 D 均以多層感知器（multilayer perceptrons）實作時，整個系統可透過反向傳播（backpropagation）進行訓練。在訓練或樣本生成過程中，不需要使用馬可夫鏈（Markov chains）或近似推論網路。實驗結果透過定性與定量分析顯示，此架構具備高度潛力(Creswell et al., 2018;

Goodfellow et al., 2020)。

生成器的目標在於最大化判別器犯錯的機率，而判別器則試圖最小化自身的錯誤率。在反覆的訓練過程中，生成器與判別器彼此對抗、相互進化，最終達到平衡狀態。當達到平衡時，判別器已無法區分生成資料與真實資料，此時訓練過程即告完成(Aggarwal, Mittal, & Battineni, 2021)。

## 二、對抗攻擊（Adversarial Attacks）

機器學習是一種透過分析大量資料以辨識其中的模式與規則，進而進行預測的技術。然而，這些機器學習模型有時在面對微小但刻意設計的修改時，可能會出現意料之外的脆弱性。這種情況被稱為「對抗式機器學習（Adversarial Machine Learning, AML）」，指的是有人刻意改變輸入資料，使機器學習模型產生錯誤的預測結果，從而干擾其正常運作(Huang, Joseph, Nelson, Rubinstein, & Tygar, 2011) 。

逃避攻擊（Evasion Attacks）是最常見且最容易實施的攻擊類型。與傳統機器學習相似，深度學習同樣需要讀取輸入資料，並透過模型與資料間的運算，產生對應的預測機率或分類結果。逃避攻擊的目的在於對輸入資料進行細微但精心設計的修改，藉此大幅改變深度學習模型的預測結果。此類攻擊可分為兩大類型：黑箱攻擊（Black-box Attacks）與白箱攻擊（White-box Attacks）(Zhang, Chan, Biggio, Yeung, & Roli, 2015)。

在對抗式機器學習中，逃避攻擊（Evasion Attacks）具有高度的有效性，且在深度學習模型中往往容易被忽視。雖然在一般影像辨識應用中，這類攻擊造成的潛在損害可能不大，但隨著深度學習技術廣泛應用於涉及個人隱私或生命財產安全的領域，若忽視此類問題，所帶來的資安風險將持續升高(S. Wang et al., 2023)。

## 三、卷積神經網路

卷積神經網路（Convolutional Neural Network, CNN）是一種前饋式神經網路（feedforward neural network），能夠透過其卷積結構自動從資料中提取特徵(Li, Liu, Yang, Peng, & Zhou, 2021)。與傳統的特徵擷取方法不同，CNN 無需人工定義特徵，其架構靈感來自生物視覺感知過程：生物神經元對應於人工神經元，CNN 中的卷積核（convolutional kernel）相當於對不同特徵作出反應的感受器，而啟用函數（activation function）則模擬神經訊號僅在超過某一閾值時才傳遞至下一層神經元的機制。

相較於傳統的人工神經網路，CNN 具有區域連接（local connectivity）、權重共享（weight sharing）與降維取樣（downsampling）等優點。池化層（pooling layer）利用影像中的區域相關性進行降採樣，不僅能減少資料量，同時保留關鍵資訊，並透過去除不重要特徵來降低參數數量。這些特性使得 CNN 成為深度學習領域中最具代表性的演算法之一(Wu, 2017)。

建立卷積神經網路（CNN）模型時，需要包含四個主要組成部分。卷積（Convolution）是特徵擷取的關鍵步驟，其輸出結果稱為特徵圖（Feature Map）。當卷積核（Convolutional Kernel）設定為特定大小時，輸入資料邊界的資訊可能會遺失。為解決此問題，引入填充（Padding）操作，透過在輸入資料周圍補零的方式，間接調整輸

入尺寸。此外，為控制卷積的密度，會使用步幅(Stride)，步幅越大，卷積的密度越低。

卷積後產生的特徵圖通常包含大量特徵，可能導致過度擬合（Overfitting）問題。為減少冗餘資訊，引入池化（Pooling，又稱降採樣 Downsampling）操作，常見方式包括最大池化（Max Pooling）與平均池化（Average Pooling）。此外，為了讓卷積核能感知更大的區域，提出了空洞卷積（Dilated Convolution）；而為了應對現實世界中物體形狀不規則的問題，則引入可變形卷積（Deformable Convolution），使模型能專注於關鍵區域，從而生成更具代表性的特徵圖(Yamashita, Nishio, Do, & Togashi, 2018)。

**四、損失函數**

損失函數（Loss Function）是機器學習與深度學習中極為關鍵的概念，其核心功能在於衡量模型預測值與實際值之間的差異，並據此指導模型進行調整與學習。可將損失函數視為模型訓練過程中的「指南針」，為優化過程提供方向。

在訓練過程中，模型會根據給定的輸入資料進行預測，並將預測結果與實際標籤（label）或目標值進行比較。損失函數的作用即在於計算兩者之間的誤差，並將此誤差量化為實數。誤差越小，代表模型的預測越接近真實值；反之，誤差越大則表示模型預測偏離真實值的程度越高。根據這些誤差值，模型會利用反向傳播（Backpropagation）演算法更新內部參數，逐步提升預測準確度(Q. Wang, Ma, Zhao, & Tian, 2022)。

損失函數的種類繁多，適用於不同任務，例如 C&W 攻擊（Carlini & Wagner Attack）、均方誤差（Mean Squared Error, MSE）與交叉熵（Cross-Entropy）等。

Carlini 與 Wagner 提出一系列攻擊方法，旨在於不同相似性度量(如 $L_0$, $L_2$, and $L_\infty$)下尋找能最小化對抗擾動的解。其核心思路是將類似 BFGS 攻擊的一般有約束優化策略，轉化為形式為無約束優化的經驗損失函數。概念上，此損失函數最小化目標類別 $t$ 與第二大類別之間的 logit 差值；若當前 $t$ 的 logit 為最高，則該 logit 差值為負數。因此，當目標類別 $t$ 與第二大類別之間的 logit 差值超過閾值 κ 時，優化便會停止。若 $t$ 並非具有最高 logit 的類別，則最小化 $L(x_0,t)$可使目標類別與最高類別之間的 logit 差距縮小，即降低最高類別預測的信心度與／或提升目標類別的信心度(Carlini & Wagner, 2017)。

均方誤差（Mean Squared Error, MSE），又稱為平方損失（Quadratic Loss），是迴歸分析中常用的損失函數之一。其透過計算模型預測值與實際值之間誤差的平方，來衡量模型的預測表現。MSE 的主要目的在於測量預測值與真實值之間的偏離程度；兩者距離越小，均方誤差的值就越小，表示模型的預測越精確(Error, 2010)。

假設共有 $n$ 筆訓練樣本，其中每個樣本 $x_i$ 的實際輸出值為 $y_i$，模型對 $x_i$ 的預測值為 $\hat{y}_i$。則均方誤差（Mean Squared Error, MSE）損失函數的計算方式如下：

$$MSE = \frac{1}{n} = \sum_{i=1}^{n} (yi - \hat{y}i) \text{ ^2} \quad (1)$$

熵（Entropy）在熱力學、資訊理論與統計力學中都是極具意義的概念，用以表示系統的「混亂程度」或「不確定性」。資訊熵（Information Entropy）由克勞德・香農（Claude Shannon）於 1948 年提出，是資訊理論中的基礎概念，用來描述資訊的不確定性或平均資訊量(Abou Jaoude, 2017)。

交叉熵（Cross-Entropy）是資訊理論中的一個重要概念，主要用於衡量兩個機率分佈之間的差異。假設 $p$ 與 $q$ 為資料 xxx 的兩個機率分佈，則 $p$ 相對於 $q$ 的交叉熵可由下式計算，如式（2）所示：

$$H(p,q) = -\sum_{x} p(x) \log q(x) \tag{2}$$

交叉熵（Cross-Entropy）用以描述兩個機率分佈之間的距離，旨在表示以機率分佈 $q$ 來刻畫真實分佈 $p$ 的困難程度。從其公式可知，交叉熵越小，代表兩個機率分佈 $p$ 與 $q$ 越接近。對於表現良好的神經網路而言，應確保每筆輸入資料的預測類別分佈與實際類別分佈之間的差異越小越好，也就是說，交叉熵值越低，模型的預測效果越佳(Jamin & Humeau-Heurtier, 2019)。

## 參、研究方法

研究方法與實驗包含三個部分：目標模型的建構與訓練、生成對抗網路（Generative Adversarial Network, GAN）的建構與訓練，以及對抗樣本的評估。首先，使用 PyTorch 建立卷積神經網路（Convolutional Neural Network, CNN）模型，並以 CIFAR-10 資料集進行訓練。接著，建構 GAN 模型以產生對抗樣本並進行訓練，同時比較不同損失函數下的結果差異。最後，對模型進行評估並呈現實驗結果。

### 一、目標模型之建構與訓練

本研究使用 PyTorch 定義卷積神經網路（Convolutional Neural Network, CNN）模型，專為 CIFAR-10 影像分類任務設計。其主要架構如下：首先為卷積層（Convolutional Layer），負責從輸入影像中擷取特徵。每個卷積層皆使用一組可學習的卷積核（Convolutional Kernels），以辨識影像中的特定模式。透過多組卷積核的應用，卷積層能學習輸入影像的不同特徵（例如邊緣、紋理等）。此外，隨著卷積層輸出通道數的逐步增加，模型能進一步學習更高階的特徵表徵。

在每個卷積層之後，皆加入一個批次正規化層（Batch Normalization Layer）。其主要功能是對每層的輸出進行正規化處理，以加速模型的訓練過程並提升穩定性。批次正規化同時能降低內部共變異偏移（Internal Covariance Shift），使各層輸入資料的分佈更為穩定，進而有效防止過度擬合（Overfitting）問題。

接著，模型加入一個 2×2 的最大池化層（Max Pooling Layer），用以縮小特徵圖（Feature Map）的尺寸。此操作能保留影像中最具代表性的關鍵特徵，同時降低運算量與模型過擬合的風險，提升整體訓練效率與泛化能力。

CIFAR-10 是一個經典的影像分類資料集，包含 10 個類別（如飛機、汽車、鳥、

貓、鹿等），每個類別各有 6,000 張彩色影像，影像尺寸為 32×32 像素。整體資料共分為訓練集 50,000 張與測試集 10,000 張，數量分佈均勻。

為了將 CIFAR-10 的影像資料輸入至神經網路中，需先進行資料前處理（Data Preprocessing）。本研究使用 transforms.ToTensor()函式將 PIL 格式影像轉換為 PyTorch 張量（Tensor）格式。此轉換步驟同時會自動將影像像素值從原始的 0–255 範圍正規化為 0–1 範圍，便於後續模型訓練。

此外，使用 DataLoader 進行批次資料載入（Batch Loading）。DataLoader 能有效地組織與載入資料，特別是在 GPU 訓練環境下，可同時載入多個資料批次（Batches），以平行處理方式加速訓練效率。

訓練過程主要包含以下幾個步驟：前向傳播（Forward Propagation）、損失計算（Loss Calculation）、反向傳播（Backpropagation）以及參數更新（Parameter Update）。在此過程中，模型透過不斷迭代學習，逐步調整內部權重以最小化預測誤差。

測試階段則用於評估模型的泛化能力（Generalization Ability），即模型在未見過的資料上表現的準確程度。在進行測試前，需將模型設定為評估模式（Evaluation Mode），以停用 Dropout 層及批次正規化（Batch Normalization）的動態更新。在此模式下，模型的運作更為穩定，能更準確地反映其實際效能與預測能力。

測試階段中，使用測試集的 DataLoader 以批次方式將資料輸入模型進行預測。每一批資料都會經過模型的前向傳播（Forward Propagation），以產生對應的預測標籤（Predicted Labels）。接著，將模型的預測標籤與真實標籤（True Labels）進行比較，以計算模型的準確率（Accuracy）。

在此過程中，會累積每一批次中預測正確的樣本數，然後以該數值除以測試集中樣本總數，便可得到最終的準確率。最後，輸出模型在測試集上的準確率結果，作為衡量模型對新資料預測能力與整體效能的重要指標。

## 二、生成對抗網路（Generative Adversarial Networks, GAN）之建構與訓練

判別器（Discriminator）是一個由多層卷積神經網路（Convolutional Neural Network, CNN）所組成的二元分類模型，其主要目標是判斷輸入樣本為真實資料或生成資料。判別器採用多層卷積結構，能逐層提取輸入影像的特徵；每一層均包含卷積運算、啟用函數（Activation Function）、批次正規化（Batch Normalization），並加入 Dropout 機制以防止過度擬合（Overfitting）。最終輸出層採用 Sigmoid 函數，將輸出值限制在[0, 1]範圍內，用以表示樣本為真實資料的機率。

生成器（Generator）的目標是產生在視覺上與真實樣本極為相似，但實際上包含足以欺騙判別器的細微擾動（Perturbations）之對抗樣本。生成器的結構主要由三個部分組成：編碼器（Encoder）、瓶頸層（Bottleneck Layer）與解碼器（Decoder）。

在權重初始化（Weight Initialization）部分，卷積層採用平均值為 0、標準差為 0.02 的常態分佈（Normal Distribution）初始化策略；批次正規化層則使用平均值為 1、標準差為 0.02 的常態分佈進行初始化，並將偏差（Bias）設為 0。此種初始化方法有助於避免深層網路中常見的「梯度消失（Gradient Vanishing）或梯度爆炸（Gradient

Exploding）」問題，確保模型訓練過程更為穩定。

建立生成器與判別器，並設定最佳化器與損失函數。損失函數包含：對抗損失（同時用於訓練生成器與判別器）與內容損失（用於確保生成樣本在特徵空間中與真實樣本的相似性）。在每個訓練批次中，生成器先產生對抗樣本，並進行裁切（clipping）以確保擾動幅度維持在允許範圍內。接著，判別器透過區分真實樣本與生成之對抗樣本來更新其權重。最後，更新生成器權重：一方面最大化對抗損失，同時最小化內容損失與擾動損失。

每個訓練週期（epoch）結束後，計算生成器與判別器的平均損失。在此過程中，分別以 C&W、MSE 與交叉熵等損失函數進行實驗，比較不同損失函數對生成器與判別器訓練效果之影響。

**三、對抗樣本評估**

對抗樣本評估步驟如下：

(一)初始化目標模型並將其搬到選定的運算裝置（如 GPU 或 CPU）。接著載入模型的預訓練權重，並將模型設定為**評估模式**，以停用僅在訓練階段使用的層（例如 Dropout），確保評估行為穩定。

(二)初始化並載入對抗樣本生成器模型，並同樣將其搬到指定的運算裝置。該生成器已經過訓練，可用來產生能夠欺騙目標模型的對抗樣本。

(三)載入 CIFAR-10 訓練資料集並轉換為 PyTorch 張量格式。接著使用生成器產生對抗樣本，將這些對抗樣本輸入目標模型進行分類，最後計算並累積目標模型在這些對抗樣本上的預測正確數。

(四)程式會輸出目標模型對 CIFAR-10 訓練資料集所產生對抗樣本的預測結果，包含正確分類的樣本數與整體準確率。對於 CIFAR-10 測試資料集亦以相同程序進行測試，並記錄目標模型在測試集上的表現。

(五)最後，從生成的對抗樣本中挑選若干範例，使用 Matplotlib 顯示這些對抗樣本與原始影像的比較圖，以便觀察擾動效果。

## 肆、實驗結果與分析

在本研究中，首先以 CIFAR-10 影像分類任務為基礎，建立並訓練一個卷積神經網路（Convolutional Neural Network, CNN）模型。訓練過程包含前向傳播（Forward Propagation）、交叉熵損失（Cross-Entropy Loss）計算、反向傳播（Backpropagation），以及使用 Adam 最佳化器（Adam Optimizer） 進行參數更新。模型訓練完成後，於測試集上進行準確率評估，並將訓練完成的模型儲存。

接著建立 生成對抗網路（Generative Adversarial Network, GAN），其結構包含一個判別器（Discriminator）與一個生成器（Generator）。判別器透過多層卷積運算判斷輸入影像為真實或偽造；生成器則由編碼器（Encoder）、瓶頸層（Bottleneck Layer）

與解碼器（Decoder）組成，用於生成對抗樣本。

在訓練過程中，生成器與判別器的權重交替更新，並採用多種損失函數進行效能評估。最終，將生成的對抗樣本輸入至目標模型進行分類，記錄模型在對抗樣本上的準確率，並以圖像方式呈現對抗樣本與原始影像之比較，以驗證生成器在生成對抗樣本上的有效性。

在目標模型（Target Model）部分，本研究分別以 10 層與 4 層之卷積神經網路模型進行實驗。模型深度的設定係依據學習率（Learning Rate）與 Dropout 層參數的組合進行調整，並選擇分類準確率（Classification Accuracy）較高者作為最終模型。

其中，10 層目標模型在學習率設為 0.0002、Dropout 比例為 0.3 的情況下，達到較高的分類準確率 0.860400，比較結果如**表 1** 所示。

而 4 層目標模型在相同學習率 0.0002 與 Dropout 比例 0.3 的設定下，則達到分類準確率 0.832300，比較結果如**表 2** 所示。

**表1**. 10 層目標模型

| | Lr0.0001 | Lr0.0002 |
|---|---|---|
| 0.3 | loss: 38.408615<br>accuracy: 0.799300 | **loss: 6.987778**<br>**accuracy: 0.860400** |
| 0.5 | loss: 40.312965<br>accuracy: 0.773300 | loss: 19.942488<br>accuracy: 0.831400 |

**表2**. 4層目標模型

| | Lr0.0001 | | Lr0.0002 |
|---|---|---|---|
| 0.3 | loss: 7.960233<br>accuracy: 0.814600 | | **loss: 6.274494**<br>**accuracy: 0.832300** |
| 0.5 | loss: 14.118930<br>accuracy: 0.796400 | | loss: 9.086166<br>accuracy: 0.814300 |

在確定最佳模型層數後，本研究進一步調整學習率（Learning Rate）參數，以尋找最適化的學習速率。經多組實驗比較後，選擇分類準確率較高的模型作為最終設定。當學習率設定為 0.0003 時，模型達到較高的分類準確率 0.863500。其比較結果如**表 3** 所示。

**表3**. 10層目標模型與其學習率 (Dropout 設為 0.3)

| Lr | 0.0001 | 0.0002 | 0.0003 | 0.0004 |
|---|---|---|---|---|
| Loss | 38.408615 | 6.987778 | **6.190508** | 5.495280 |
| Accuracy | 0.799300 | 0.860400 | **0.863500** | 0.862200 |

在目標模型訓練完成後，本研究分別以 C&W、MSE 與交叉熵（Cross-Entropy）三種損失函數進行實驗，比較其對模型分類準確率的影響，並以分類準確率最低者作

為對抗攻擊效果最佳之判斷依據。實驗結果顯示，C&W 損失函數的表現較佳，其對目標模型的分類準確率分別為 0.019180 與 0.027100，比較結果如**表 4** 所示。

接著，透過設定閾值參數 C 以限制擾動（Perturbation）的大小，選擇分類準確率較低者為最佳結果。限制擾動幅度後，實驗結果顯示模型表現更佳，目標模型的分類準確率分別為 0.006420 與 0.011200，比較結果如**表 5** 所示。

在生成器（Generator）部分，透過比較不同 Dropout 層參數設定下的模型分類準確率，選擇分類準確率較低者作為最佳設定。實驗結果顯示，Dropout 比例 0.1 與 0.3 分別於不同測試中表現最佳，綜合觀察訓練過程中的損失值變化後，選擇在訓練中表現出較穩定且損失值較佳的 0.3 作為生成器 Dropout 層的最終設定。比較結果如**表 6** 所示。

此外，實驗亦顯示在擾動範圍（Disturbance Range）中，分類準確率越低表示攻擊效果越佳。結果指出，當擾動範圍設定為-0.5 至 0.5 時，能獲得更佳的對抗效果，目標模型之分類準確率分別為 0.005180 與 0.008300，其結果如**表 7** 所示。

**表4**. 損失函數

| | loss_D | loss_G_fake: | loss_perturb: | loss_adv: | Accuracy | Accuracy |
|---|---|---|---|---|---|---|
| C&W | 0.037 | 0.979 | 8.551 | 0.656 | **0.019180** | **0.027100** |
| Cross-Entropy | 0.028 | 0.966 | 22.131 | -116.145 | 0.099940 | 0.100100 |
| MSE | 0.014 | 0.976 | 36.710 | -3278.694 | 0.100000 | 0.100000 |

**表5**. 擾動大小限制

| C&W | loss_D | loss_G_fake: | loss_perturb: | loss_adv: | Accuracy | Accuracy |
|---|---|---|---|---|---|---|
| Restricted Perturbation | 0.006 | 0.976 | 8.913 | 0.360 | **0.006420** | **0.011200** |
| Unrestricted | 0.037 | 0.979 | 8.551 | 0.656 | 0.019180 | 0.027100 |

**表6**. 生成器的Dropout

| C&W | loss_D | loss_G_fake: | loss_perturb: | loss_adv: | Accuracy | Accuracy |
|---|---|---|---|---|---|---|
| 0.1 | 0.015 | 0.943 | 7.128 | 0.278 | **0.004860** | 0.018600 |
| 0.3 | **0.006** | **0.976** | 8.913 | 0.360 | 0.006420 | **0.011200** |

**表7**. 限制擾動範圍

| C&W | loss_D | loss_G_fake: | loss_perturb: | loss_adv: | Accuracy | Accuracy |
|---|---|---|---|---|---|---|
| -0.1~0.1 | 0.026 | 0.992 | 3.841 | 12.830 | 0.019180 | 0.027100 |
| -0.3~0.3 | 0.006 | 0.976 | 8.913 | 0.360 | 0.006420 | 0.011200 |
| -0.5~0.5 | 0.010 | 0.988 | 8.157 | 0.285 | **0.005180** | **0.008300** |

# 伍、結論

本研究旨在開發一個生成對抗模型來模擬惡意攻擊，以評估目標模型是否存在資安風險。研究中採用三種不同的損失函數──C&W、交叉熵（Cross-Entropy）與均方誤差（MSE）──來建構生成器與訓練流程，並於實驗中透過觀察結果不斷微調模型參數。三種損失函數的實驗結果顯示，C&W 損失函數能生成效果最佳的對抗樣本，對目標模型之分類準確率造成的降低分別為 0.019180 與 0.027100。

關於是否進一步限制擾動（perturbation）的大小，實驗顯示受限擾動的結果更佳，目標模型之分類準確率為 0.006420 與 0.011200。在生成器 Dropout 層參數的選擇上，Dropout 比例 0.3 表現最佳（分類準確率亦為 0.006420 與 0.011200）。至於擾動範圍的設定，範圍 -0.5 到 0.5 為最優，對應之目標模型分類準確率為 0.005180 與 0.008300。

綜合實驗結果可見，採用 C&W 損失函數能產生較具破壞力的對抗樣本；其他參數的最佳設定則為：限制擾動、生成器 Dropout = 0.3，以及擾動範圍為 -0.5 至 0.5（在該組合下目標模型之分類準確率為 0.005180 與 0.008300）。

本研究亦選擇在相同架構下，透過堆疊層數將網路深度由原先的四層逐步增加至十層。雖然模型複雜度提升會帶來過擬合風險，但增深網路可改善資料解析度不足所導致的問題，避免 advGAN 所產生的對抗樣本因目標模型解析度低而失效。此外，我們亦在目標模型中加入 Dropout 層以抑制過擬合，進一步提升模型的分類準確率與穩健性。

隨著深度學習（Deep Learning）技術在各領域的廣泛應用，模型的安全性已成為備受關注的議題。生成對抗網路（Generative Adversarial Networks, GAN）在產生對抗樣本（Adversarial Samples）方面展現出強大的能力，為評估目標模型的**穩健性（Robustness）**提供了一種有效的方法。

本研究旨在透過 GAN 生成對抗樣本，以評估目標模型是否存在潛在的安全風險。基於 GAN 的對抗樣本生成技術已展現出在模型安全評估上的廣泛潛力。隨著相關研究持續發展，GAN 不僅被視為攻擊工具，也將成為強化模型安全與穩定性的重要方法。

藉由生成對抗樣本，研究者能揭示模型在真實應用情境中可能面臨的潛在威脅，進而促使開發出更具防禦力的演算法與策略。未來，GAN 技術預期將被廣泛應用於各領域，不僅限於影像辨識，亦可延伸至自然語言處理（NLP）、語音辨識（Speech Recognition）以及多模態資料分析（Multimodal Data Analysis）等範疇，為不同應用場景下的模型建立更完整且嚴謹的安全評估框架，協助人工智慧系統在動態環境中維持更高的適應性與穩健性。

## 陸、引用文獻

Abou Jaoude, A. (2017). The paradigm of complex probability and Claude Shannon's information theory. *Systems Science & Control Engineering, 5*(1), 380-425.

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights, 1*(1), 100004.

Carlini, N., & Wagner, D. (2017). *Towards evaluating the robustness of neural networks.* Paper presented at the 2017 ieee symposium on security and privacy (sp).

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine, 35*(1), 53-65.

Error, M. S. (2010). Mean squared error. *MA: Springer US, 5*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011). *Adversarial machine learning.* Paper presented at the Proceedings of the 4th ACM workshop on Security and artificial intelligence.

Jamin, A., & Humeau-Heurtier, A. (2019). (Multiscale) cross-entropy methods: A review. *Entropy, 22*(1), 45.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems, 33*(12), 6999-7019.

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Annals of Data Science, 9*(2), 187-212.

Wang, S., Ko, R. K., Bai, G., Dong, N., Choi, T., & Zhang, Y. (2023). Evasion attack and defense on machine learning models in cyber-physical systems: A survey. *IEEE communications surveys & tutorials, 26*(2), 930-966.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China, 5*(23), 495.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging, 9*(4), 611-629.

Zhang, F., Chan, P. P., Biggio, B., Yeung, D. S., & Roli, F. (2015). Adversarial feature selection against evasion attacks. *IEEE transactions on cybernetics, 46*(3), 766-777.