基於卷積神經網路及對抗樣本之數位浮水印機制

組員: 周固廷 張治尹 廖晨維

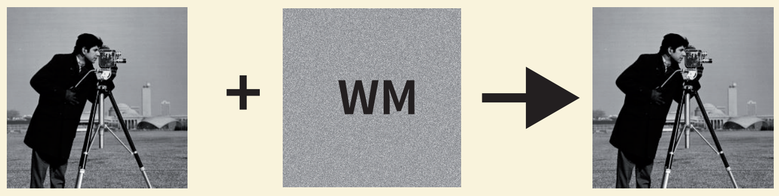
指導教授: 丁培毅 博士

摘要

本專題研究分析對抗樣本對於深度學習模型的攻擊方法，轉換為全新形式的數位浮水印嵌入與擷取方法，此方法嵌入在數位內容的浮水印是肉眼不可見的、影像處理軟體不易偵測、無法輕易移除的、且浮水印擷取的正確性極高。由於浮水印的擷取是透過深度學習的分類網路實作，網路本身對於輸入具有一定的容錯性，可以抵抗一般數位浮水印的攻擊，結合錯誤更正碼以後更具有極佳的強健性。本專題研究透過實驗分析此方法隱藏資訊的容量以及抵抗攻擊的能力。

研究動機與目的

在過去設計數位浮水印系統的研究中，絕大部分先設計一個嵌入浮水印的演算法，也許在空間域、頻率域、或是兩域來操作修改掩護圖片。

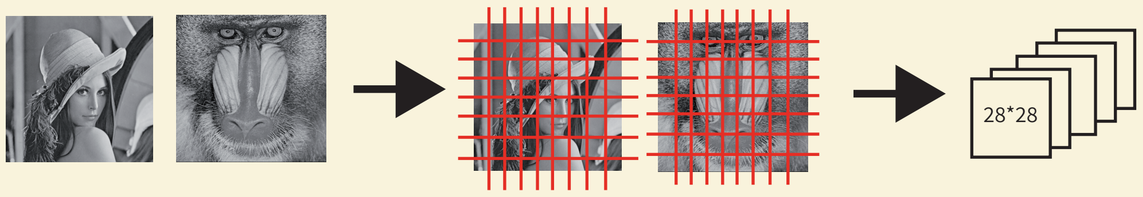


本研究分析對抗樣本的產生機制以及對系統的影響，由基於投影梯度下降(Projected Gradient Descent, PGD)之對抗樣本產生方法出發，以“調整輸入圖片改變網路輸出”作為設計的核心概念，直接訓練一個深層網路作為擷取隱藏資訊(浮水印)的工具。

研究材料與方法

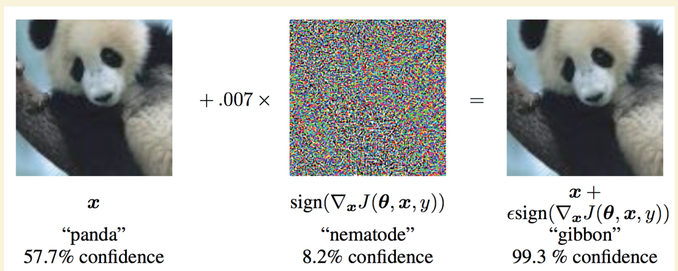
資料集

本研究使用之資料集為MNIST、FASHION　MNIST、CIFAR10、CIFAR100、及自組的CUSTOM資料集，其中CUSTOM資料集是由影像處理常用的Lenna及Baboon圖片切成的多張28x28的小圖組成。此外CIFAR10及CUSTOM資料集都經過前處理轉成灰階影像，才進行訓練與測試。



對抗樣本

透過對輸入圖片的細微修改去攻擊指定的網路，影響該網路的輸出，利用這種方式修改產生的圖片就是「對抗樣本」。研究指出所有深度學習的模型(甚至很多傳統機器學習的模型)對於這樣的攻擊都是相當脆弱的。



FGSM : I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," , 2014.

投影梯度下降攻擊方法(PGD)

PGD法運用限制性最佳化方法產生對抗樣本，雖然產生對抗樣本需要多次迭代，所需要的時間比較長，但是每一步都在指定範圍內沿著梯度方向找到損失函數上升最大的樣本，是一個廣泛使用的方式，它可以看成是迭代多次的FGSM攻擊指定模型，方法如下:

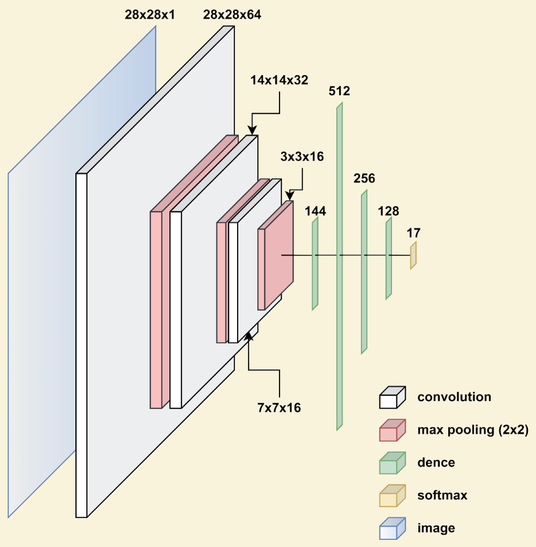
$$x^{t+1} = \prod\left(x^t + \alpha\ sgn\left(\nabla_x L(\theta, x, y)\right)\right)_{x+S}$$

浮水印擷取網路

這是本研究所提出系統中，用來擷取隱藏資訊(浮水印)的擷取網路，此「浮水印擷取網路」的架構為可用來分辨不同類別的卷積神經網路，其目的包括:

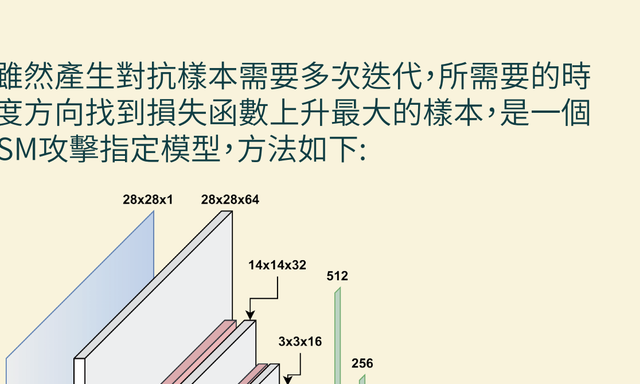
1.分辨輸入圖片中是否有嵌入浮水印

2.如果有嵌入浮水印的話，正確地解讀藏入之浮水印資訊



浮水印嵌入方法

對圖片嵌入浮水印的方式是透過PGD來攻擊浮水印擷取網路，未嵌入浮水印的圖片由浮水印擷取網路處理後輸出為類別0，假設要在圖片中嵌入四個位元的資訊，則嵌入0000四個位元時需要修改圖片使得浮水印擷取網路的輸出成為類別1，嵌入0001四個位元時需要修改圖片使得浮水印擷取網路的輸出成為類別2，以此類推，嵌入1111四個位元時需要修改圖片使得浮水印擷取網路輸出類別16



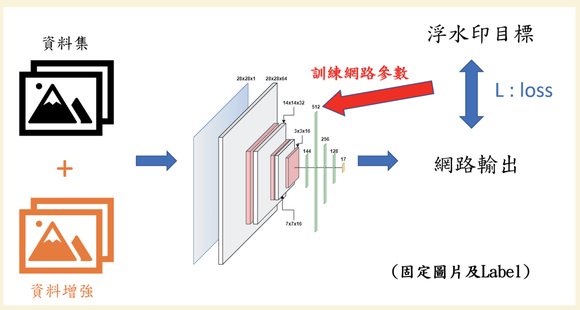
浮水印擷取網路訓練方法

1.將訓練資料集裡每一張圖片以PGD演算法沿著損失函數梯度下降方向做修改，目標是讓修改後的圖片進到浮水印擷取網路後，輸出指定的類別。

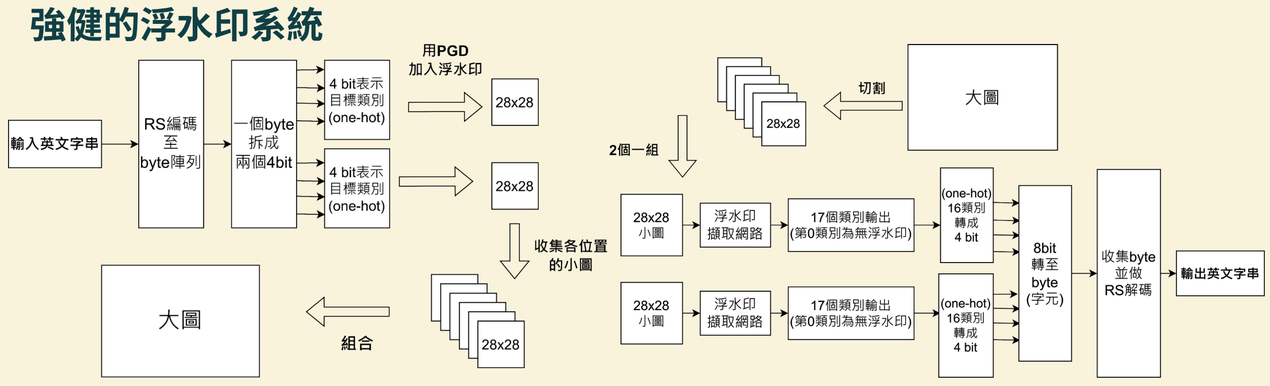
2.如右圖所示，再用資料增強的方法擴充步驟1中浮水印擷取網路各個類別的訓練資料集。並且指定監督式學習時對應的類別標籤(label)。

3.以步驟2中所得到的所有類別的訓練資料運用隨機梯度下降法訓練浮水印擷取網路的參數 。

4.重複步驟1~3，以這個演算法做50次迭代。



強健的浮水印系統



我們繼續擴充浮水印擷取網路，就可以在一張大圖中放入一個完整的字串作為浮水印資訊，上圖顯示了完整的浮水印系統架構，包含嵌入過程及擷取過程。如果浮水印資訊是原始資訊經過Reed-Solomon　code編碼的資訊，就可以更進一步強化浮水印擷取時的容錯能力，以確保這個系統的強健性。

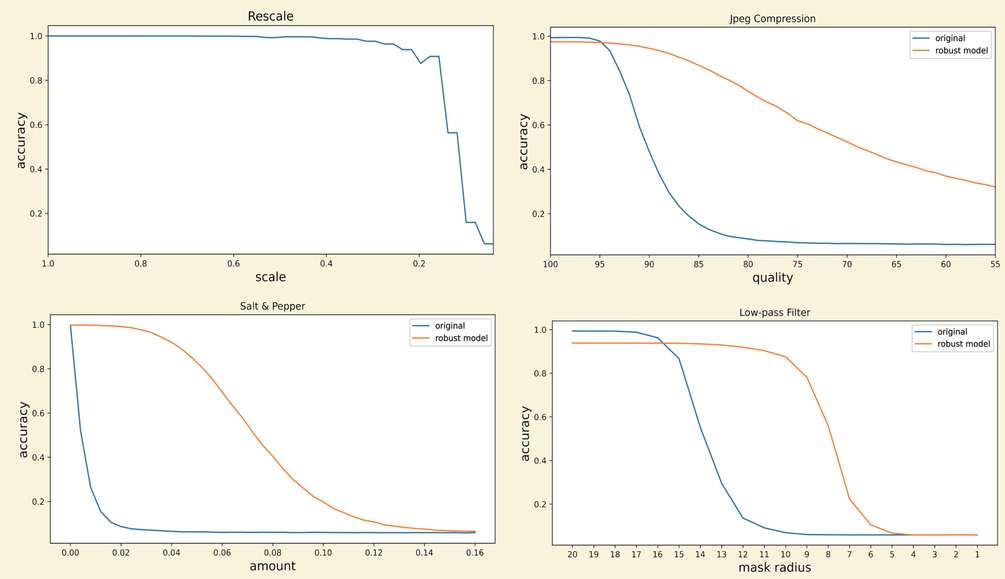
實驗結果

浮水印擷取網路在不同資料集下的訓練及測試正確率

metric	MNIST	FASHION MNIST	CIFAR 10	CUSTOM
TRAIN accuracy	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.012
TEST accuracy	1.00±0.00	0.99±0.0003	0.99±0.002	0.992±0.03
PSNR	41.47±0.76	43.21±0.64	41.70±0.28	42.50±1.432

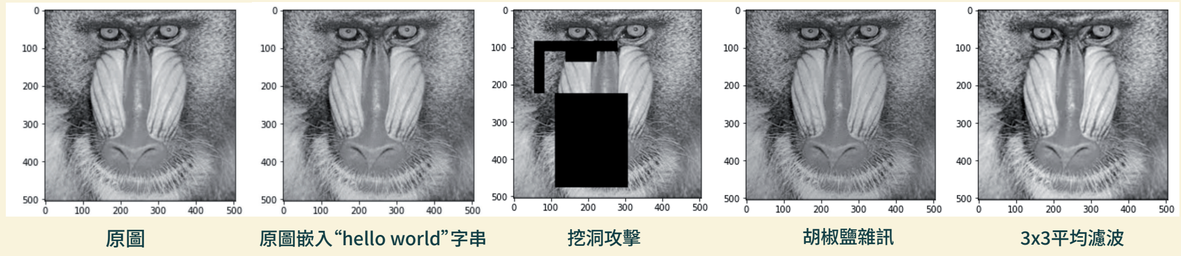
表中的TRAIN　accuracy為以各個資料集訓練後，所有訓練浮水印圖片的擷取正確率。TEST accuracy則為以各個資料集測試所得到的正確率。PSNR是各個資料集的圖片經過PGD嵌入不同資訊之後，與原圖的平均相似程度，40以上通常代表與原圖的差異無法以肉眼察覺。

在各種情況下浮水印擷取網路之正確率



在CIFAR 10資料集的圖片裡，用PGD嵌入不同資訊作為浮水印，並使用縮放、JPEG壓縮、胡椒鹽雜訊、低通濾波等攻擊來攻擊圖片，最後再測量浮水印擷取網路在不同程度攻擊下的正確率(藍線)。 使用資料增強等技術可以近一步提高浮水印擷取網路的強健性(橘線)。

大張圖片的強健浮水印系統



我們在完整的Baboon圖(512x512)中加入“hello　world”字串，加入字串以後和原圖視覺差異很小，PSNR大於40。

建構出的系統可以抵抗各種浮水印系統的攻擊，圖中分別顯示挖洞、雜訊、低通濾波等攻擊後的浮水印圖片，這些圖片都能被浮水印擷取網路正確地解碼出“hello world”字串，顯示本系統的實務應用價值。

結論

本專題研究利用產生對抗樣本的特性設計出一套全新的浮水印系統，同時訓練出一個能夠滿足正確性及容錯性要求的浮水印擷取網路

本專題研究也結合 Reed-Solomon 錯誤更正碼設計完整的浮水印嵌入與擷取系統

本專題研究成功地將對抗樣本這個對於深度學習強健性的巨大威脅轉換為正面的數位版權應用。

這些結果也點出一個很重要的設計信念 - 深度學習網路雖然容易受很小的擾動而改變輸出，如果善加運用這樣的敏感性，還是可以在許多不同地方得到正面的應用。