國立屏東大學資訊學院

資訊科學系

Department of Computer Science,

National Pingtung University

108學年度

專題研究期末報告

題目

利用深度學習模擬生成對抗網路架構之人臉影像偽造偵測研究

組員：

孫浩倫（CBE105004）

孫芷榆（CBE105044）

林佳儀（CBE105045）

指導老師：林義凱 博士

中　華　民　國　108 年 12 月

摘要

近年來由於硬體效能大幅的提升，使得深度學習受到重視。然而在技術的快速發展下，也出現許多挑戰道德議題的作品如人臉偽造系統(DeepFake)，使人們漸漸開始無法分辨現今多媒體時代資訊之真假。故本人臉偽造偵測系統將探討如何有效檢測一張人臉影像是否為偽造，並採用生成對抗網路(GAN)做為檢測的工具，期望透過 GAN 生成與對抗的機制，來協助模型可以更準確地找出確切的偽造區域。

關鍵字 ： 人臉偽造系統、人臉偽造偵測、生成對抗網路、DeepFake

目錄

圖目錄

表目錄

一. 緒論

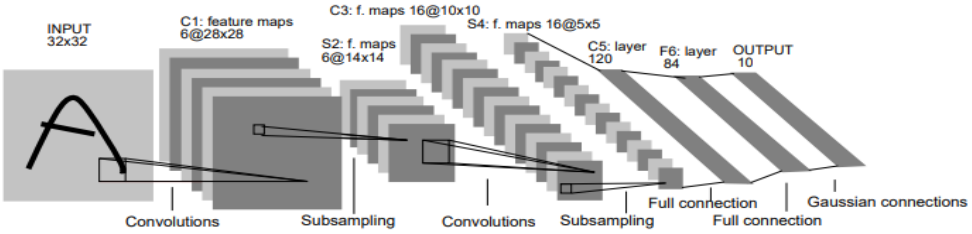
近年來在深度學習蓬勃的發展下，許多強大的演算法以及工具漸漸的被提出來。其中最具代表性的技術，無疑地就是在 2014 年由 Goodfellow 所提出的 Generative Adversarial Network(以下簡稱 GAN)[2]。GAN 的出現，大大的改變了在非監督學習(Unsupervised Learning)上的看法以及運用。舉例來說，GAN 可以用於生成各種影像，如由 Nvidia 所提出的 StyleGAN[3]即為一種可以生成不存在於這個世上的人臉的演算法；又如在資安上可利用 GANs 來設計金融防詐欺模型等。由此可知，GAN 已無所不在地應用在現實各種的日常生活中。然而隨著技術的發展，不免也會出現許多挑戰道德議題的作品被設計出來，如 DeepFake、FaceSwap、Face2Face 等人臉偽造系統的出現，讓人們漸漸開始無法分辨現今多媒體時代的資訊是真是假，使得人們對於社會的不信任感慢慢提高。故本次專題將探討如何有效的檢測出一張影像是否為偽造，並且採用 GAN 做為本系統的檢測的工具，其目的就在於希望透過 GAN 生成與對抗的機制，來協助模型可以更準確地找出確切的偽造區域。

二. 文獻探討

(1) CNN(Convolutional Neural Network)

近幾年來，卷積神經網路(以下簡稱 CNN)已經成為主流的影像處理

以及分類技術，不管是在各種的影像處理競賽、醫學影像辨識或者人流影像辨識等等，都會運用到 CNN 的技術。CNN 為一種監督式學習網路，而早在 1998 年，Yann LeCun 等學者就提出了 CNN 的基礎架構 LeNet-5(如圖一所示)[4]。在 CNN 的架構中，有可以接受類似空間相似像素值的卷積層(Convolution Layer)，與在輸出特徵圖(Output Feature Map)中，選取特定區塊最佳特徵值的池化層，而其卷積就是一個特徵擷取的方法。這兩個特性，使得 CNN 在影像辨識上具有相當大的能力。



圖一、LeNet-5 的架構圖。一個卷積神經網路[4]。

(2) GAN(Generative Adversarial Network)

GAN[2]是近幾年來最具代表性的深度學習演算法，而其中有關論文數也是逐年的上漲。GAN 是一種非監督式學習的神經網路，其中包含兩大模組，一為生成模組(稱 Generator)以及辨別模組(亦稱 Discriminator)。生成模組一開始會隨機產生雜訊(noise)，之後生成模組會將該雜訊送入辨別模組中。另一方面，辨別模組將會嘗試著辨別真圖與假圖，亦即，辨別模組會替送來的圖片打上分數，若辨別模組認為該輸入圖為假圖，則給予該圖一接近 0 之分數，反之，則給予該圖一接近 1 之分數。生成模組會依據辨別模組的結果來自我調整使得能夠產生更真實的圖片來騙

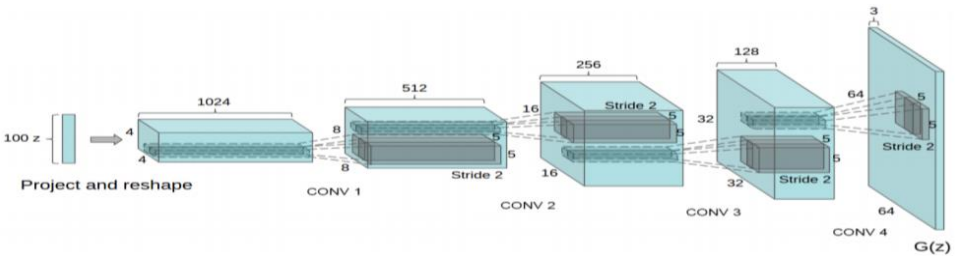
過辨別模組，而辨別模組也會自我學習而不被生成模組所欺騙，在這樣的生成與對抗之下，直到生成模組的機率分布與辨別模組的機率分布相當接近，即完成訓練。

其公式如下:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim pdata(x)}[\log D(x)] + E_{z\sim pz(z)}[\log(1 - D(G(z)))]$$

(3) DCGAN(Deep Convolutional Generative Adversarial Network)

DCGAN[7]是一種變形版的 GAN，其在 GAN 當中加入了 CNN 技術，由於 CNN 在影像處理上有著顯著的能力，故 DCGAN 可以有很好的影像生成效果。DCGAN 中的 Generator 採用反卷積的方式，將輸入的亂數值展開成一張影像，而 Discriminator 則是採用一般的卷積運算。而在本系統中即採用 DCGAN 做為主要架構來做延伸應用。



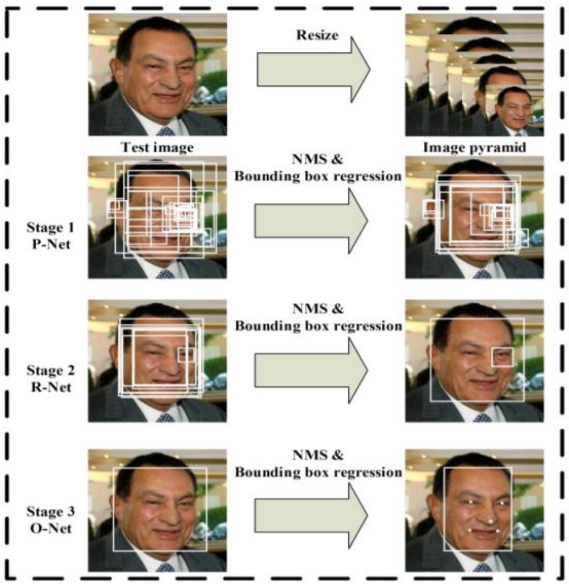
圖二、DCGAN 的架構圖[7]。

（4）MTCNN(Multi-task Cascaded Convolutional Networks)

MTCNN[9]為一種可以實現人臉偵測與對齊的工具，其回傳值包含了人臉範圍、鼻子位置、右邊嘴巴位置、右眼位置、左眼位置以及左邊嘴巴位置(如圖三所示)。

圖三、MTCNN 成果展示[9]。

MTCNN 中包含了三個主要的網路，第一個網路層 Proposal Net(P-Net)的目標為生成人臉的框架，採用了 Fully Convolutional Network(FCN)來進行偵測;第二個網路層為 Refine Net(R-Net)，其目標為對前一個網路層的結果做更進一步的偵測與判斷；第三個網路層為 Output Net(O-Net)，其目標為產生最終的 Bounding Box 以及臉部的對齊位置。而每一層網路層都會經過 Non-Maximum Suppression (NMS)以及 Bounding Box Regression 運算(如圖四所示)，使得最終可以得到最符合期望的結果。在本專題研究中，我們利用 MTCNN 將模型運算結果做人

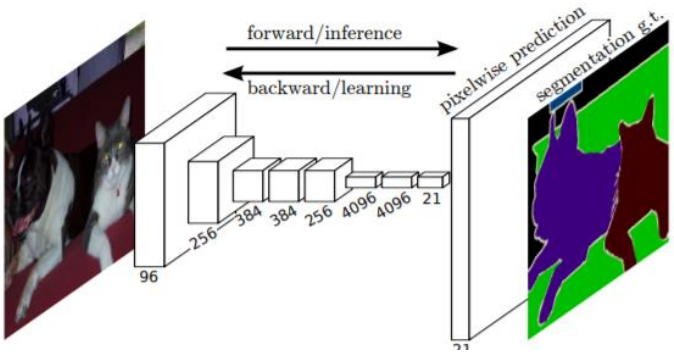


圖四、MTCNN 計算流程[9]。

臉提取，並藉由提取結果來計算偵測結果的準確率(accuracy)。

(5) FCN(Fully Convolutional Network)

全卷積網路(FCN)[6]是一個 CNN 的變形，其與 CNN 最大的差異即在於，FCN 將 CNN 的隱藏層全部替換成卷積層，故在整個 FCN 架構中，完全由卷積層建構起來。由於 FCN 可對每一個 pixel 做類別預測，故 FCN 的主要功能為影像語意分割，而相關應用則常見於道路的影像分割等等(如圖五所示)。

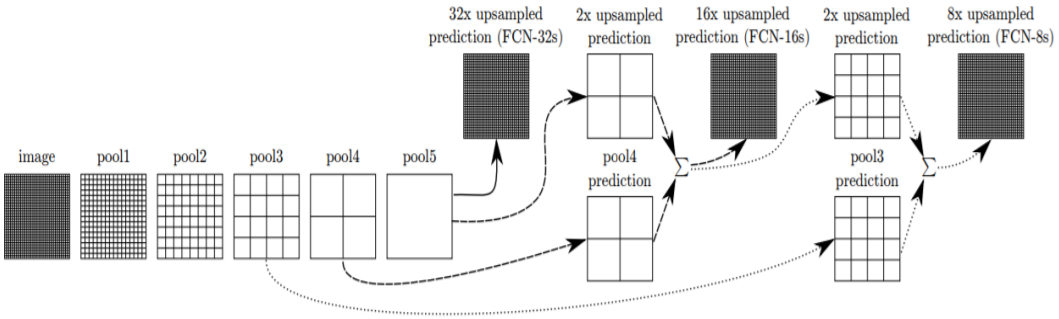


圖五、FCN 模型架構[6]。

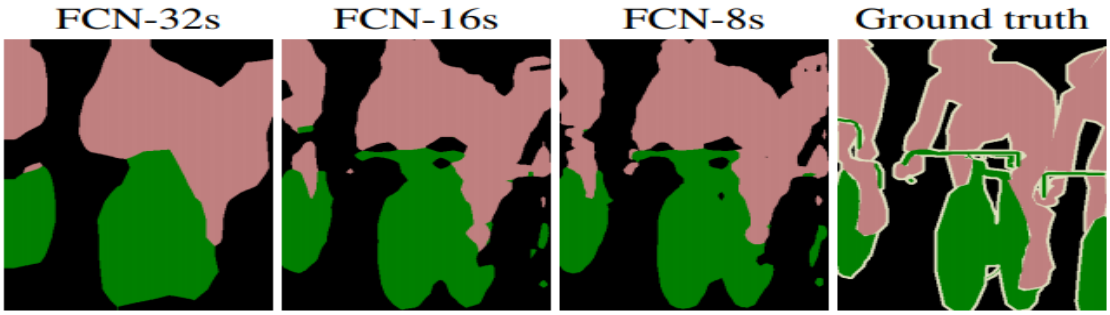
FCN 的主要架構如圖六所示，FCN 總共會經過 5 次的 max pooling 層，使得輸入的圖片縮小成原來的$\frac{1}{32}$倍(此時的影像稱為 pool5)，若直接對 pool5 的影像做向上採樣 32 倍時，得到的結果則稱為 FCN-32s；而若先對 pool5 向上採樣 2 倍並與 pool4 相加，隨後再向上採樣 16 倍的結果則稱為 FCN-16s；而若對前一次相加的結果並加上 pool3，而再向上採樣 8 倍後，其結果稱為 FCN-8s。

由圖七可知，FCN-8s 有最好的結果，而 FCN-32s 則得到最差的結

果。在本專題研究中，我們在 Generator 中實做 FCN-8s 來生成輸入影像之分數圖(簡稱 Score Map)。



圖六、3 種 FCN 架構圖[6]。



圖七、FCN 成果展示[6]。

三. 研究方法與步驟

在本系統當中，我們建構出一個神經網路模型，而此模型可以針對輸入影像之每個像素(pixel)做真假值的預測，而模型的輸出則為該圖所有像素的真假預測值，我們將此輸出稱為分數圖(簡稱 Score Map)，最後我們將利用 Score Map 的結果，透過 MTCNN 提取人臉並且計算該輸入圖之偽造之比例，再藉由我們制訂的門檻值(threshold)來判斷該圖片之真偽。以下將說明

本研究之各項方法細節以及研究步驟流程圖。

(1) 資料集的選擇

在資料集的選擇上，我們總共採取了兩個方式，第一種方式為自行製作，我們先使用了 DeepFake 這個開源的換臉程式來簡易的製作出本系統所需的資料集(如圖八所示)；第二種方式則為上網收集，我們取得了 FaceForensics++[8]這篇論文所使用的資料集，其中此資料集總共蒐集了 1000 部的影片，而每部影片都經過了 DeepFake、FaceSwap、Face2Face 這三種工具偽造，並且可以選擇想要的壓縮率。故在本次研究中，我們可以取得相當大量的資料集做訓練以及效能評估，來強化本系統的模型以及提高模型可靠度。

在資料集運用的部分，我們將取得到的資料集先依照偽造工具的種類做分類，而每一類下再分成兩大集合，第一個集合為訓練集(簡稱$D1$)，在訓練集當中，我們隨機選取原始圖集(raw)以及偽造圖集做組合，並且利用該訓練集來做模型的訓練；第二個集合為驗證集(簡稱$D2$)，其中驗證集必須與訓練集互斥，故驗證集的選取是從非訓練集當中隨機挑選原始圖集以及偽造圖集作為組合，而驗證集即為檢驗我們的模型好壞時判斷的依據，倘若在驗證集能有相當好的偵測結果，則認定這個模型是好的，反之，則這個模型是不好的。

這些資料集也提供了偽造的遮罩區(mask)，而這些遮罩上記錄著該圖片的任一個像素是否經過偽造，若經過偽造的區塊則會用不同的方式或顏色來標示，而未被偽造的區塊則顯示為黑色。當我們讀入這些遮罩時，我們會將未經偽造的區塊標記成 0，而偽造的區塊標記成 1，並且使用獨熱編碼(one-hot encoding)，使得後續模型可以正確的利用這些標籤(label)做訓練。

圖八、使用 DeepFake 所製造的訓練集。

(2) 神經網路的模型

本系統中以 DCGAN 做為整體的模型架構。由於 DCGAN 中包含了 Generator 以及 Discriminator 兩個網路，故在訓練時，本系統採用先訓練 Discriminator 三次，後訓練 Generator 一次，使得兩種網路能夠相互抗衡。以下將個別介紹這兩個網路的模型以及訓練細節。

甲、 Discriminator

Discriminator 是以 CNN 做為基本架構的網路，其輸入可為$x \circ x'$或者$x \circ x''$。其中 Discriminator 會依據我們所設定的批次大小(batch size)，從X(data space，簡稱X)中抓取在這一次訓練 Discriminator 所需的圖片集X'，而x為X'中的一張圖片，x'為x經過 Generator 運算後的輸出、x''為 x 之正確答案的 Score Map。

Discriminator 的工作在於當對 Discriminator 輸入$x \circ x'$時，Discriminator 可以透過卷積運算並輸出一個 logits(簡稱L)，這個 logits 需要愈接近 0 愈好，象徵著 Generator 產生出來的x'還不夠精確，使得 Discriminator 認定這是一個不好的結果。相反的，當對 Discriminator 輸入$x \circ x''$時，由於夾帶正確答案x''，故 Discriminator 會透過卷積運算得到一個 logits(簡稱L')且須讓此 logits 接近 1，即代表這是一個精確的結果。即$\mathrm{D}(x, x') = L$ 或 $\mathrm{D}(x, x'') = L'$。

乙、 Generator

Generator 是一個由 VGG16 FCN-8s 為主所建構出來的神經網路，其以一圖片集(data space，簡稱X)做為輸入，這些圖片X摻雜著原始圖與偽造圖，而 Generator 會依據我們所設定的批次大小(batch size)，從X中抓取在這一次訓練 Generator 所需的圖片集(簡稱X')，而 Generator 會針

對X'中的每個圖片(簡稱x)去做卷積與反卷積的運算，並在反卷積後對圖上的每個像素做真偽的分析。亦即，Generator 會對圖上的每個像素打上一個分數(其分數應介於 0~1 之間，$[0, 1]$)。而最終 Generator 會輸出一張標滿分數的分數圖(簡稱 Score Map，x')，即$\mathrm{G}(x) = x'$。

而 Generator 的工作在於其需要產生可以騙過 Discriminator 的 Score Map。亦即，由 Generator 所產生的x'在進入 Discriminator 中做運算後，其產生之 logits 應要接近 1，代表 Discriminator 認為這是一個接近標準答案的結果。

(3) Loss function 的計算

在 loss function 的計算當中，由於 Generator 以及 Discriminator 兩個神經網路都需要計算 logits 與期望值的差距，故本系統採用了 sigmoid cross entropy with logits 交叉熵公式來計算所需結果。其公式如下:

$$L(x, z) = z * -\log\big(sigmoid(x)\big) + (1 - z) * -\log\,(1 - sigmoid(x))$$

其中x為輸入之 logits，z為標籤。

而 Discriminator 的 loss function 可定義為:

$$L_D(x, y) = L\big(D\big(x, G(x)\big), 0\big) + L(D(x, y), 1)$$

其中x為輸入之圖片，y為x正確的 Score Map。

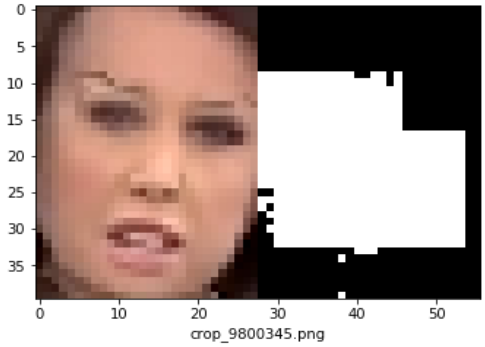
而 Generator 的 loss function 可定義為:

$$L_G(x,y) = L(G(x),y) + L\big(D\big(x,G(x)\big),1\big)$$

x為輸入之圖片，y為x正確的Score Map。本研究中與原始的DCGAN比較不同的是，在Generator中的loss中本系統多加入了$L(G(x),y)$這個交叉熵。亦即，透過$G(x)$ 以及y去做運算後，其結果可以幫助Generator朝向正確的方向做輸出，使得不會出現無法預期的結果。而原先的$L\big(D\big(x,G(x)\big),1\big)$則是透過與Discriminator的對抗中，協助產生出更好的結果。

(4) 模型的評估

在模型的評估上，本系統藉由計算模型在驗證集上的Accuracy來判斷這個模型是否符合期待。而在計算Accuracy前，本系統會先利用MTCNN將人臉給提取出來，並且透過MTCNN回傳的座標值，將預測的Score Map切割出與提取結果符合之大小，最後再經由計算偽造的區域佔人臉的百分比，並檢測此百分比是否超過本系統所制定的門檻值(threshold)，若超過則判斷此圖為假圖，否則即為原始圖(如圖九所示)。

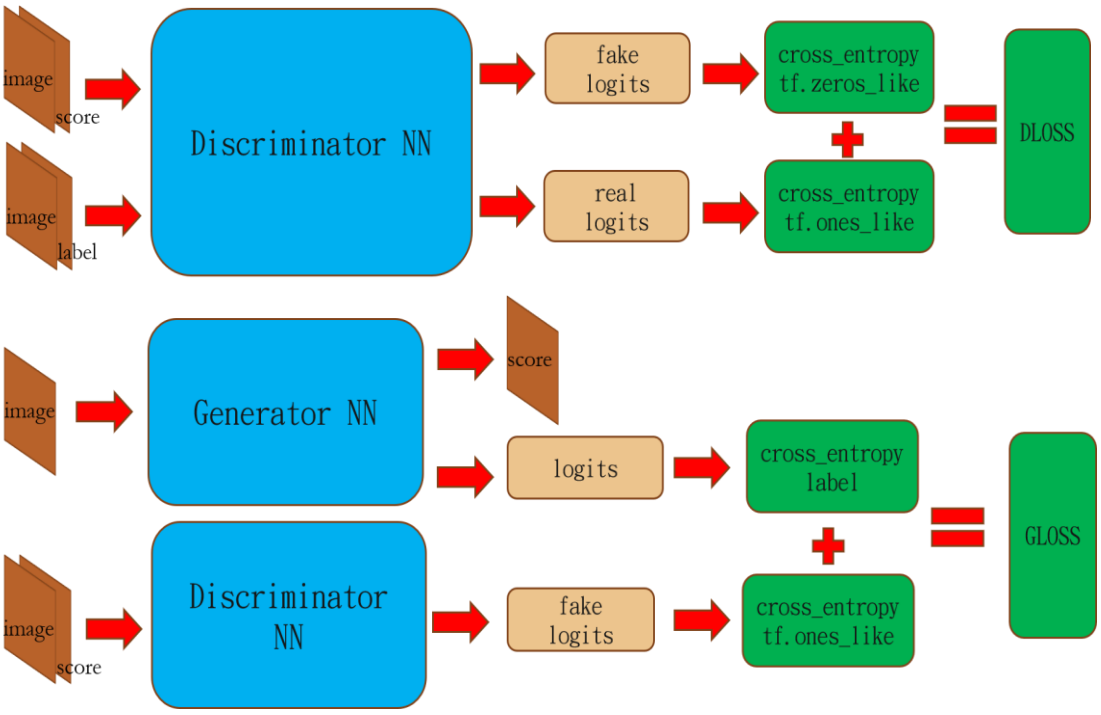


圖九、模型偵測結果。

(5) 程式流程圖

本專題研究之程式主要流程如圖十所示，而兩個神經網路之流程圖如圖十一所示。

圖十、程式流程圖

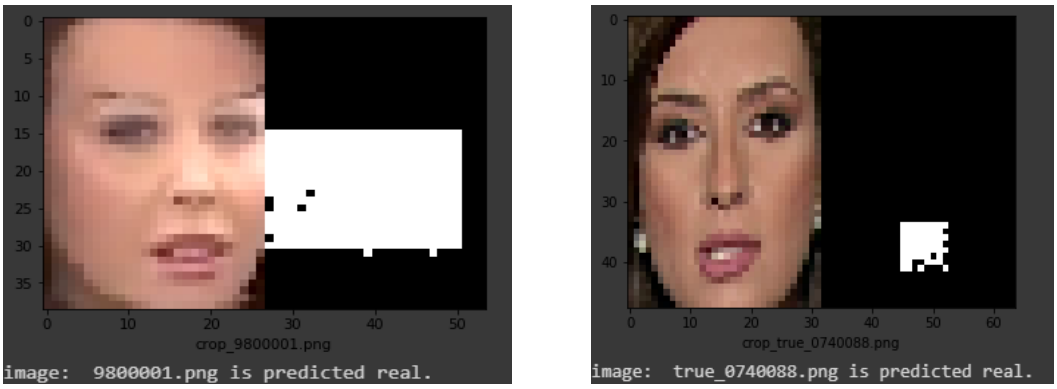


圖十一、神經網路流程圖。

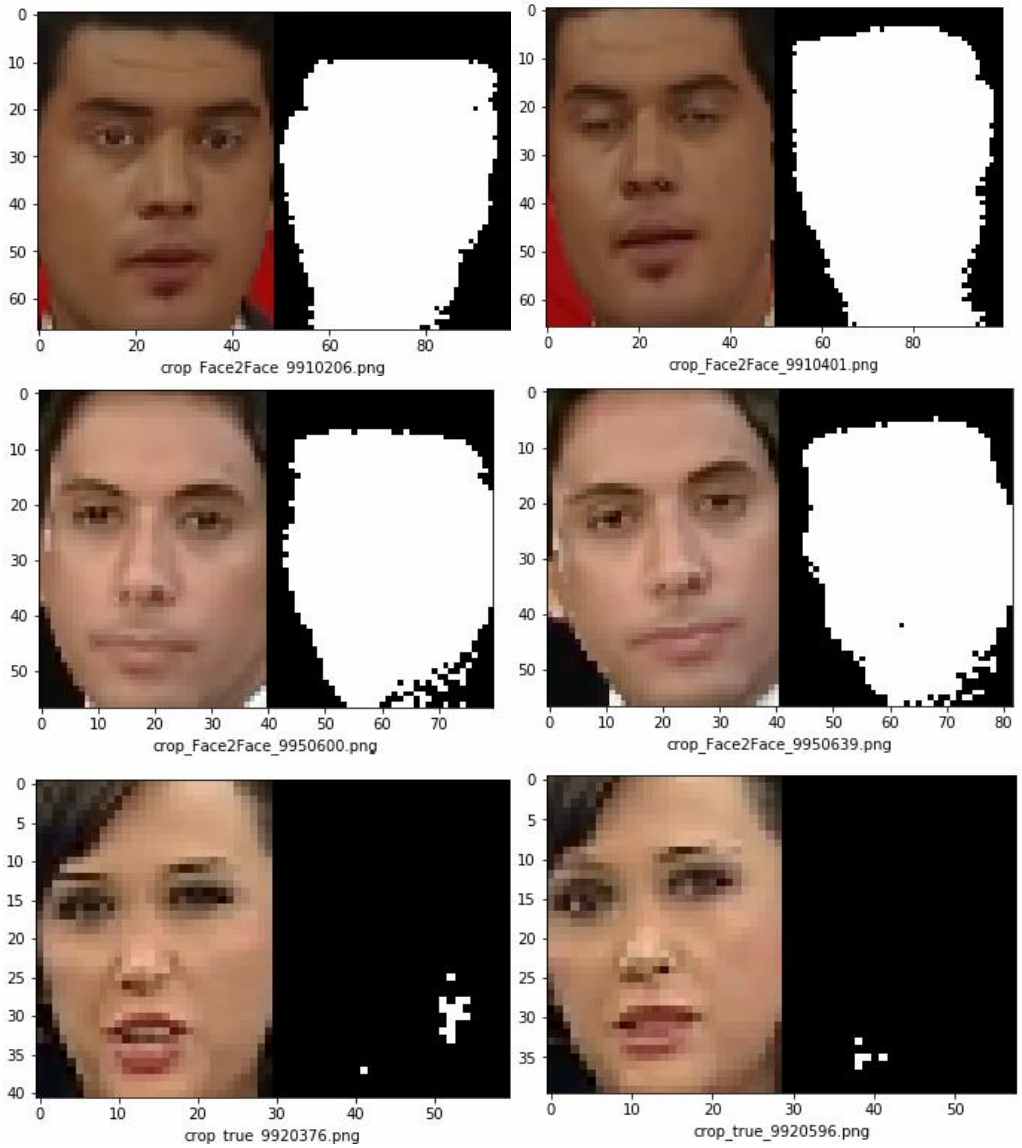
四. 研究成果

在本專題研究完成時，由於受限於資料集的取得、訓練時間的長短以及硬體運算速度等等因素，故目前僅初略完成 Deepfake 以及 Face2Face 的測試。故以下內容將以以上兩種的偽造集以及訓練結果進行介紹。

在本次實驗中，我們總共完成了兩個測試。首先，我們在 DeepFake 資料集上總共取用了 2000 多張的影像做為訓練集，而約 700 多張的影像做為驗證集；在 Face2Face 資料集上取用了 30000 多張的訓練集以及 1700 多張的驗證集，其中兩者之訓練集以及驗證集都混合了原始圖以及偽造圖。最終在測試下，兩種偽造方式在本系統都能獲得九成以上的辨識結果(見圖十二至圖十三、表二至表四)。圖十二以及圖十三分別為本模型在 DeepFake 以及 Face2Face 部分之驗證集上的偵測成果，其中圖的左半部為透過 MTCNN 偵測出的人臉區塊，右半部為透過 MTCNN 預測之人臉範圍所切割出之 Score Map。由圖十二可以得知，若白色的區塊愈多，則代表模型認為偽造的區塊較多。而在門檻值的部分，我們將其設定為 40%。亦即，若模型預測之偽造區域佔了人臉大於 40%的比例，則認定該圖為假圖，反之則為原始圖。



圖十二、DeepFake 驗證集上的成果。



圖十三、Face2Face 驗證集上的成果。

而相關的訓練參數如下表一所示。

表一、訓練相關參數。

影像尺寸	學習率	偽造門檻值	Discriminator ： Generator 訓練次數
256x192	0. 0001	40%	3 ： 1

表二、DeepFake 驗證集測試結果。

訓練集大小	驗證集大小	Batch size	Epoch	Accuracy
2391	728	64	270	0.971

表三、Face2Face 驗證集測試結果。

訓練集大小	驗證集大小	Batch size	Epoch	Accuracy
30020	1795	10	80	0.934

表四、DeepFake 驗證集之其他測試結果。

Batch Size	Epoch	Accuracy
8	**900**	**0. 960**
8	600	0. 915
8	550	0. 926
8	500	0. 595
8	400	0. 473
8	200	0. 172
10	200	0. 848
64	300	0. 446
64	280	0. 628
64	**270**	**0. 971**
64	250	0. 886
64	200	0. 815
90	200	0. 474

五. 問題探討

(1) Discriminator 的深度會影響 Generator 收斂

我們在測試的時候發現，倘若對 Discriminator 設計較多層的卷積層

時，會出現 Generator 遲遲無法收斂的情況，也就是 Generator 的梯度永遠無法下降，使得本系統無法得到預期的結果。目前猜測的主要原因為當 Discriminator 有太好的辨別能力時，則 Discriminator 可以很輕鬆的辨識出何者為Generator所產生出之結果，亦即，Discriminator會將Generator所送來的結果全部當成不好的結果，如此一來，Generator 則無法透過 Discriminator 的辨識結果來使自己的梯度下降。故在設計 Discriminator 時，也應該考慮何種的模型才不會造成上述問題的發生。

(2) 門檻值的設定會影響測試結果

在本次研究中，本系統之偽造門檻值設定在 40%左右，40%是目前相較穩定且較合理的比例值，然而在不同的門檻值之下可能造成不一樣的結果出現，如以下整理所示:

甲、 較低的門檻值:

較低的門檻值會使得偵測的條件放寬。亦即，當門檻值較小時，將會增加判別為假圖的機率。在這種情況下，有機會可以提高準確度，然而若設定出不當的門檻值時，很可能會因為些許的誤差使得原始圖也被偵測為假圖，進而造成準確度降低。

乙、 較高的門檻值:

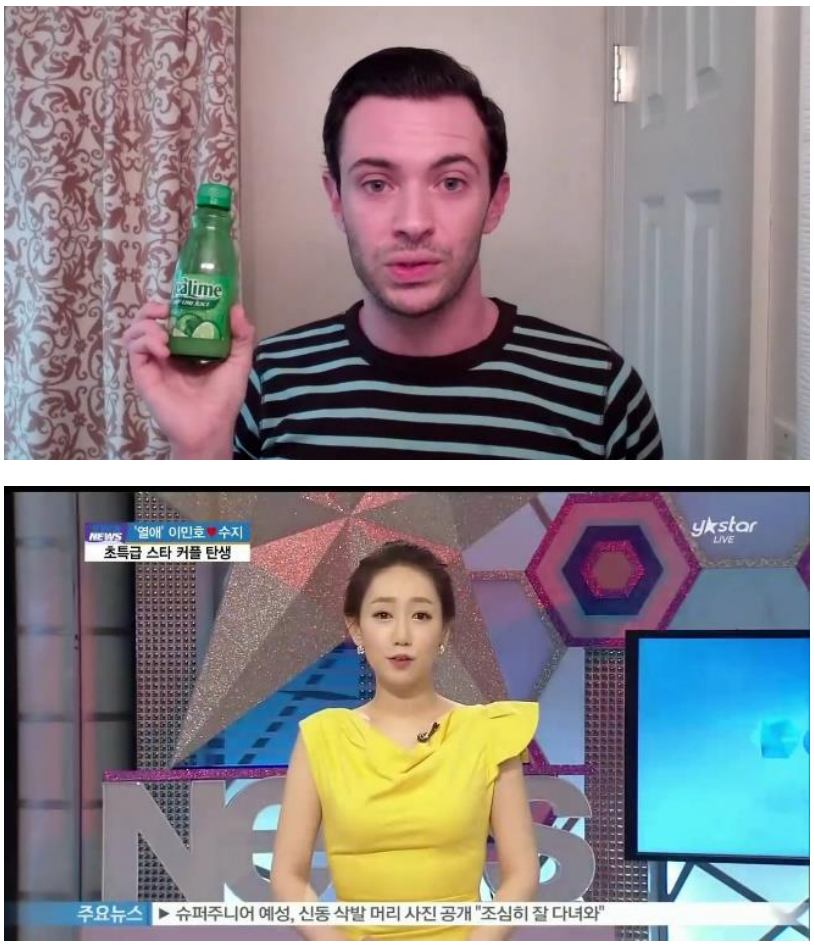
較高的門檻值會使得偵測的條件變得更嚴苛。亦即，當門檻值較大

時，將會降低辨別為假圖的機率。在這種情況下，準確率會因為門檻的提高而下降。所以若要設定較高的門檻值，則必須同時加強模型的辨識能力，才能在模型與門檻值的相互合作下，不會因此而失衡，並且可以有效的提高使用者對於模型的信心度。

(3) 影像的輸入尺寸

在深度學習的訓練上，硬體的規格是一個相當大的限制，倘若設定較大的影像尺寸時，很可能會同時造成主記憶體以及顯示卡的記憶體不足，使得訓練時崩潰；然而較小的影像尺寸可能會造成影像的特徵消失，造成模型無法取得良好的影像特徵。在我們取得的 FaceForensics++資料集中，由於每部影片的解析度、人臉的位置(如圖十四所示)以及人臉佔影像之比例(如圖十五所示)皆有差異，故如何調整合適的影像大小做為訓練將會是一大挑戰。而在未來我們將嘗試各種自動化的方式，如自動切出合適的人型大小使背景的比例縮小，如此一來便能降低縮小的比例。

圖十四、人臉的位置並非在正中。



圖十五、由圖可得知，儘管兩者影像解析度相同，但人臉佔影像比例卻有很大的不同。

六. 結論與建議

隨著硬體的快速進步，同時也帶動了深度學習的快速發展。然而在這些技術的發展之下，也不停的在挑戰我們人類的道德議題[5]。現今很多學者們也不停的在研究如何與這類的偽造程式做抗衡。然而道高一尺，魔高一丈，儘管學者們很努力的製作偵測模組，也趕不上有心人士製作偽造圖片的速度，而目前的所有偵測工具也無法有效的偵測所有的偽造圖集。在未來的發展上，

偵測與偽造仍舊會是一個強烈的對抗，而如何製造出一個全能的偵測系統，想必也是深度學習在資訊安全上一個重要的技術發展。

在本專題研究當中，我們深刻的體會到本系統仍然有許多需要改進的地方，以下列出三點未來本系統可以改進以及研究的方向。

甲、 加大訓練集以及驗證集:

FaceForensics++的資料集總共提供了高達 1000 部的影片供使用，故在未來本系統將會更有效的利用這 1000 部影片，並且針對三種偽造技術(DeepFake、FaceSwap、Face2Face)持續研究，使得本系統的模型可以有更好的訓練資料以提升本系統模型的能力。

乙、 將三種偽造方式(DeepFake、FaceSwap、Face2Face)的資料集做混合:

由於目前大多數的偵測方式都是個別做訓練，故若在偵測前未能得知輸入的影像為何種偽造方式，則可能因為使用錯的權重而使得模型無法有效偵測影像是否為偽造。故在本系統的未來研究中，將嘗試將三種資料集做混合，如此一來我們僅需訓練出一個權重，則可以用於偵測三種偽造方式且不需事先得知輸入之影像之偽造方式，而這種偵測方法較貼近現實的日常生活。

丙、加強 Generator 網路的能力:

在此階段中，本系統將研究其他神經網路模型在本系統上的運作能力，並且透過測試，將最好的模型納入系統中。其中可以嘗試的神

經網路包含了 SegNet[1]、PSPNet[10]等。

七. 參考文獻

[1] Badrinarayanan, V., Kendall, A., Cipolla, R.:SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. arXiv preprint arXiv:1511.00561v3(2016)

[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.:Generative Adversarial Networks. arXiv preprint arXiv:1406.2661v1(2014)

[3] Karras, T., Laine, S., Aila, T.:A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv preprint arXiv:1812.04948v3(2019)

[4] LeCun, Y., Bottou, L., Bengio, Y., Haffiner, P.: Gradient-Based Learning Applied to Document Recognition. In: Proc. Of The IEEE. (November 1998)

[5] Lee, D.: Deepfakes porn has serious consequences. In: BBC News.(February 2018). https://www.bbc.com/news/technology-42912529

[6] Long, J., Shelhamer, E., Darrell, T.:Fully Convolutional Networks

for Semantic Segmentation. arXiv preprint arXiv:1411.4038v2(2015)

[7] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434v2(2016)

[8] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv preprint arXiv:1901.08971v3(2019)

[9] Zhang, K., Zhang, Z., Li, Z., Qiao, Y:Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. arXiv preprint arXiv: 1604.02878v1(2016)

[10] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.:Pyramid Scene Parsing Network. arXiv preprint arXiv:1612.07705v2(2017)