基於遠程光電容積脈搏波特徵及機器學習的深度偽造影片偵測方法

Video Deepfake Detection using Remote Photoplethysmography Feature and Machine Learning

專題組員：呂霈芳、謝采軒、林翊婷、施巧軒

專題編號: PRJ-NTPUCSIE-111-004

執行期間：111 年 06 月 至 112 年 06 月

一、摘要

我們從影片的臉部抓取數學及圖像特徵，來訓練模型以達到判斷影片真偽的結果。以數學特徵作為模型輸入的部分使用了三種模型： Support Vector Regression (SVR)、Support Vector Classification (SVC)以及 Multilayer Perceptron (MLP)；而以圖像特徵作為輸入的部分則是使用了 Convolutional Neural Network (CNN)、 Long Short-Term Memory (LSTM)以及 EfficientNet。我們採用了著名的 deepfake 資料集 Deepfake Detection Challenge(DFDC)、Celeb-DF 和 FaceForensics++當作訓練及測試資料。實做時將影片數量切成五等份，四份為訓練資料，一份為測試資料，將每等份輪流當作測試資料的結果記錄下來後計算其平均準確率。

二、簡介

隨著人工智慧和機器學習技術的進步，製作高度逼真的虛假內容變得相對容易。這種技術在開放的網絡平台上迅速流行，引起了虛假資訊和侵犯隱私權的嚴重問題，也由此技術帶動了 deepfake 偵測工具快速的崛起。

最初我們在決定題目時，正巧碰上網紅小玉事件的發酵讓台灣開始重視 deepfake 的問題。除了小玉事件之外，還有許多有關於 deepfake 濫用的新聞，例如歐巴馬的影片被重製成辱罵川普的假新聞，也對社會造成了重大影響。

此外儘管目前已經有付費的 deepfake 偵測服務，但公開的 deepfake 偵測技術一直都沒有一個較好的解決方案，如著名的 deepfake 偵測比賽 Deepfake Detection Challenge(DFDC)第一名的模型在黑箱測資的表現也只有 65.18%的準確率，所以我們最終決定做 deepfake 偵測這個題目。

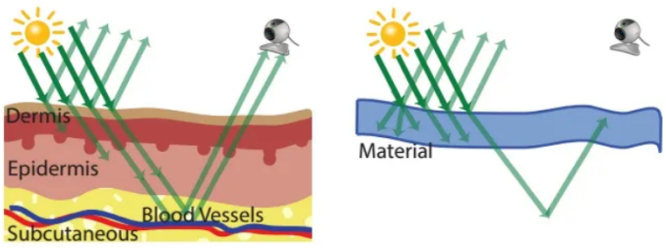
三、 專題進行方式

1.技術背景

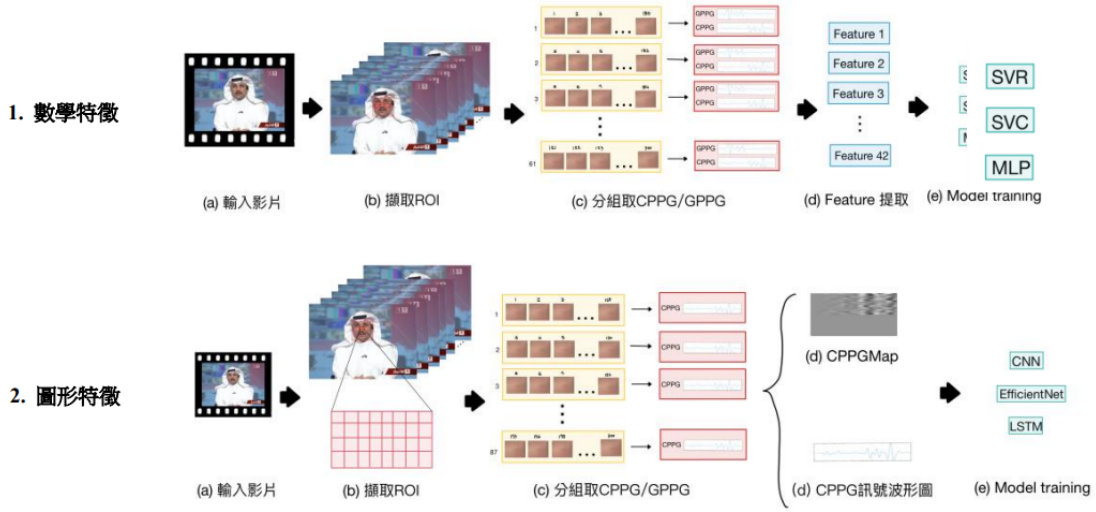
在本專題中，我們採用人體特徵[1]作為偵測影片真偽的主要依據。人體特徵是利用人的呼吸、心跳或者是面部微血管，透過光反射出來的膚色等是否連續來判斷，經過調查後發現這些生物特徵是較難被仿造的。我們使用的生物偵測方法是 remote Photoplethysmography(rPPG)[2]領域中的 Chrominance-based rPPG(CPPG)及 GRGB rPPG(GPPG)。

rPPG 是使用反射的光線來測量因心臟跳動造成的皮膚細微亮度變化的技術，可以透過 rPPG 訊號去預測心率、呼吸等生理特徵。

我們是利用影像中真實臉及假造臉反射訊號不同來進行 deepfake 影像分類。影像中的真臉會有環境光穿過皮膚到達血管，然後反射回相機，因此相機可以檢測到血液流動導致的亮度變化；而假臉由於材質不同，吸收以及反射回來的資訊就會與真臉的訊號差異大，原理如圖一所示。



圖一：rPPG 檢測的原理示意圖



圖二：整體架構圖

根據研究顯示，對整個影像提取 rPPG 的雜訊比較高，因此我們使用 CPPG 跟 GPPG 取代 rPPG。CPPG 是單獨提取影像中 HVC 顏色系統中彩度通道(chroma channel)rPPG，GPPG 則是提取 RGB 色彩系統綠色通道的 rPPG。

2.系統架構

我們的架構是以 intel 的 FakeCatcher[3]為原型進行變化的，架構圖如圖二，根據模型的輸入可將架構分成兩部分，分別是用訊號的數學特徵以及用訊號的圖像特徵去進行訓練。兩邊流程分別會先進行影片的輸入、擷取所需 ROI、取 ROI 的 CPPG/GPPG 數值、進行 feature/CPPGMap/CPPG 訊號波型圖的提取，最後採用機器學習或深度學習的模型訓練及偵測是否為深偽影片。機器學習方法包含 SVM（Support Vector Machine)、 SVC (Support Vector Classifier)、SVR (Support Vector Regression)，深度學習模型包括 CNN(Convolutional Neural Network)、EfficientNet[4]、LSTM(Long Short-Term Memory)，各細部流程說明如後。

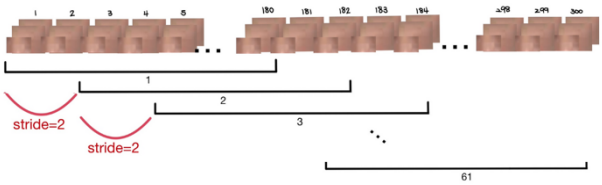
2-1. 數學特徵

整體架構輸入的影片固定每秒 30 幀，輸入後使用 Google 開發的人臉檢測框架 mediapipe[5]抓取臉部位置，並用它標出來的 landmark 定位及切割每幀中人臉的左臉頰、右臉頰及面中 ROI(圖三)，接著對所有 ROI 進行分組提取 CPPG 及 GPPG

圖三：擷取臉部 ROI

訊號。

舉例來說一個 10s 的影片會有 300 組左臉、右臉及面中 ROI，我們會以 180 幀為一組，每組之間間隔 2 幀。因此第一組為編號 1~180 的 ROI、第二組為編號 3~182 的 ROI，以此類推。接著對每組的三種 ROI 分別進行 CPPG 以及 GPPG 的提取，一組會生成 6 個訊號，整部影片就會有 61*6 個訊號(圖解請見圖四)。



圖四：將 ROI 分組

再來會對每組的 6 個訊號進行 42 種特徵計算，如表一所示，我們將 feature 分為四大組，分別為 F1、F3、F4 與 Last，其中有*標記 9 個 feature

的為結果分類準確率較高的 feature，個別的介紹如下：

A. GPPG narrow pulse first (3)

計算 GPPG 訊號 spectral correlation 的最大值。

B. GPPG narrow pulse second (4)

計算 GPPG 訊號 spectral correlation 的第二大值。

C. GPPG STD (15)

計算 GPPG 訊號的標準差。

D. GPPG mean std of differences (19)

先計算一秒 30 幀的 GPPG 訊號標準差，再位移 1 幀後計算標準差直到第 180 幀，最後計算所有標準差的平均值。

E. GPPG RMS STD (20)

計算一秒內 GPPG 訊號的 root mean square，每次位移 1 秒直到第 180 幀，最後計算所有 root mean square 的平均值。

F. CPPG std of mean values of 1 sec windows (24)

計算 CPPG 訊號每 30 幀的平均值，每位移一幀之後算一次平均值，最後取所有平均值的標準差。

G. CPPG mean std of differences (26)

作法與 D 相同，僅差在輸入訊號為 CPPG。

H. CPPG RMS STD (28)

作法與 E 相同，僅差在輸入訊號為 CPPG。

I. CPPG Shannon entropy (30)

計算 CPPG 訊號的 Shannon entropy。

最後是機器學習模型的訓練。這部分測試的模型有 SVR、SVC 及 MLP，主要著重於 SVR 跟 SVC。不同模型參數的挑選主要是用 sequential forward selection 的方法挑選輸入的特徵。目前這部分準確率最高是由輸入為 9 個特徵的 SVC 測出來的 76%準確率。

2-2. 圖像特徵

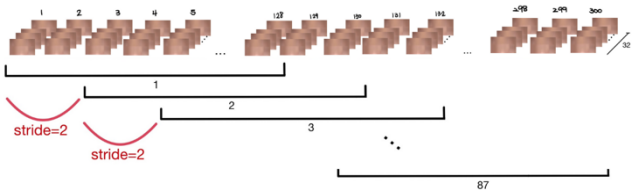
以圖像特徵作為輸入的前半部跟數學特徵的部份的步驟相似，只是這次只提取面中 ROI(圖五)，且將此 ROI 切成 32 等分去做分組。

Feature編號	組別	Feature名稱
1	F1	meanCsd
2	F1	maxCsd
*3	F3	GPPG narrow pulse first
*4	F3	GPPG narrow pulse second
5	F3	GPPG avg energy
6	F3	GPPG max spectral autocorrelation
7	F3	CPPG narrow pulse first
8	F3	CPPG narrow pulse second
9	F3	CPPG avg energy
10	F3	CPPG max spectral autocorrelation
11	F3	CSD narrow pulse first
12	F3	CSD narrow pulse second
13	F3	CSD avg energy
14	F3	CSD max spectral autocorrelation
*15	F4	GPPG STD
16	F4	GPPG std of mean values of 1 sec windows
17	F4	GPPG std of differnces
18	F4	GPPG RMS
*19	F4	GPPG mean std of differnces
*20	F4	GPPG RMS STD
21	F4	GPPG mean of autocorrelation
22	F4	GPPG shannon entropy
23	F4	CPPG STD
*24	F4	CPPG std of mean values of 1 sec windows
25	F4	CPPG std of differnces
*26	F4	CPPG mean std of differnces
27	F4	CPPG RMS
*28	F4	CPPG RMS STD
29	F4	CPPG mean of autocorrelation
*30	F4	CPPG shannon entropy
31	F4	CSD STD
32	F4	CSD std of mean values of 1 sec windows
33	F4	CSD std of differnces
34	F4	CSD mean std of differnces
35	F4	CSD RMS
36	F4	CSD RMS STD
37	F4	CSD mean of autocorrelation
38	F4	CSD shannon entropy
39	Last	GPPG auto correlation mean
40	Last	GPPG auto correlation max
41	Last	CPPG auto correlation mean
42	Last	CPPG auto correlation max

表一：42 種 features[6][7]

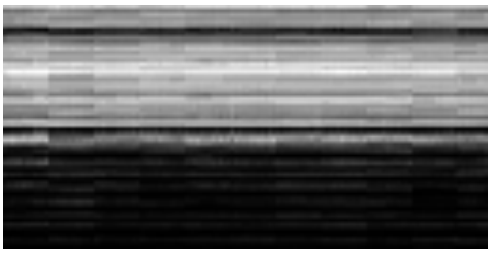
圖五：擷取臉部 ROI

這次我們以 128 幀為一組，每組之間間隔 2 幀，編號 1~128 為一組，編號 3~130 為一組，然後對每個小組提取 CPPG 訊號，一個小組會產生 32 個訊號，一部影片就會有 87x32 個訊號(圖解請見圖六)。



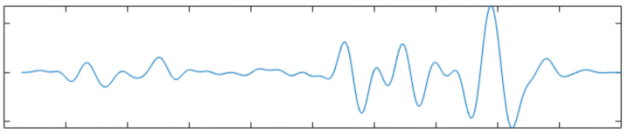
圖六：將 ROI 分組

最後是 model training 使用的模型，這個階段採用的有一般的 CNN、EfficientNet 及 LSTM。我們主要使用的是有三層卷積層及一層 Dropout 的 CNN 及 EfficientNet。模型輸入的部分主要分為兩種，CPPG 訊號波型圖及 CPPGMap。CPPG 訊號波形圖即是將得到的 CPPG 訊號畫成圖片表示。CPPGMap 則是將每小塊的 CPPG 訊號，與套上功率譜密度計算的 CPPG 訊號結果合併，一起 normalize 到 0-255 的區間成為一個 CPPGMap(圖七)。



圖七：CPPGMap

其實原論文中只有使用 CPPGMap 作為輸入，但我們無論換哪種 CNN 架構都很容易過擬合，在 EfficientNet 的表現也不是很好，所以直接改用面中的 CPPG 訊號波形圖(圖八)去訓練，所幸最後的結果也提高到七成，為輸入 CPPG 訊號波形圖的 EfficientNet，準確率為 73%。



圖八：CPPG 訊號波形圖

3.困難與解決方式

3-1. 資料集篩選

部份我們蒐集到的資料集真假影片的比例不均，例如我們能成功提取出 ROI 的 DFDC 資料集真影片原本有 70 部，但假影片卻有 278 部，因此訓練出來的模型就會偏向將影片分類為假影片。最後是採取自行尋找影片及合併資料集的方式來解決這個問題。這樣雖然能成功將真假影片的比例變得更加均衡，但不同資料集製作出的 deepfake 影片品質不一，也可能會造成其他影響。

再來是資料集中的影片每秒的幀數不固定，例如 FaceForensics++的影片大多是 25fps，而其他資料集如 DFDC 跟 Celeb-DF 就是 30fps，由於我們是將影片分組取 rPPG 的，這樣不同幀數的影片取得的資料量就會差很多。針對這個問題我們就固定只使用每秒 30 幀的影片作為輸入。

3-2. 影像預處理部分

原本使用的臉部定位程式抓取速度非常慢，在嘗試了多種方法後，最後改用了能 realtime 抓取的 mediapipe。而對於影片的限制，無論是使用何種臉部定位方式，在光線不佳、一人以上、側臉為主及臉部過小的影片，還是難以定位臉部位置。

3-3. 模型部分

剛開始我們訓練出來的數學特徵模型準確率一直停在 60%左右無法上升，嘗試了多種參數及特徵組合都無法改善，最後發現是我們參考的 rPPG open source code 提取出來的數值不合理，在更換成其他參考程式碼後，準確率才突破七成。

而在圖像特徵的部分，由於我們參考的原論文並沒有清楚地描述 CNN 架構，所以是在部分參考的基礎下自己建，因此可能沒有發揮論文自己研發出來的 CPPGMap 跟 CNN 配合的最大功用。

四、 主要成果與評估

我們用到的資料及有 DFDC[8]、Celeb-DF[9]、Face Forensics++[10]以及自己混合製作的 DFDC+Celeb-DF 資料集，這幾個資料集是 deepfake detection 領域中著名的資料集，表二為我們使用到的部分資料集真假影片數量。

資料集	真影片數	假影片數
DFDC	70	278
Celeb-DF	282	287
FaceForensics++	239	238
DFDC+Celeb-DF	267	278

表二：資料集真假影片數量

以數學特徵作為輸入時的部分測試組合請見表三，特徵欄位為了呈現方便僅列出特徵編號，這部分最佳的結果為用 Celeb-DF 資料集提取出的 9 個特

徵訓練的 SVC 模型，準確率為 0.76。

Model	資料集	特徵	Accuracy	Precision	Recall	F1
MLPRegression	DFDC	1,7,8	0.62	0.6	0.67	1.75
MLPRegression	DFDC	1,7,8	0.59	0.58	0.58	1.65
SVC	Celeb	15,20,19,28,24,4,3	0.75	0.81	0.73	0.74
SVC	FF	15,20,19,28,24,4,3	0.72	0.71	0.75	0.73
SVC	FF	15,20,18,19,28,24,4,3	0.73	0.72	0.75	0.73
SVC	Celeb	15,20,19,28,24,4,3,30	0.76	0.82	0.72	0.75
SVC	Celeb	15,20,19,28,24,4,3,30	0.76	0.82	0.73	0.75
SVC	Celeb	15,20,19,28,24,4,3,30,26	0.76	0.8	0.76	0.76

表三：數學特徵的部分成果

圖像特徵部分的測試結果如表四，我們嘗試過多種的 EffiicientNet 模型，其中準確率最高的組合就是 dropout 設為 0.3 的 EfficientNetV2M，準確率為 73%。

Model	Dropout	資料集	Input	Accuracy
EfficientNetV2B0	0.2	Celeb	Cppgmap	50%
EfficientNetV2S	0.2	Celeb	Cppgmap	59%
EfficientNetV2S	0.5	Celeb	Cppgmap	50%
EfficientNetV2M	0.3	Celeb	Cppg波形圖	73%
EfficientNetV2L	0.3	Celeb	Cppg波形圖	62%
EfficientNetV2L	0.3	FF	Cppg波形圖	49%
CNN		Celeb	Cppg波形圖	70%
LSTM		Celeb	Cppg波形圖	65%

表四：圖像特徵的部分成果

五、結語與展望

我們這次的專題嘗試了許多方法，數學特徵部分使用了三種模型：SVR、SVC 以及 MLP，輸入的話測試了多種 feature 組合，最後是採用 sequential forward selection 選出來結果最好的 9 個參數。圖像特徵部分使用三種模型架構：CNN、LSTM 以及 EfficientNet，以及兩種輸入：Cppg 訊號波形圖及 Cppgmap。最終兩邊最好的結果分別是 SVM 的 76% 以及 EfficientNet 的 73%。

若我們有更多的時間能繼續做研究的話，希望能嘗試看看 Vision Transformer(VIT)，以及研讀更多 feature 的挑選方法。讓我們能夠開發出準確率更高 deepfake 偵測軟體給民眾使用，讓 deepfake 的鑑定能擴展到民間，更高程度的避免錯誤資訊的傳播。

六、銘謝

在這一年的專題製作當中我們遇到許多困難，非常感謝指導教授以及實驗室學長姐在各種資源與技術上的協助，也感謝各位組員之間的合作，使本次的專題能夠順利完成。

七、參考文獻

[1] Umur Aybars Ciftci, Ilke Demir, Lijun Yin. (2020). How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. arXiv:2008.11363v1 [cs.CV]

[2] Daniel, M., & Ethan, B. (2019). iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox. arXiv preprint arXiv:1901.04366 [cs.CV].

[3] UmurAybars, C., Ilke, D., and Lijun, Y. (2020)"FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. X, no. X, pp. 1-1.

[4] Mingxing Tan, Quoc V. Le. (2021). EfficientNetV2: Smaller Models and Faster Training. arXiv:2104.00298v3 [cs.CV]

[5] Umur Aybars Ciftci, Ilke Demir, Lijun Yin. (2020). How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. arXiv:2008.11363v1 [cs.CV]

[6] Daniel, M., & Ethan, B. (2019). iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox. arXiv preprint arXiv:1901.04366 [cs.CV].

[7] UmurAybars, C., Ilke, D., and Lijun, Y. (2020)"FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. X, no. X, pp. 1-1.

[8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg and Matthias Grundmann, Google Research (2019). MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172v1 [cs.DC]

[9] H. Hu, Y. Wang, and J. Song. (2008). “Signal classification based on spectral correlation

analysis and svm in cognitive radio," in 22nd International Conference on Advanced Information Networking and Applications (aina 2008), pp. 883–887.

[10] A. Kampouraki, G. Manis, and C. Nikou. (2009). "Heartbeat time series classification with support vector machines," IEEE Trans. on Information Technology in Biomedicine, vol. 13, no. 4, pp. 512–518.