資料探勘109學年度下學期

利用AI模型判讀肺部X光圖片
以偵測新冠肺炎

組別

第三組

組員

資管碩一 R09725045 宋亭遠

資管碩一 R09725051 張智鈞

資管碩一 R09725059 吳昀蔚

資管碩一 R09725060 黃瀚陞

壹、 研究動機與架構

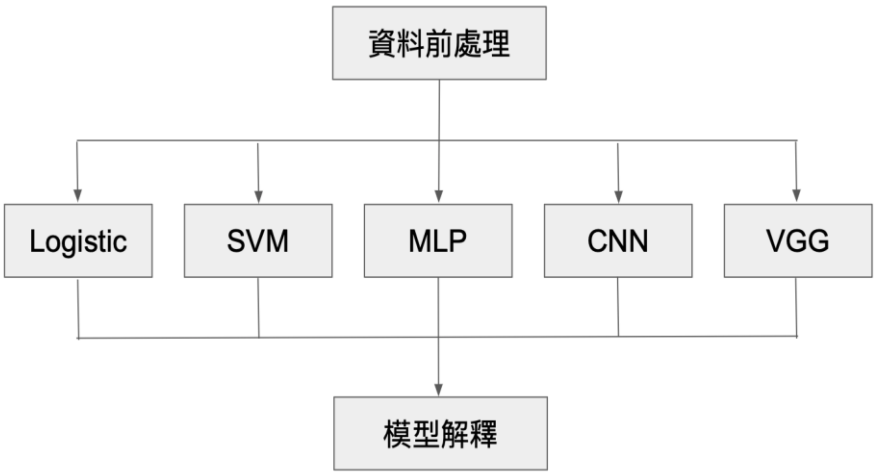
一、研究動機

自 2020 年以來爆發的 COVID-19 疫情至今已經造成全世界超過三百萬人死亡、無數人染疫，這使得如何控制此病毒流行已成為各國政府與公衛專家們刻不容緩的議題。要防止 COVID-19 病毒大規模的擴散與流行，目前各國大多採用的是採檢並隔離確診者的方式。

因此，如何準確又快速地找出確診者就變得非常重要。目前用來篩檢是否患有 COVID-19 的核酸檢測（Reverse transcriptase polymerase chain reaction，簡稱 RT-PCR）不僅成本高又耗時，而且還需要專業的設備和醫事人員執行。若是可以透過 AI 模型判斷是否感染 COVID-19，不僅速度較快而且準確性高，甚至還可以節省檢測時的人力成本。如此一來，將可以有效而且快速地偵測出患有 COVID-19 的病患，進而進行隔離，防止疾病大規模擴散。

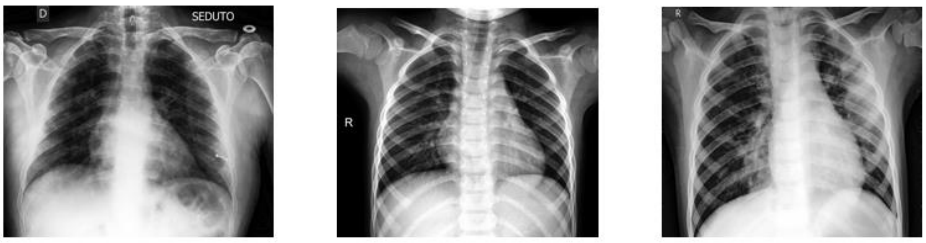
二、研究架構

因為 COVID-19 病毒主要是攻擊人類的肺部功能，因此若是可以觀察一人的肺部狀況，也許就可以藉此判斷此人是否染疫。在本次報告延伸過去報過的論文，我們希望透過 AI 模型針對肺部 X-Ray 照片進行辨識，判斷此人是否感染 COVID-19。由於資料本身較少，我們利用生成資料及 Transfer learning 的方式來建模，並且除了辨識是否患病之外，並以 LIME 的技術試著針對圖片進行解釋，供醫學人員進行此疾病更為深入的探討，以下為本研究架構的示意圖：



貳、 資料收集

本研究所使用之資料取自 Kaggle 網站上的「COVID-19 Image Dataset」(https://www.kaggle.com/pranavraikokte/COVID19-image-dataset)，此資料集的資料皆為 X 光圖片，而這些圖片又可再分為 COVID、Normal 以及 Viral Pneumonia 三種。其中，COVID 一類為感染新冠肺炎的病患的肺部 X 光照片；Normal 一類代表的是健康人的肺部 X 光片；Viral Pneumonia 則是感染一般病毒肺炎的病患 X 光照片，下圖為各類別照片的實例：



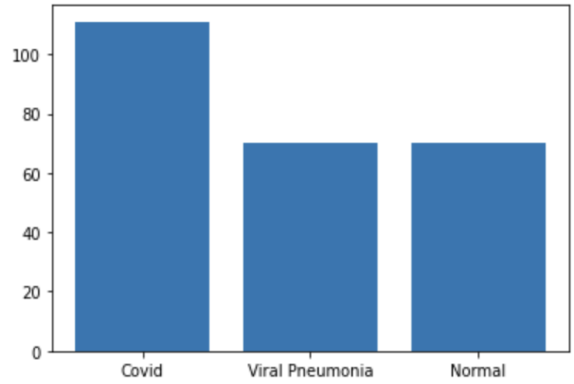
COVID　　Normal　　Viral Pneumonia

資料數量部分，此份資料集的提供者已將資料切分為訓練資料集與測試資料集，訓練資料集共 251 張照片（COVID 類別 111 張、Normal 類別 70 張以及 Viral Pneumonia 類別 70 張），而測試資料集共 66 張照片（COVID 類別 26 張、Normal 類別 20 張以及 Viral Pneumonia 類別 20 張）。

參、研究方法

一、資料前處理

1. 資料不平衡



上圖為各資料類別的數量示意，有上圖可知，原訓練資料集存在些微各類別資料量不平衡（111, 70, 70）的問題，為避免該問題造成模型訓練的偏頗，我們利用 Keras 套件中的 ImageDataGeneratoy 來生成資料，其中的參數包含旋轉與縮放，以最多資料的類別 Coivd 為基準，將另外兩類資料平衡至同樣數量的 111 筆。

2. 資料樣本小

因為本次訓練資料較少（共 251 筆），即便經過平衡，亦只有 333 筆，在層數較多的模型中可能因為樣本數過少導致訓練表現較差，故我們又再生成資料至各類別各 500 筆。

3. ZCA 白化

在初步觀察資料時發現不同類別的 X 光片的黑白（陰影）部分不同，COVID 的相對較白，且下半部大多是白的，而正常的則顏色較深（陰影較少），故推測若將圖片經 ZCA 過白化處理，不僅有降噪效果，亦能較清楚分辨顏色深淺分界，可能有助於辨識，故再加入了 ZCA 白化的參數，並生成資料至各類別各 500 筆。

原圖片

ZCA白化、旋轉、縮放

由於 ZCA 白化轉換需大量的 RAM，故在本次訓練中，在白化過程 resize 成 80*80 的大小，故上圖 ZCA 圖片的解析度較差。

4. 轉一維陣列

由於部分模型無法直接用圖檔訓練，故對於各資料集，除了原先的圖檔，亦將圖檔轉成大小一致的一維陣列，供模型訓練之用，ZCA 白化過的圖檔大小轉為 6400（80*80）維，其餘圖檔則是轉為 65536（256*256）維。

5. 小結：經過資料前處理後，共產生四種資料集，包含：

A. Original Dataset：原資料集。

B. Balanced Dataset(111)：各類別平衡至各類 111 筆。

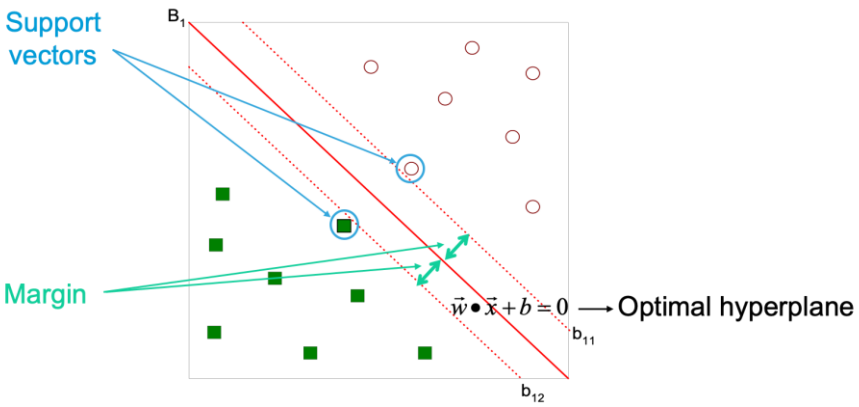
C. Balanced Dataset(500)：各類別平衡至各類 500 筆。

D. ZCA(80x80)：將圖檔經過 ZCA 白化處理，且各類別平衡至各類 500 筆。

二、建立模型

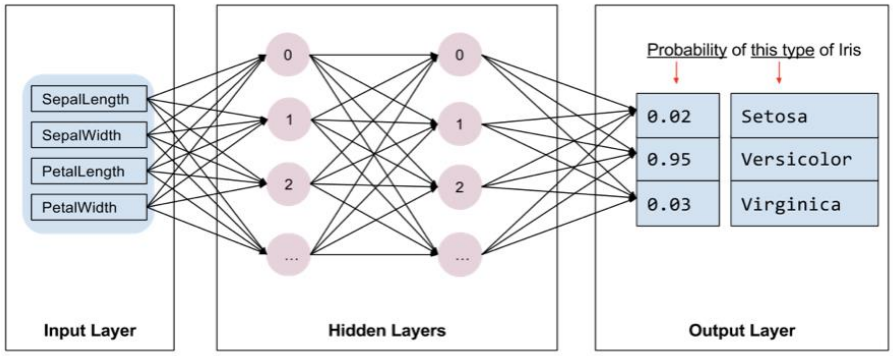
1. SVM

SVM 的目的在於找到能使區隔分類的邊界（Margin）最大化的 Optimal hyperplane，而邊界取決於最靠近分類界線的樣本資料們，這些資料稱為「支持向量」，此模型亦因此被命名為支持向量機，訓練過程中，亦可利用 kernel 的選擇，如 linear、polynomial、rbf 等，來符合資料分類的特性與需求。下圖為 SVM 在二維資料分類上的示意圖：



2. MLP(Multiple Layer perceptron)

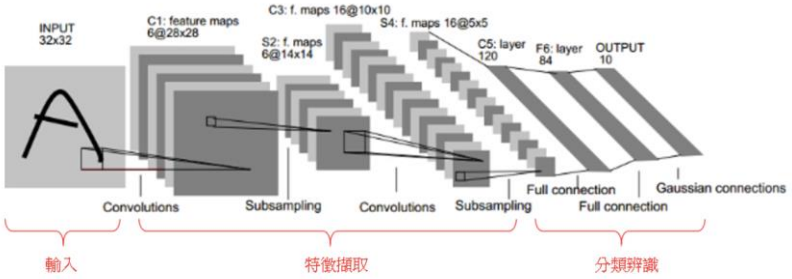
MLP 是 NN(Neural Network)的一種，NN 透過類神經元的方式來做非線性的特徵轉換，使用一組 W 給予一組特徵向量 X 不同的權重，並透過 activation function 進行非線性的轉換來處理非線性的一些分類問題，而 MLP 則是使用多層的 hidden layer 來學習抓取複雜的分類重要特徵，藉此來達到良好的分類效果。其簡單的架構如下圖所示：



（圖片來源：https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/685913/）

3. CNN(Convolution Neural Network)

CNN 是 Deep Learning 中眾多方法中的一種，常用於圖像辨識或 NLP(Natural Language Processing)領域，透過特殊的模型設計使得機器能抓到辨認圖片的重要特徵，藉此來達到很好的辨識圖片/理解字詞之能力。



LeNet Architecture, 1998

(圖片來源：https://medium.com/%E9%9B%9E%E9%9B%9E%E8%88%87%E5%85%94%E5%85%94%E7%9A%84%E5%B7%A5%E7%A8%8B%E4%B8%96%E7%95%8C/)

上圖為 CNN 的示意圖，CNN 運作的重要步驟大致可分為：

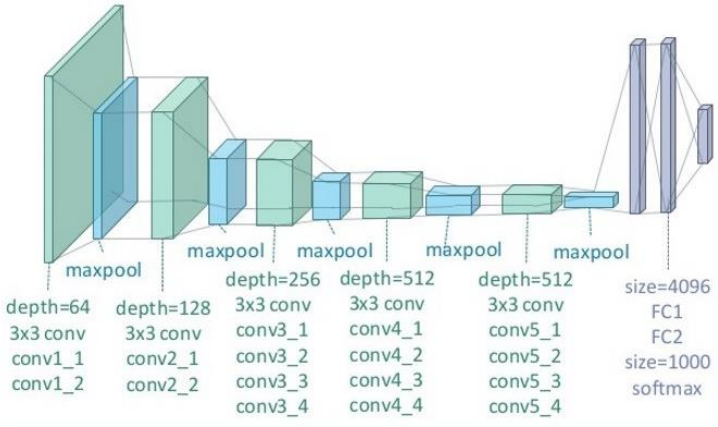
A. Convolution（卷積）：卷積就是在對圖片去做擷取特徵的動作，找出最好的特徵最後再進行分類

B. Pooling（池化）：對特徵圖(Feature map)降維，並且保留重要的特徵，參數減少，可防止 Overfitting。

4. VGG

VGG(Visual Geometry Group)是神經網路中非常經典的模型，較常見的又可以根據層數分為 VGG 16 以及 VGG 19。VGG 模型是在 2014 年由英國牛津大學的 Visual Geometry Group 所提出，此模型設計在 ImageNet 的比賽當中獲得非常好的成績。

本次報告中使用的 VGG 19 包含了 16 層卷積層和 3 層全連接層。除了模型深度更深之外，VGG 的模型設計上有別於 VGG 之前以前的 CNN 模型，使用了更小尺寸(3*3 conv)的濾波器(filter)從圖片當中獲得更多的資訊量並且這些多層的非線性層來增加模型非線性的程度。另外在池化層的部分，VGG 的提出者認為使用更小的 pooling 尺寸可以亦可保留更多的資訊量，因此在 VGG 模型中採用 2*2 的池化核，下圖為 VGG 19 模型的示意圖。



（圖片來源：https://www.slideshare.net/ckmarkohchang/applied-deep-learning-1103-convolutional-neural-networks）

除了對於模型的卷積層和池化層做調整，VGG 也有對於訓練及和測試及資料進行處裡。訓練集的部分，VGG 有採用 Multiple scale training，會將訓練資料較短的一邊利用亂數隨機縮放為 256 至 512 區間中的一個尺寸，接著再取圖片中間 224*224 的範圍輸入訓練網路。測試資料的部分則有採用 Multiple Crop Testing 的方式，會先把圖片放大至長寬皆是 280 的尺寸，接著分開針對四個角落還有中間 224*224 的範圍進行測試，最後才利用 Softmax 函式的輸出值平均作為預測結果。

使用 VGG 的好壞處方面，VGG 的優點是它的模型設計非常簡單易懂，每個卷積層都是使用 3*3 的濾波器，每個池化核的尺寸則都是 2*2，而且 VGG 也透過實驗證明了深的模型會優於淺的模型。不過，VGG 的缺點則是因為加深了模型深度，所以造成權重的數量提升，訓練上需要耗費較多的運算時間。

5. 模型比較

模型	Logistic Regression	SVM	MLP	CNN	Transfer learning (VGG)
速度	超快	快	中	中偏慢	較慢
Deterministic	Yes	Yes	No	No	No
缺點	不適用於非線性可分之資料	易受離群值的影響	解釋性差、易受模型架構影響	解釋性差、易受模型架構影響	解釋性差、模型深度深
優點	解釋性佳	泛化特性佳	可處理複雜問題、擴充性較佳	適合圖形辨識	沿用訓練好的權重

肆、研究結果

一、模型表現

以下為各資料集套用不同模型的訓練結果。

		Logistic Regression	SVM	MLP	CNN	VGG
Original Dataset	Accuracy	84.80%	86.40%	81.80%	89.40%	92%
	Precision	84.90%	86.00%	84.10%	90.30%	94%
	Recall	83.70%	86.20%	81.50%	89.50%	92%
	F1-score	0.83	0.86	0.81	0.89	0.92
Balanced Dataset (111)	Accuracy	78.80%	90.90%	75.70%	92.80%	91%
	Precision	78.20%	91.20%	63.90%	88.40%	93%
	Recall	77.80%	90.80%	63.20%	86.50%	91%
	F1-score	0.78	0.91	0.58	0.86	0.91
Balanced Dataset (500)	Accuracy	84.80%	94.00%	83.30%	93.90%	98%
	Precision	84.30%	94.40%	85.60%	93.70%	99%
	Recall	84.10%	94.10%	83.20%	93.30%	98%
	F1-score	0.84	0.94	0.83	0.93	0.98
ZCA (80x80)	Accuracy	77.30%	70.00%	65.10%	68.20%	89%
	Precision	76.50%	70.20%	65.20%	68.60%	92%
	Recall	76.50%	69.40%	64.40%	67.30%	89%
	F1-score	0.76	0.69	0.65	0.67	0.89

根據上表中的結果，我們得出以下幾點小結：

1. Logistic Regression、SVM 模型就有不錯的結果
2. VGG 相較表現最好，可能由於有訓練好的權重。
3. MLP 表現最差，可能是超參數設定或模型架構設計不良，相較之下 CNN 比較適合圖形辨識。
4. Balanced Data 500 訓練效果都比 Original data 來得好。
5. 以 confusion matrix 來看，COVID 類別的分類普遍都有最好的表現，但另外兩類有時會發生混淆的狀況。以下圖（使用 VGG 模型，Original Dataset）為例，編號 0 的是 COVID 類別，編號 1、2 則分別是 Normal 和 Viral Pneumonia 類別，可看出後兩類的分類結果較差。

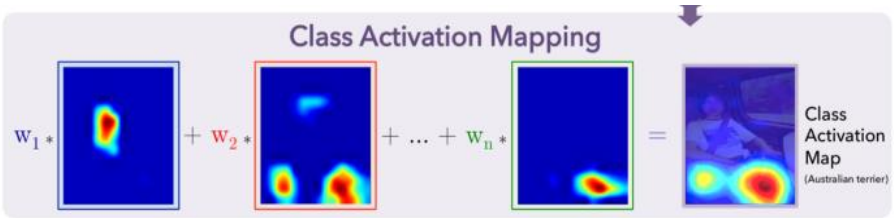
6. ZCA 的處理對於分類表現並沒有改善，可能 X 光照片較不適合這樣處理（只會剩下骨頭），並且這邊因為計算時間有進行縮圖(256 -> 80)，可能也造成資訊減少。
7. SVM 相較自己訓練的 MLP 有較好的表現，可能因為 SVM 比較適合少量的樣本，又由於 MLP 模型設計及參數的選擇上較複雜，導致較差的結果。

二、模型解釋

1. GRAD-CAM (Gradient-weighted Class Activation Mapping)

透過 Grad-CAM 我們可以了解一個執行圖像分類任務的 CNN 在它自身的網路裡是因為看重照片中的哪一個區域，進而做出該 CNN 模型的分類判斷。

CAM (Class Activation Map)，其做法是在最後的卷積層之後接上 GAP 層(Global Average Pooling Layer)，取代一般會使用的攤平全連接層 (Flatten Fully Connecting Layer)，經過 GAP 轉換後的每一個神經元分別對應到了最後一層的某一張特徵圖，而 GAP 層所連接的權重即可視為每一張特徵圖對於模型預測類別的重要性，最後將每一張特徵圖依照其對應的權重進行加權即得到 CAM (Class Activation Map)。



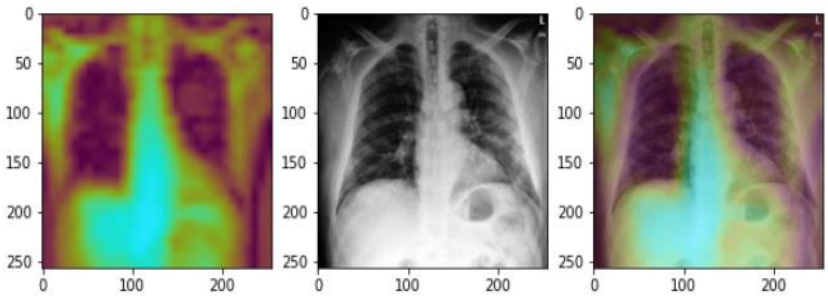
（圖片來源：https://medium.com/ai-academy-taiwan）

從 CAM 的介紹我們知道要能夠實踐 CAM 最後的卷積層輸出一定要接上 GAP Layer，可是這無疑的限縮了網路架構的設計方

式，而 Grad-CAM 關鍵是能夠透過反向傳播(Back Propagation)計算在 CAM 中使用的權重 w，不論模型在卷積層後使用的是何種神經網路，不用修改模型就可以實現 CAM。

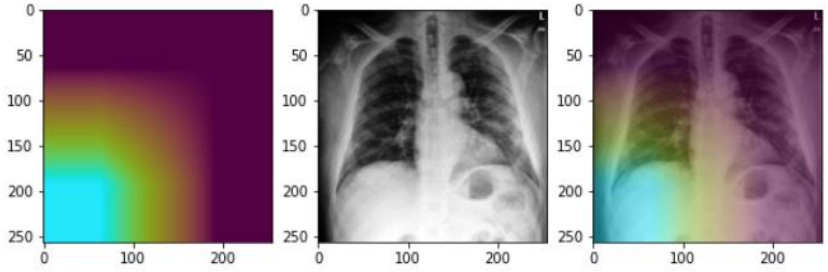
A. CNN 結果（由類別是 COVID 的第 25 張為例）

以下為第一層卷積層結果：



由 GRAD-CAM 結果可以看出，CNN 中第一層的卷積層重要的部分比較像是整體骨頭的部分，由他的熱點圖可以看出 X 光中脊椎及心臟的部份，對此層來說是最為重要的。

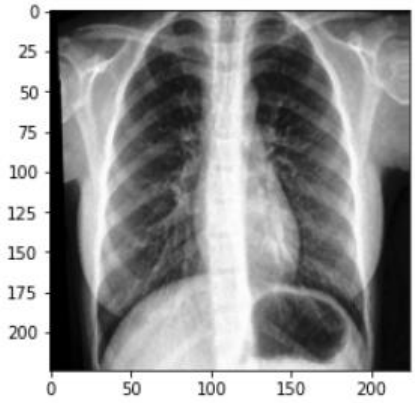
以下為第二層卷積層結果：



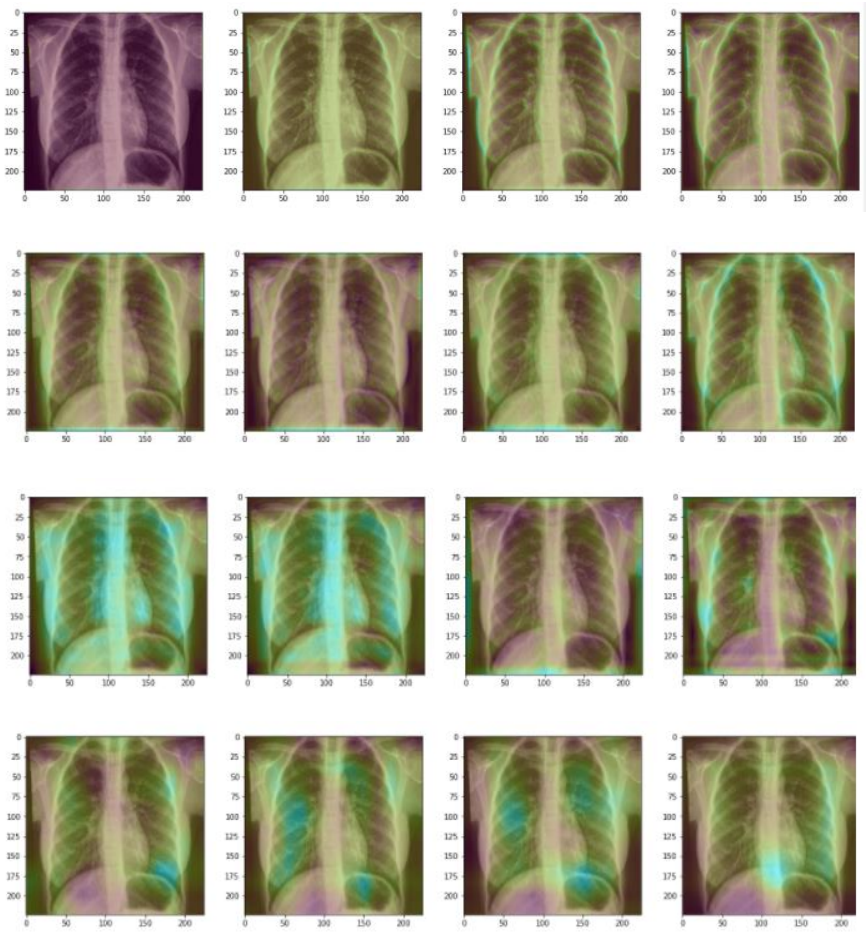
由 GRAD-CAM 結果可以看出，CNN 中第二層的卷積層重要的部分轉變成了左肺的下半部，相較第一層看到較細微，可能患有 COVID 左肺的下半部有特別的病變。

B. VGG-16 結果（由類別是 Normal 的第 40 張為例）

以下為 Normal 第 40 張照片的原圖：

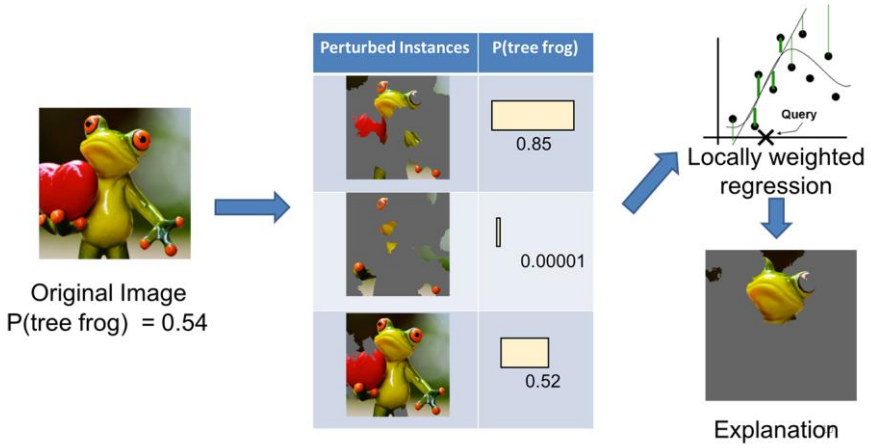


VGG-16 共有 16 層卷積層，以下是照順序各層的熱點圖：



前面幾層主要是看整體以及骨骼的部分，後面會去細看肺部的不同部分，如左肺上下半部，右肺上下半部，以及脊椎心臟部分，可以配合其他 有 Domain 的醫療單位，了解到這些部位分別是哪裡，協助偵測是否染疫以及後續的治療。

2. LIME (Local Interpretable Model-Agnostic Explanations)



（圖片來源：https://medium.com/@kstseng）

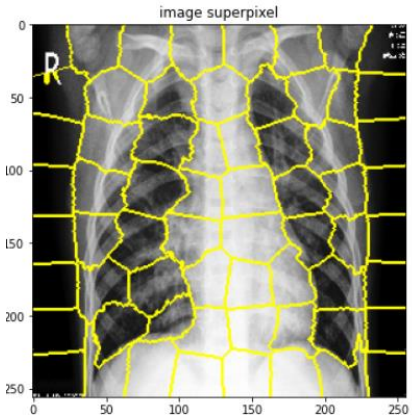
除了 Grad-CAM 比較粗略的熱點圖來看重要的圖片部分，我們也應用 LIME 希望可以探討不同類別分別重要的特定圖片部分。

圖片進行 LIME 之前會先把圖片中有相似特徵的像素進行聚合，形成較大的超像素（Super Pixel），我們這邊利用的是 SLIC (simple linear iterative clustering)，採用類似 K-mean 的算法生成超像素，不同的是他只在一定的有限區域中計算各點的距離，減少計算量，並且可以形成較規則整齊的像素分割。

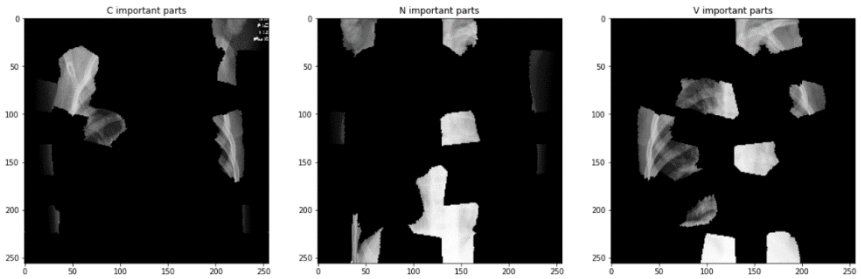
而 LIME 會利用這些 Super Pixel 隨機生成不同圖片樣本，產生不同的排列組合（perturb the input），之後利用已經訓練好的複雜模型預測這些被擾動過的資料，計算預測出各類別的機率，最後透過簡單的線性迴歸模型，X 是不同超像素的排列組合(1/0)，Y 是某類別的機率，透過這個簡單的線性迴歸，可以了解到對於特定類別，哪些超像素的權重較大，也代表著他出現與否會很大的影響大模型預測的機率變化，藉以找到重要的超像素。

A. CNN 結果（由類別是 COVID 的第 150 張為例）

圖片切割的超像素：



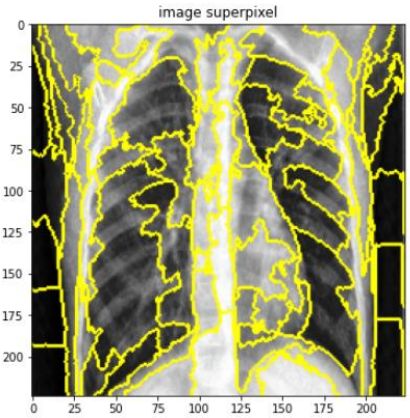
不同類別重要的超像素：



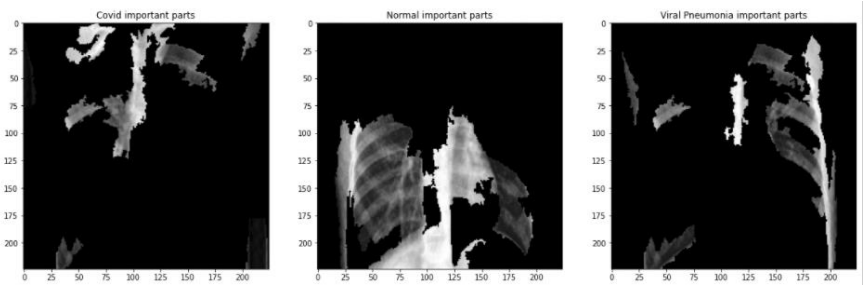
COVID 重要的超像素主要是肺部的上方，代表這幾塊對於預測 COVID 有較大的影響力，Normal 則是中間脊椎以及心臟的部分，對於病毒式肺炎這要的是左肺部的部分。

B. VGG-16 結果（由類別是 Normal 的第 25 張為例）

圖片切割的超像素：



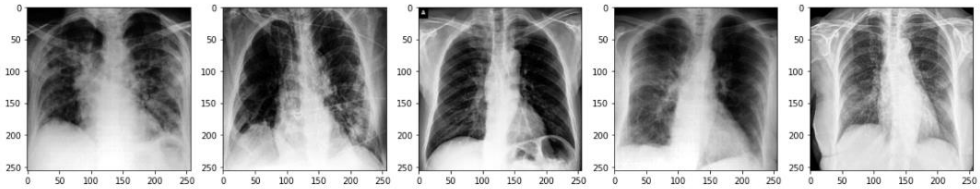
不同類別重要的超像素：



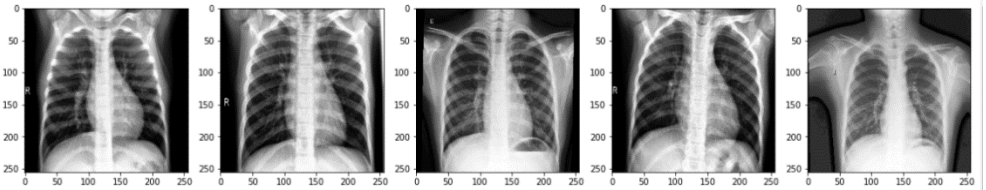
COVID 重要的超像素主要是肺部的上方，代表這幾塊對於預測 COVID 有較大的影響力，Normal 則是中間脊椎以及心臟的部分還有左肺的下半部，對於病毒式肺炎這要的是右肺的上半部的部分。

3. 其他解釋

我們也對照原本的圖片，分析一下模型結果，以下為感染 COVID 者的肺部圖片：



以下則是健康的肺部圖片：



由前面兩組圖可以看出相比之下染疫者可能因為肺部的病變，心臟脊椎的部分皆不清楚，並且肋骨的部分也較無感染者來說不清晰，並且有明顯的白斑狀況，有些肺部也有萎縮的狀況，這可能可以解釋為何模型的效果那麼好，以及為何模型認為特定部分較為重要。

伍、 結論

在本篇報告中，我們使用了羅吉斯迴歸、SVM、MLP、自行設計的 CNN 架構以及 transfer learning(VGG)的方式針對 X 光片做預測，預測類別共有 COVID、Viral Pneumonia 以及 Normal 三種。資料預處理部分，由於原始的資料集的數量較少（僅包含 251 筆訓練資料、66 筆測試資料），我們透過旋轉、縮放的方式再生成了另外三種種資料集進行預測，包括了 Balance Dataset(111)以及 Balanced Dataset(500)兩種資料集，以及 ZCA 白化技術後生成的 ZCA(80x80)資料集。

透過我們上述的分析流程及方法，我們可以高機率（98%）的檢驗出染疫與否，了解模型中那些肺部的特定區塊較為重要，增加模型可信度。另外，在資料前處理的部分，透過旋轉及縮放等機制增加圖片數量，可以有效的增加模型準確度，比較相關模型，機器學習需要較大的數據來訓練模型以及驗證，未來若能在臨床上搜集更多數據，能增加深度學習模型的準確度與可靠性。

除此之外我們也透過可解釋 AI 的部分，了解到 X 光中心臟脊椎的部分較為重要，某些肺部部分染疫與否差異較大，並且也發現資料集的 COVID 可能都是症狀嚴重的患者，所以肺部狀況才明顯，但對於輕症的感染者可能就無法預測出來是否染疫，總結透過這樣的架構，有很良好的預測能力，也可以增加模型的可信度，可解釋 AI 的部分也可能協助醫療人員的治療，但如果要實用還需多考量更多資料，以及輕重症感染者的部分。

陸、 參考文獻與圖片來源

Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M., & Shakhawat Hossain, M. (2020). COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. Machine Learning and Knowledge Extraction, 2(4), 490-504.

Mark Chang. (2016, October 27). Applied Deep Learning 11/03 Convolutional Neural Networks. slideshare. https://www.slideshare.net/ckmarkohchang/applied-deep-learning-1103-convolutional-neural-networks

François Chollet. (2020, April 26). Grad-CAM class activation visualization. Keras. https://keras.io/examples/vision/grad_cam/

分類演算法-多層感知機 Multi-layer Perceptron。IT 人。2020 年 1 月 19 日。https://iter01.com/454359.html

[機器學習 ML NOTE]Convolution Neural Network 卷積神經網路。雞雞與兔兔的工程世界。2018 年 12 月 21 日。https://medium.com/%E9%9B%9E%E9%9B%9E%E8%88%87%E5%85%94%E5%85%94%E7%9A%84%E5%B7%A5%E7%A8%8B%E4%B8%96%E7%95%8C/

VGG_深度學習_原理。JT。2018 年 4 月 16 日。https://danjtchen.medium.com/

孫曉恩(2020 年 7 月 4 日)。了解 CNN 關注的區域:CAM 與 Grad-CAM 的介紹分享。AI 台灣人工智慧學校。https://medium.com/ai-academy-taiwan/

曾凱聲(2017 年 12 月 27 日)。LIME - Local Interpretable Model-Agnostic Explanation 技術介紹。Kai-Shen Tseng (kst)。https://medium.com/@kstseng/

JNingWei(2017 年 10 月 14 日)。影像處理: 超圖元(superpixels)分割 SLIC 演算法。CSDN。https://blog.csdn.net/JNingWei/article/details/78236098