智慧化企業整合

Skin Cancer

Using CNN model to classify

Group 9

108034401 蘇靜琦

108034553 尤子維

108034555 蔡沛洹

108034557 陳宣任

目錄

一、 背景介紹

二、 問題定義

三、 模型架構

四、 分析與改善

五、 結論

一、背景介紹

癌症一直來都是在全台灣當中的10大死因之一，而其中罹患皮膚癌比例佔了所有癌症的四成。除此之外，由台灣癌症登記小組的資料，我們更可以看到皮膚癌的發生率逐年上升，由民國68年至民國85年，全國皮膚癌每年的申報人數由248人增至1220人，上升了將近5倍；至95年已經到達2457人，上升了將近10倍之多。若皮膚癌能早期發現，並且確定病徵，都能夠透過手術以及藥物得到改善，因此我們透過建構神經網路模型的方式，提高辨認皮膚癌病徵的準確率，使得病患能夠提早發現病徵，得到最適當的治療。。

建構神經網路的過程大致可以分為七個步驟

1.資料收集、分類
2.資料前處理
3.建立神經網路架構
4.訓練神經網路模型
5.透過測試集預測神經網路成效
6.不斷評估模型，優化此模型
7..提出結果

二、問題定義

1.5W1H分析法

5W1H分析法是一種思考方法，對選定的項目、工序或操作，都要從對象（何事WHAT）、時間（何時WHEN）、人員（何人WHO）、地點（何地WHERE）、原因（何因WHY）、方法（何法HOW）等六個方面提出問題進行思考，且對問題進行綜合分析研究，從而得到更具建設性設的決策。

利用此方法，我們主要想針對辨認皮膚癌的病徵來進行改善。

(1) What?: 要解決甚麼問題?
解決單獨由醫生判斷錯誤，可能準確率較低的問題。

(2) When?: 在甚麼時候要辨別病徵?
在醫生確認並辨別病人皮膚病徵的時候。

(3) Who?: 由誰來改善目前診卻率不夠高的問題?
醫院的研究部門提出此模型，看診醫生同時採用此模型，增加判斷準確率。

(4) Where?: 在哪種地方可以進行問題改善?
醫院的皮膚科

(5) Why?: 為什麼要進行改善此問題?
避免錯誤診斷導致病人看診時間延誤、提高醫院皮膚科的名聲、給予病人更好的治療。

How?: 如何解決此問題?

建構神經網路模型，協助醫生進行病人皮膚癌的病徵判斷

三、模型架構

1. 資料輸入(Input)

為了解決辨認皮膚癌病徵的問題，我們使用了 isic 所提供的資料集，共有 7 種不同病徵的皮膚癌照片，共 10015 張。其中包括黑色素細胞癌(melanocytic nevi,nv)6705 張、光化性角化病或腺癌或鱗狀上皮癌(Actinic keratoses and intraepithelial carcinoma / Bowen's disease,akiec)327 張、基底細胞癌(basal cell carcinoma,bcc)514 張、曬班或脂漏性角化症或扁平苔癬樣角化病(solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl)1099 張、皮膚纖維瘤(dermatofibroma,df)115 張、黑色素瘤(melanoma,mel)1113 張、血管病變(vascular lesions,vasc)142 張，我們將其作為模型的投入。

2. 資料前處理

由於資料各種病徵的張數分布不均，且照片容易受到拍攝角度、光線、大小等等的影響，造成我們模型訓練的效果不好，因此我們透過各種不同的前處理方式，將資料進行處理，使模型的準確率能夠提高，原始模型的前處理可分為下面幾個步驟:

(1)將放置在各個不同的資料夾的圖片進行合併。

(2)在 ISIC 所提供的資料中，在年齡這個項目有 57 格為空值，我們填補為 0

(3)將所有圖片裁減為相同大小，使每張的像素點數一致，才能夠放入模型中

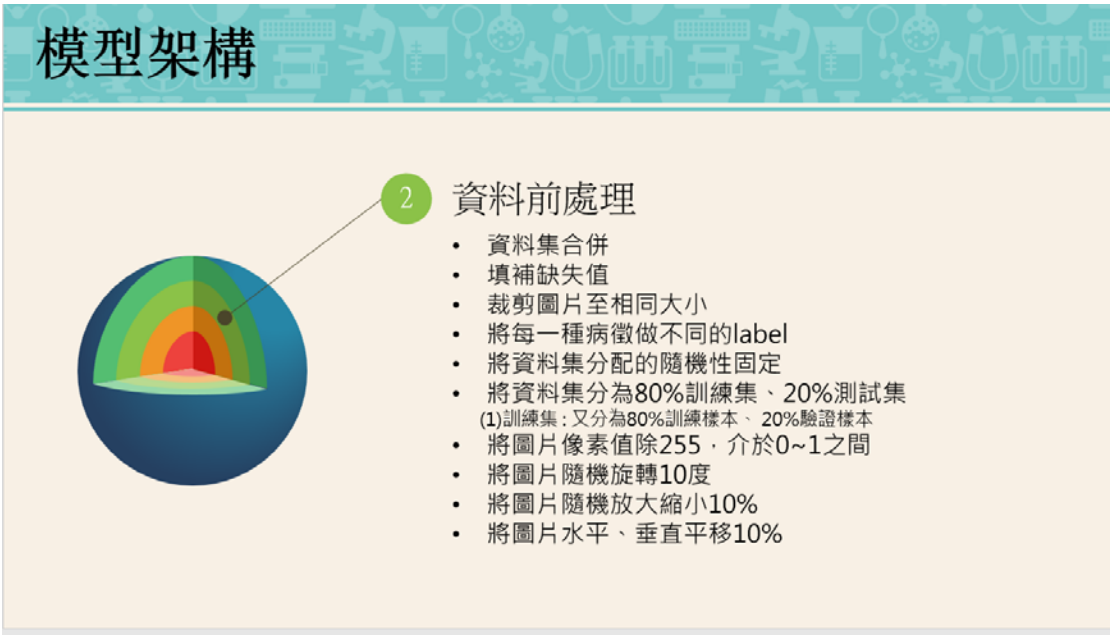
(4)所有照片共分為 7 種病徵，因此先將圖片 label 為 7 類。

(5)將區分資料集的隨機性固定(為了後續模型比較，避免因為抽取不同樣影響訓練出來的結果。

(6)將資料集的 80%做為訓練集，20%做為測試集。訓練集當中再取出 10%做為驗證集。

(7)將圖片的像素值除上 255(像素值介於 0-255 之間)，使像素值介於 0-1 之間

(8)在訓練過程，將圖片隨機旋轉 10 度，隨機放大縮小 10%，圖片水平垂直平移 10%，加強模型的訓練。



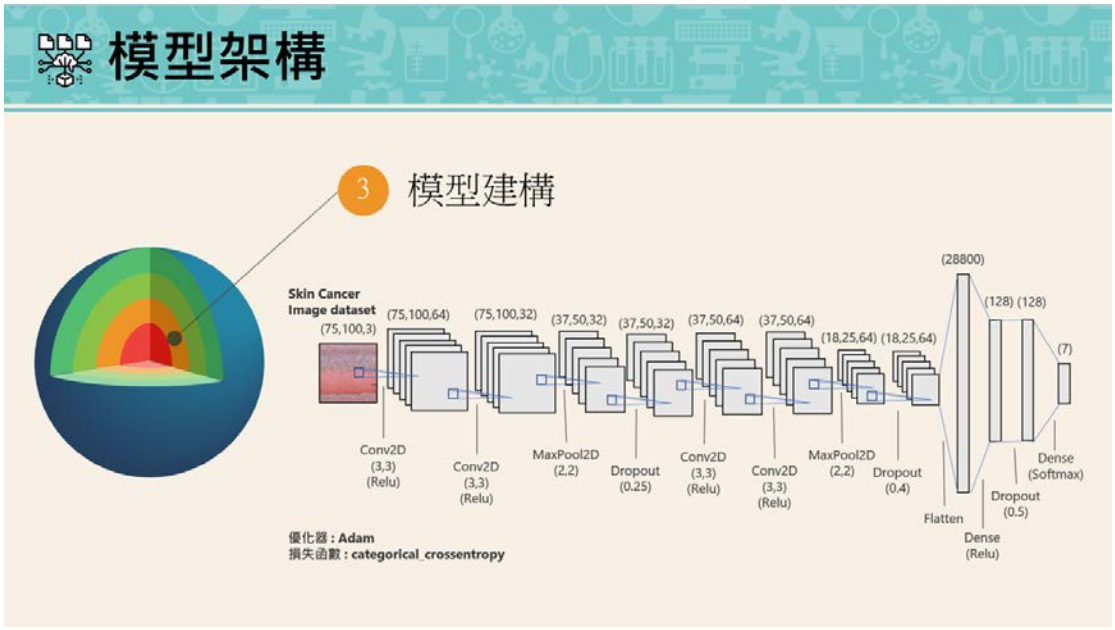
3. 模型建構

將資料做完前處理後，必須決定使用的神經網路類型以及其內部架構，由於我們投入神經網路模型的項目為圖片，使用卷積神經網路模型能夠進行圖片特徵的截取，幫助我們保留各種不同病徵的圖片特徵，從而進行病徵分類。

此章節說明之模型架構為 kaggle 網站中他人所提出的原始模型，而我們這次 porject 著重在於如何提高這個模型的準確率，分析過程及最終模型，將在第

四章、分析與改善提出。

原始模型當中，卷積神經網路的架構為:
(1)照片輸入(75,100,3)
(2)一層卷積核大小為 3×3，個數為 64 的卷積層
(3)一層卷積核大小為 3×3，個數為 32 的卷積層
(4)一層池化大小為 2×2 的池化層
(5)進行比率為 0.25 的 Dropout
(6)二層卷積核大小為 3×3，個數為 64 的卷積層
(7)一層池化大小為 2×2 的池化層
(8)進行比率為 0.4 的 Dropout
(9)一層隱藏單元為 128 個的全連接層
(10)進行比率為 0.5 的 Dropout
(11)個數為 7 的輸出層(分類 7 種不同病徵)
(13)各層激活函數:relu
(14)輸出層激活函數:softmax
(15)損失函數:categorical_crossentropy



訓練集共:7210 張 驗證集共:802 張 測試集共:2003 張

根據此架構，訓練 50 個 Epoch、batch size= 10，每訓練完一次 Epoch 驗證一次結果，可以根據訓練狀況來確認是否有過擬合的情況產生。

4. 結果輸出

根據此模型的訓練結果，最後一次訓練集的準確率為 0.7257，驗證集準確率為 0.7182，測試集準確率為 0.7159。

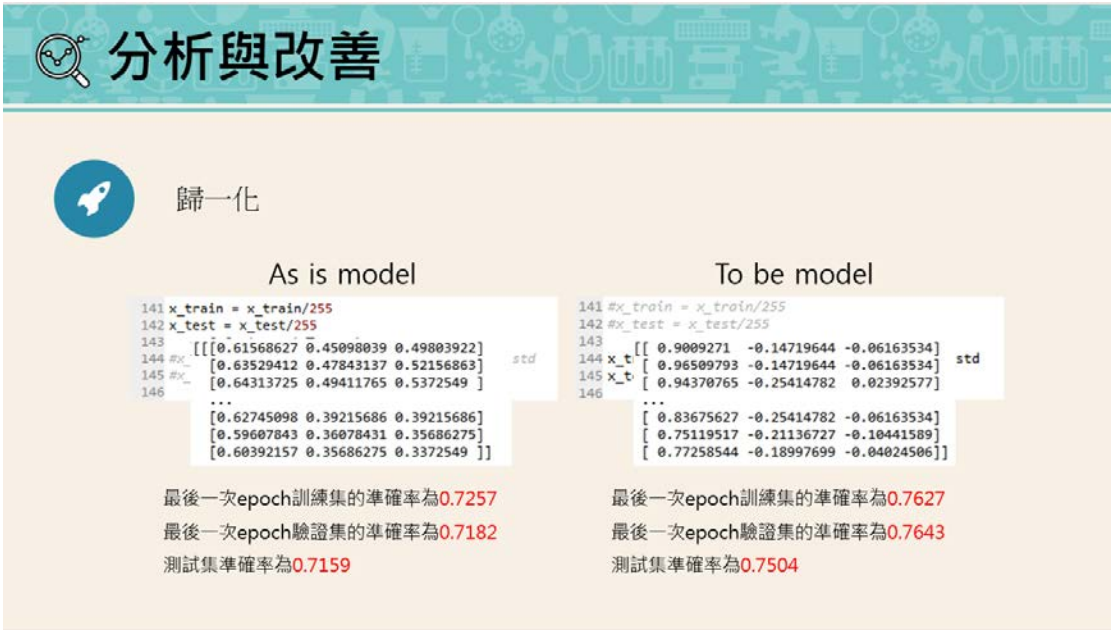
從原始模型當中的結果我們可以看出病徵分類的準確率仍偏低，因此我們

將透過進行不同的前處理、優化器、損失函數、激活函數、參數、層數等等，使模型的分類準確率能夠提高。

四、分析與改善

1. 前處理-歸一化

首先我們先將原本圖片是像素值除 255 的方式投入神經網路，改為使用歸一化 Z-score 的方式，將像素值減去平均在除標準差，得到又投入神經網路模型的值，使用歸一化的方式有助於我們模型收斂的速度更快，同時也能降低圖片每張拍攝光線不同的問題。



2. 決定損失函數和激活函數的組合

原先的模型激活函數是使用 softmax，而損失函數使用 categorical_crossentropy，categorical_crossentropy 通常是用於多分類，而 binary_crossentropy 通常是用於二分類，但事實上在多標籤多分類，使用 categorical_crossentropy 做為損失函數會使模型難以收斂，但這兩個損失函數也有其輸出層適合的激活函數，因此我們採取實驗設計的方式，將 4 種模型進行比較:

激活函數	損失函數	優化器	訓練集準確率	驗證集準確率	測試集準確率
softmax	Categorical crossentropy	adam	0.7627	0.7643	0.7504
softmax	Binary crossentropy	adam	0.9340	0.9186	0.9136
sigmoid	Binary crossentropy	adam	0.9400	0.9371	0.9345
sigmoid	Categorical crossentropy	adam	0.7598	0.7444	0.7374

我們可以看出與我們的分析結果一致，binary_crossentropy 在訓練結果明顯比 categorical_crossentropy 好很多，同時使用 sigmoid 做為激活函數結果最好。

3. 決定優化器

原先使用的優化器為 adam，我們再加入了 adamax、Nadam 兩種不同的優化器來進行比較，adamax 為 adam 的一種變體，改變了 adam 的學習率調整方式，而 Nadam 則是帶有 Nesterov 動量項的 adam。

激活函數	損失函數	優化器	訓練集準確率	驗證集準確率	測試集準確率
sigmoid	Binary crossentropy	Adam	0.9400	0.9371	0.9345
sigmoid	Binary crossentropy	Adamax	0.9364	0.9291	0.9317
sigmoid	Binary crossentropy	Nadam	0.9263	0.9252	0.9232

根據訓練出來的結果可以看出，仍然是 adam 做為優化器在這個模型的結果為最佳。

4. 調整卷積層數

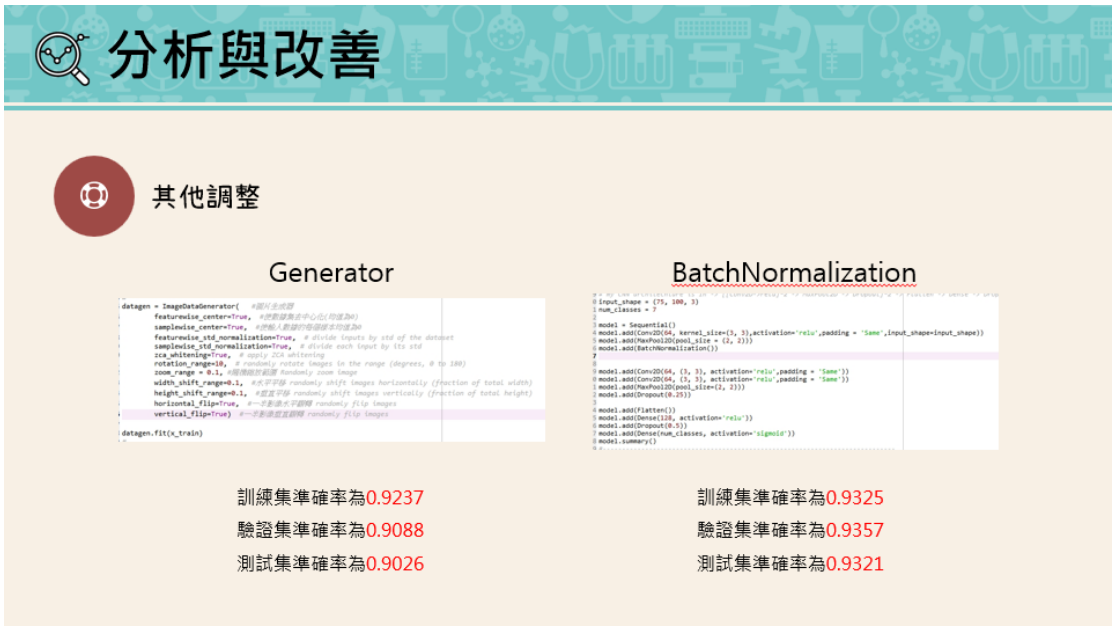
由於我們不知道卷積神經網路模型應該要有多少層數才能使分類效果最好，因此我們透過一層一層增加卷積層的方式，找到模型分類效果最好的層數

第1次池化前的卷基層數	訓練集準確率	驗證集準確率	測試集準確率
2	0.9400	0.9371	0.9345
1	0.9383	0.9361	0.9367
3	0.9337	0.9323	0.9304
4	0.9346	0.9330	0.9320
5	0.9325	0.9339	0.9341
6	0.9060	0.9088	0.9026

根據訓練出來的結果，我們得到在進行第一次池化前的卷積層數為 1 層時，訓練效果最好。

5. 其他調整

除了上述那些顯著影響預測結果的項目之外，我們同時也做了其他調整:
(1)利用圖片生成器(Generator)，生成各種與原圖不同的照片
(2)增加 BatchNormaliztion 標準化層，有人認為使用 Dropout 在卷積網路模型中會讓在擷取特徵時，遺失一些重要特徵，反而使訓練效果不好，因此我們也同樣使用 BatchNormaliztion 來勳練我們的模型。



最後可以看出，我們使用的這兩種方式對於這個模型，都沒有辦法提高模型的準確率，同時還使訓練模型的時間更長，因此後續就沒有將此兩種方法納入考慮之中。
我們所改善過後的最終模型，架構如下:
(1)照片輸入(75,100,3)
(2)一層卷積核大小為 3×3，個數為 64 的卷積層
(3)一層池化大小為 2×2 的池化層
(5)進行比率為 0.25 的 Dropout
(6)二層卷積核大小為 3×3，個數為 64 的卷積層
(7)一層池化大小為 2×2 的池化層
(8)進行比率為 0.4 的 Dropout
(9)一層隱藏單元為 128 個的全連接層
(10)進行比率為 0.5 的 Dropout
(11)個數為 7 的輸出層(分類 7 種不同病徵)

五、結論

1. 分析與改善小結

根據我們改善後所提出的模型，訓練集準確率達 0.9384、測試集準確率達 0.9367，比起原先所提出的模型訓練集準確率 0.7257、測試集 0.7159 有著非常顯著的提升，可見這次針對模型的改善，的確有使模型的預測能力提升。根據我們這次所建立出的模型，有著 93.67%的準確率，可以協助醫生在判斷病人皮膚癌病徵時，能夠協助醫生做出正確的判斷，讓病人能夠接受正確的治療。

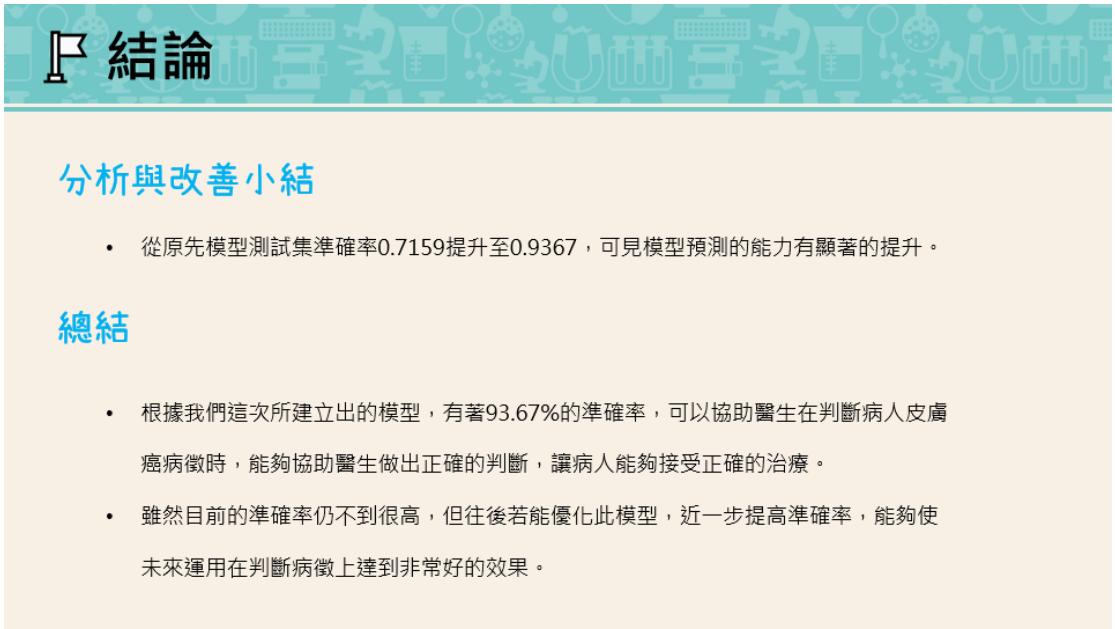
雖然目前的準確率仍不到很高，但往後若能優化此模型，近一步提高準確率，能夠使未來運用在判斷病徵上達到非常好的效果。

2. 模型及研究限制

這次我們透過實驗設計所訓練出的模型，即使經過了幾十次的不同的架構，仍然無法使模型測試集的準確率提高到 0.94 以上，主要受到下面幾點的原因所影響:

(1)各種病徵種類的照片張數分布不均，資料集當中，屬於黑色素細胞癌的病徵就佔了超過所有資料集 6 成的圖片，而像是皮膚纖維瘤、血管病變、光化性角化病僅僅只有 115、142 及 327 張，而這樣會導致某些病徵模型學習的次數較少，導致分類預測不準確。

(2)這次訓練模型的過程，由於訓練一次(50 個 Epoch)需要 6-8 個小時，且每次修改只能變動一個指標，因此沒辦法將所有可能優化模型的狀況皆跑過一次，因此這次所得到的結論只是根據我們所有試過的模型當中，所推得的結果，並不是所有可能組合當中最好的結果。



3. 後續優化方向

由於目前的研究結果準確率仍無法達到最佳，因此若要繼續優化此模型，可以藉由下面的幾種方式，來加強卷積神經網路模型的準確率，甚至是提高訓練效率。

(1)增加圖片張數較少的病徵種類數量，讓每個分類結果都有被完整訓練到

(2)使用幫助提高訓練效果的方法:

AdaBoost

AdaBoost，是一種集成學習的方式，它會將訓練集當中訓練錯誤的樣本，提高在前幾個分類錯誤樣本的權重，這樣可以使下一次訓練能夠更聚焦於分類錯誤的訓練樣本上，優化模型的訓練結果。

(3)提高訓練速度的方式:

Shuffle Net

CNN 模型若計算量過大，則可能導致模型難以在筆電等設備上運行，對此有人提出 Shuffle Net 這個輕量級網路的架構，且為了達到降低計算複雜度的功能，用到了 pointwise group convolution 和 channel shuffle 兩種作法，並透過實驗證實了有此架構的效果。

(3)使用 autokeras

Autokeras 是一個新的套件，只需要給定資料集，並將模型投入前處理完成，給定總訓練時間，autokeras 就會在此時間內不斷尋找最好的模型結構，雖然其找到的模型架構並不一定是最好的，但若能在手動調整參數前，先找到一個較好的模型結構，有助於我們大幅縮短改善模型的時間。