Small Data Training for Medical Images

Sponsored by HTC DeepQ

Team Name: 泱泱溫妮琳妲凡凡千千提莫

Members: 駱奕霖 蘇建銓 劉凱翔 曾柄元

Graduate Institute of Computer Science & Information Engineering
National Taiwan University

Graduate Institute of Communication Engineering
National Taiwan University

Graduate Institute of Electronics Engineering
National Taiwan University

ABSTRACT

我們針對 National Institutes of Health (NIH) 所提供的 Chest X-Ray Dataset 來判定對於 Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia 這 14 種疾病的發生與否提出建議方法。這個方法是基於 Imagenet 上預先訓練好的參數權重並搭配著 classifier & CNN layers dropout, regularization, augmentation, class weight, self-ensemble on testing, model ensemble 對於在 Imagenet 上 InceptionResnetV2, InceptionV3, Densenet121 這 3 種不同的模型架構下做訓練。此方法中的這些技巧對於避免在訓練資料上過度擬合以及 Area Under the Receiver Operating Characteristic curve (AUROC)的準確率提高上都有非常好的效果。

I. INTRODUCTION & MOTIVATION

胸部 X 光檢查是最常見和最具成本效益的醫學影像檢查之一。然而，胸部 X 射線的臨床診斷其實是非常具有挑戰性的，有時比通過胸部電腦斷層掃描診斷更加困難。在缺乏大型公開且具註解的數據的情況下，要在醫療站點中實現臨床相關的電腦輔助檢測和診斷（CAD）仍然是非常困難的。建立大型 X 射線圖像數據的一個主要障礙是缺少大量被標記的圖像資源。

在醫學發達的現今，許多疾病已經被證實，如果能提早發現是可以提早預防與提早治療來避免嚴重後果。X 光圖像是常用來判斷疾病的依據，但要能從 X 光片判讀該病患是否有疾病，醫生的專業判斷是不可缺少的，在醫療上金錢與時間的成本是很大的。在這次的期末專題中，我們希望可以藉由在這堂課所學到的知識來預測病患可能罹患 14 種疾病的機率來幫助醫生做輔助上的診斷，降低金錢與時間的成本，讓疾病可以提早被發現與解決。

該 NIH 胸部 X 射線資料包含 112,120 張 X 射線圖像，其中包含來自 30,805 名獨特患者的疾病標籤。在這次期末專題中提供了其中 10,002 張含有標籤的圖像與 68,466 去除標籤的圖像當作訓練資料，如何在如此少量具有標籤的圖像下達到很好的預測疾病的模型成為這次期末專題的重點。許多論文也針對 NIH 這次所提供的資料提出不同的做法，[1]提出了 densenet121 對於 NIH 胸部 X 射線資料偵測 14 種疾病效果非常好，比先前其他的 state of art 還要優越。[2]使用了 4 種不同的架構模型來觀察對於 14 種疾病中的 8 種疾病預測的結果比較，包含 AlexNet、GoogleNet、VGGNet-16 和 ResNet-50，其中又以 ResNet-50 表現最出色，另外在架構中使用了 weighted cross entropy loss (W-CEL)而非一般的 CEL，論文中也提出了 Log-Sum-Exp (LSE) pooling 的方法並找出最適合此 pooling 的最佳參數，最後對於疾病的關聯性有深入的探究。

此份報告整理如下，所有技巧的做法與其對應想解決的問題或原因將呈現在第 II 節。曾經嘗試但效果沒有很顯著的方法將在第 III 節做解說，包含 pseudo-label, auto-encoder, single disease training。如何進一步改良作法與提升效果不如預期的方法將在第 IV 節做討論。結論與可能的未來工作將呈現於第 V 節和第 VI 節，引用說明於第 VII 節。

II. Proposed Method

A. Extractor & Classifier Architecture and Dropout Method

我們將整個模型架構分為兩部分，convolution layers extractor 和 dense layer classifier。使用 Imagenet 的 pre-trained weights 來當作 extractor 的初始值參數，使 extractor 在訓練之前就存在著 Imagenet 上圖片的多種特徵。由於分類的目標改變為胸部 X 光圖像的疾病分類，我們必須在所有的模型架構的最後兩層接上 classifier，包含一層長度為 1024 的 dense layer 和一層輸出 14 種疾病機率長度為 14 的 dense layer，如 Fig. 1，最後一層的主要目的為將 14 種疾病做最妥善的分類。這次期末專題的訓練目標為 multi-label, i.e., 每一個 data 都可能同時罹患 14 種疾病的任一種，而不是 multi-class, i.e., 每一個 data 都只可能罹患 14 種疾病的其中一種，因此我們必須挑選適合的 loss function，而 binary cross entropy 是我們用來實現 multi-label 的訓練依據；相反的，若是將 loss function 誤選用了 categorical cross entropy，最後長度為 14 的 dense layer 則會預測出 14 個總和為 1 的疾病機率值，這樣就與本次專題訓練的目標不符。

我們針對 Neural Network 中所有 CNN Layers 和 classifier 中長度為 1024 的 dense layer 加上 dropout，其中 dropout rate 範圍大致落在 0.2 到 0.5 之間，此技巧主要是避免訓練模型過度擬合於訓練資料上。由於 classifier 中長度為 14 的 dense layer 的每一個 neuron 都會個別 output 其對應疾病的機率值，因此在此層加上 dropout 是不恰當的，此舉會在訓練過程中因為 neuron 的 dropout 而缺少機率值的輸出。

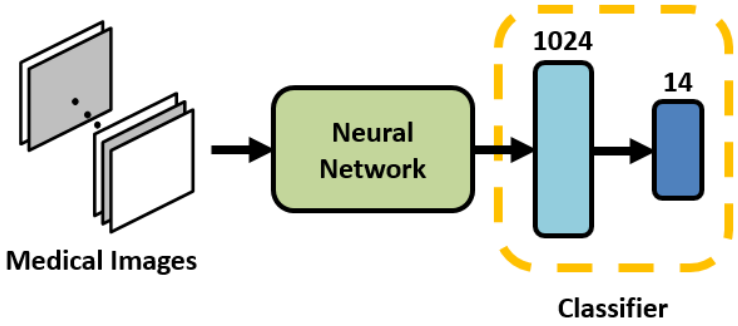


Fig. 1. General Model Structure

B. Regularization

由於我們所使用在 Imagenet 上的 InceptionResnetV2, InceptionV3, Xception, DenseNet201, Densenet121 的模型架構的層數都很多，多層數的模型架構擁有更多的參數可以完美的去擬合訓練資料，這是我們非常不樂見的狀況。另外，對於多層數的模型架構而言，其更容易擬合於少量的訓練資料上，而如何使用少量的訓練資料來預測大量的資料又是期末專題的主要目標，如果不適時的在每一層加上 regularization，在每一層加成情況下的結果如前述是非常容易造成對於訓練資料的過度擬合，並且訓練過程中的前 2~3 個 epoch 就達到 validation loss 的最低點。在實作中我們 extractor 中每一層以及 classifier 都加上 L2 norm，其值大致介於 0.005 到 0.05 中。L2 norm 又名 weight decay，因其在每次計算完 loss function 的梯度後都會再加上所有的 weights 的總和並將此乘上一小於 1 的數，此舉可以抑制過大權重的產生，強制使模型架構不要過度擬合於訓練資料上，適度的加上 regularization 除了能避免過度擬合訓練資料之外，因較平滑的擬合程度，對於模型架構未見過的 validation 以及測試資料的準確率上也會有很大的幫助。

C. Augmentation

在訓練資料中含有許多與正常的胸部 X 光圖像非常不一樣的資料，如 Fig. 2。由 Fig. 2 可以觀察出圖像大致有 Zoom (縮放)、Shift (位移)、Rotation (旋轉)、Brightness (明亮度)、Coarse Dropout (具有黑空格)、Salt & Pepper (胡椒鹽雜訊)，其中被模擬為 Coarse Dropout 的圖像是由於那位婦人身戴項鍊造成 X 光圖像上的雜訊，而模擬為 Salt & Pepper 的圖像中病人身上或身體內部含有凌亂的細管，因此用胡椒鹽雜訊去模擬此情況。因為在訓練過程當中這些比較奇異且相對少數的 X 光圖像一樣會當作輸入下去做訓練，如果沒有多增加這類類似的資料，模型對於較奇異的圖片是非常不熟悉的。尤其 augmentation 的目標是希望能用少量的具標籤資料去製造出類似的訓練資

料，以大家熟悉的手寫數字辨識為例，我們可以經由簡單的旋轉與縮放去模擬出不同人手寫數字的筆劃，同樣的在這次胸部 X 光圖像辨識疾病的專題中，我們也試著依觀察去模擬出類似的 X 光圖像的資料。另外，如前敘所述 Imagenet 上的模型架構層數大部分都很多，其實是需要相對大量的圖像來去做訓練，如果又受限於能訓練的 X 光圖像僅有 10002，增加資料量是必須的。增加資料量已經被視為機器學習中最有效的方法之一，尤其適用於高度複雜或學習能力極強的模型架構，隨機的 augmentation 也能使之避免過度擬合而正確的發揮其效用，由此可看出 augmentation 對於這次期末專題的重要性。

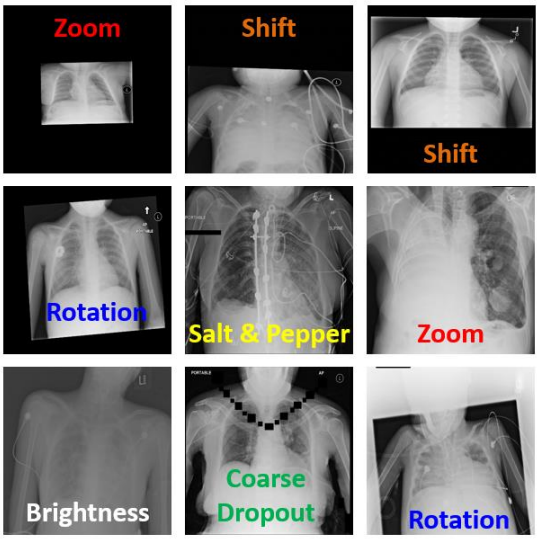


Fig. 2. Diversity of Medical Images

以上這 6 種情況是我們所觀察到大多數奇異的狀況，其中 Salt & Pepper 和 Coarse Dropout [5] 又比 Rotation、Zoom、Shift、Brightness 發生的機率低，儘管如此，訓練過程中調整 Shift 或者 Brightness 均會造成我們結果的準確率下降，因此這兩種 augmentation 的情況並未採納入我們最終 augmentation 的流程裡。由以上觀察到的結果，我們設計了一套 augmentation 的流程，如 Fig. 3，將原始 1024x1024 的 X 光圖像縮小成 256x256 的圖像，接著 256x256 的圖像有 100%的比例會做 Horizontal Flip、Rotation Range、Zoom Range，後端繼續接著做 RandomCrop，此功能也是 100%的比例，RandomCrop 後有 50%比例的圖像直接輸出成訓練資料，有 25%比例的圖像會加上胡椒鹽雜訊再輸出成訓練資料，有 25%比例的圖像會加上黑格子再輸出成訓練資料。

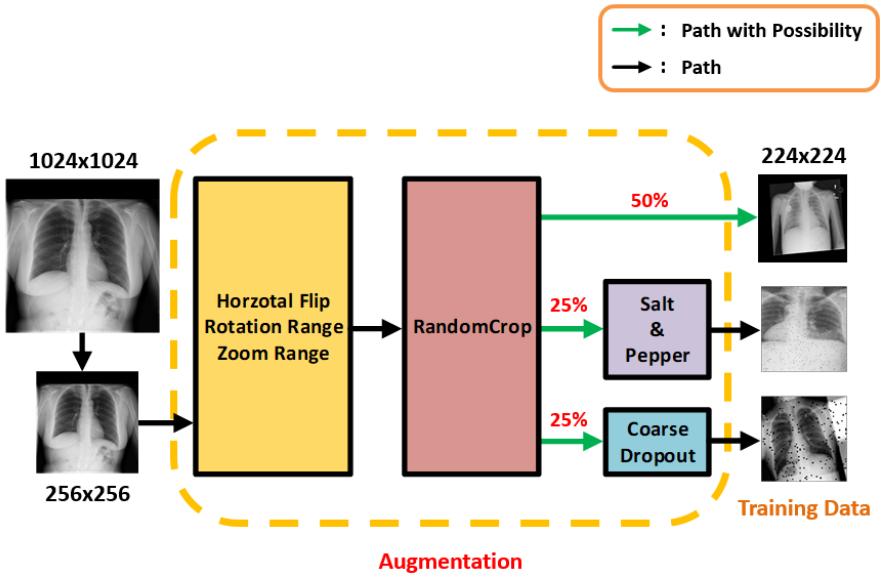


Fig. 3. Block Diagram of Augmentation

D. Class Weight

我們在訓練過程中發現 Hernia 在訓練的過程中很容易產生極大的震盪，如 Fig. 4，其原因為對於 14 種疾病如果都給予相同的權重訓練的話，發生機率低的疾病容易在訓練過程中產生很大的震盪。經由統計 10,000 筆訓練後，我們發現 Hernia 所發生的機率遠小於其他 13 種疾病，如 Table 1，Hernia 在 10,000 訓練資料中只發生了 22 次，而其他 13 種疾病發生的次數為 117~1666 次，發生機率次低的疾病其機率依然為 Hernia 發生機率的 5 倍以上。

很明顯的在這次的期末專題中 14 種疾病是非常不平均的，class weight 非常適合應付這樣的情況，因此我們給予 Hernia 的權重為 10，而其他 13 種疾病的權重為 1 或 2 來突顯 Hernia 的重要性。對於具有權重較大的疾病，模型一旦對此疾病預測錯誤其對於 loss function 的處罰也會較大，強迫模型在訓練中對於此疾病的重視度增加。

Disease	No. of 1 in 10k Training Data	Class Weight
Hernia	117~1666	10
Others	22	1~2

Table 1. Number of Diseases and Given Class Weight

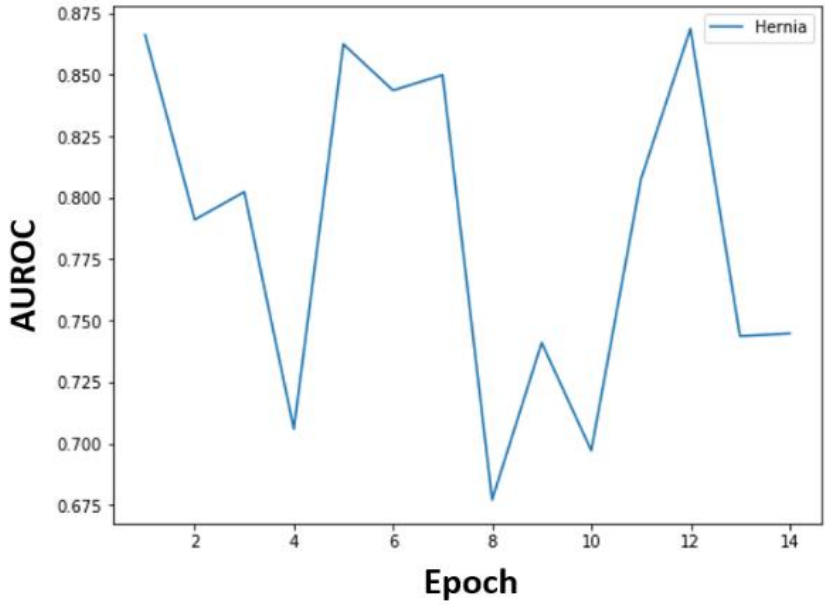


Fig. 4. Unstable AUROC of Hernia on Validation Set

E. Self-ensemble on Testing

在 testing 的階段，我們用了 Self-ensemble (N-Crop)的技巧，所謂的 N-Crop 是在 256x256 的影像範圍裡選取 N 個 224x224 的圖像，如 Fig. 5，這個方法出發點在於影像和影像之間的差異往往只是身體骨架在影像上位移的現象，因此利用此方法可以增加測試資料穩定性。因此在 testing 時，一張測試影像會經過 N-Crop 產生出 N 張影像再輸入訓練好的模型架構做測試並產生 N 條對於 14 種疾病的預測值，並把這 N 條預測值平均才是最終的預測結果。這個的行為與多個模型的 ensemble 是同樣的概念，將個別模型架構輸出的結果平均作為最終輸出結果，此舉亦可以減少 variance。套用此方法後，對於我們整體的 AUROC 表現提升了約 0.003，在實作上我們的 N 為 10，如將 N 繼續調大應有更好的表現。

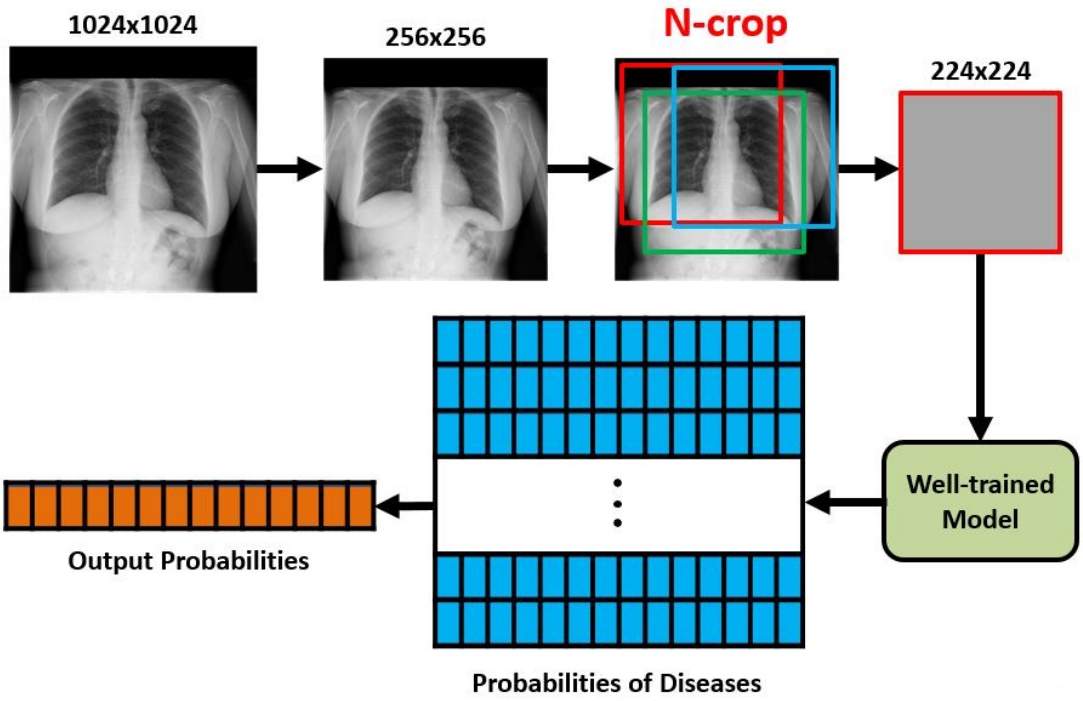


Fig. 5. Flow of Self-ensemble on Testing

F. Model Ensemble

Model ensemble 對於多個具高度複雜性的模型架構來說效果會很好，因為一個高度複雜的模型其 bias 會很小但 variance 會很大，而 model ensemble 的主要功能為縮小 variance，因此多個複雜的模型經由 model ensemble 過後就可以同時具有很小的 bias 以及很小的 variance。

在這次的期末專題中，我們使用了 Imagenet 上 InceptionResnetV2 (IRV2), InceptionV3 (IV3), DenseNet121 (D121)這 3 種不同的模型架構下做訓練，所 ensemble 的模型架構數量為 13，依照每個模型架構的強弱再給予不同的 ensemble 權重，由於每個模型都具有高度複雜度，與 model ensemble 適用前提非常吻合，也因此成為這次期末專題中的最大幫手。Table 2 為 model ensemble 中所使用的模型架構與對應的 AUROC (非同參數)。

Model	D121	D121	IRV2	IRV2	D121	D121	D121
AUROC	0.75876	0.76098	0.76140	0.75834	0.76500	0.76119	0.74961
Model	IRV2	IRV2	IRV2	IRV2	IRV2	IV3	
AUROC	0.76271	0.76262	0.75649	0.76990	0.77634	0.75284	

Table 2. Selected models and corresponding AUROC

其中 DenseNet 在近年已被證實在影像辨識比 Imagenet 上大多數的模型架構還要好[3]，在早年影像的辨識常用的架構裡通常都含有 CNN，而 DenseNet 簡單來說是一種強化版的 CNN，其正式名稱為 Densely Connected Convolutional Networks，依照複雜度的不同 DenseNet 的種類包刮 DenseNet121、DenseNet169、DenseNet201、DenseNet161，個別對應的架構如 Fig. 6。

Layers	Output Size	DenseNet-121($k = 32$)	DenseNet-169($k = 32$)	DenseNet-201($k = 32$)	DenseNet-161($k = 48$)
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 36
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 24
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Fig. 6. Model Structure of DenseNets

III. Trial Method

A. Pseudo-label

有鑑於這次的訓練資料有 10,002 筆具標籤的資料以及 68,466 筆不具標籤的資料，如何利用大量的未標籤資料成為設計模型架構的關鍵。這次期末專題同時給予具標籤的資料與未標籤資料，這是一個很典型的 semi-supervised 的問題，因此我們想先套用 semi-supervised 學習中最基礎的 pseudo-label 來嘗試實作這次的題目。

首先，為了觀察此 14 種疾病分布的情形，我們針對 10,002 筆具標籤的資料做統計，如 Table 3，其列述了此 14 種疾病在 10,002 筆具標籤的資料中各別被標示為 1 的比例。以表中第一項疾病 Atelectasis 為例，10,002 筆具標籤的資料中有 10.0%的資料被標示為得了 Atelectasis。由此統計結果我們假設在不具標籤的資料中每種疾病個別的得病比例與都 Table 3 相似。

基於這樣的假設下，我們將 68,466 筆不具標籤的資料餵入我們用 10,002 筆具標籤的資料所訓練好的模型架構，並記錄下所有不具標籤的資料經過 model 後所輸出對於此 14 種疾病得病的機率值，並將每個疾病對於不具標籤的資料的機率由大排到小，機率前幾大的測試資料其所對應的疾病將被標示為 1。以 Table 3 第一項疾病 Atelectasis 為例，在先前假設測試資料下也大約有 10.0%的 Atelectasis 的前提下，我們將不具標籤的資料中 Atelectasis 得病機率前 10.0%高的測試資料標示為 1，其餘則的則標示為 0。將所有不具標籤的資料中 14 種疾病都標示完成後，就得到了更多具標籤的資料來增加具有標籤的資料量，降低過度擬合原本少量具標籤的資料的機會。

Disease	**Atelectasis**	**Cardiomegaly**	**Effusion**	**Infiltration**	**Mass**	**Nodule**	**Pneumonia**
%	**10.0**	**2.5**	**11.0**	**16.7**	**4.8**	**5.5**	**1.2**
Disease	**Pneumothorax**	**Consolidation**	**Edema**	**Emphysema**	**Fibrosis**	**Pleural**	**Hernia**
%	**3.9**	**4.0**	**2.3**	**2.2**	**1.6**	**2.5**	**0.2**

Table 3. Percentage of Label 1 in 10k Training Data

B. Auto-encoder

除了在 semi-supervised 中最基礎的 pseudo-label，我們希望建立一個 auto-encoder，並將訓練好的 encoder 裡的 weights 當作我們訓練具標籤資料的 pre-trained weights。其做法為先將原本大小為 1024x1024x3 的圖像縮小為 224x224x3，再將其編成灰階圖片 224x224x1，經由 4 層 convolution layers 與一層 512 的 dense layer 將圖像轉成一 512 的 code，接者以完全相反的方式建立 4 層 deconvolution layer 並輸出一 224x224 的圖像，架構示意圖如 Fig. 7，而完整的架構表如

Fig. 8。我們希望輸出的 224x224 的圖像與原輸入的 224x224 可以有最小的 reconstruction error，而我們所採得 loss function 為每一個對應圖像位置上 pixel 的 mean square error。得到訓練好的 encoder 後，我們在其後再接上一新的 classifier，接著訓練具有標籤的資料針對 encoder 裡的 weights 做一個 fine tune。期望中間 512 的 code 能學習到這些 X 光圖像不同疾病的特徵。

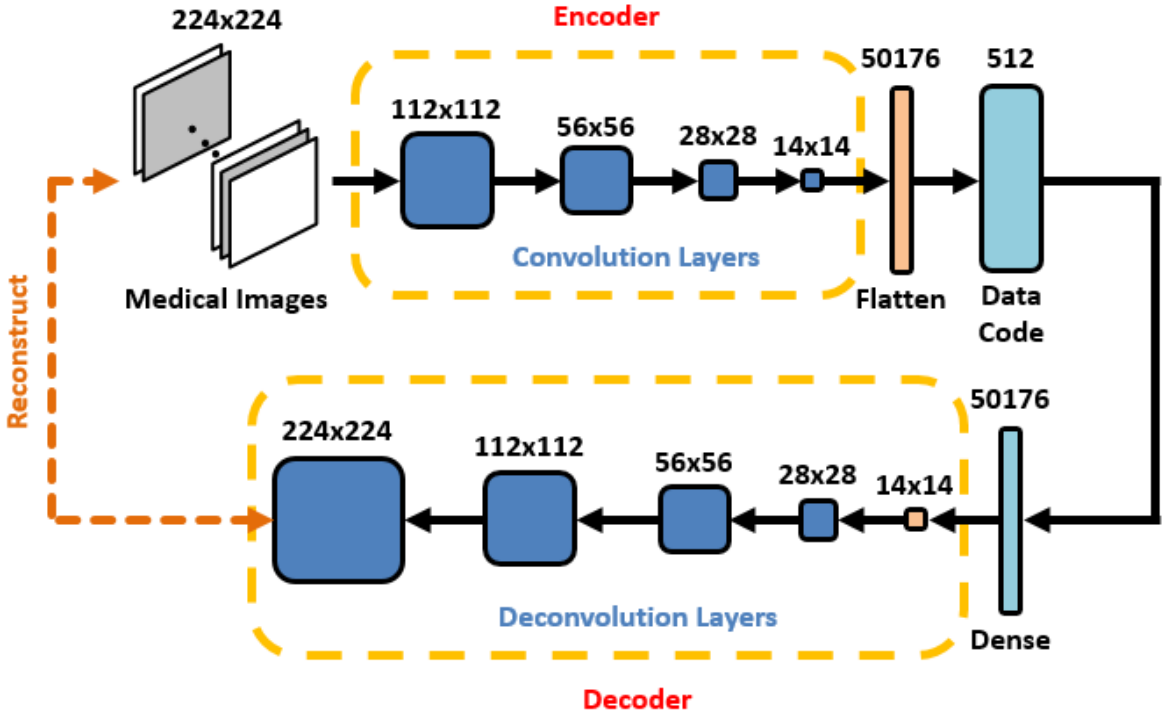


Fig. 7. Simple Architecture of Auto-encoder

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 1)	0
conv2d_1 (Conv2D)	(None, 224, 224, 64)	1664
average_pooling2d_1 (Average	(None, 112, 112, 64)	0
conv2d_2 (Conv2D)	(None, 112, 112, 128)	73856
average_pooling2d_2 (Average	(None, 56, 56, 128)	0
conv2d_3 (Conv2D)	(None, 56, 56, 128)	147584
average_pooling2d_3 (Average	(None, 28, 28, 128)	0
conv2d_4 (Conv2D)	(None, 28, 28, 256)	295168
average_pooling2d_4 (Average	(None, 14, 14, 256)	0
flatten_1 (Flatten)	(None, 50176)	0
dense_1 (Dense)	(None, 512)	25690624
dense_2 (Dense)	(None, 50176)	25740288
reshape_1 (Reshape)	(None, 14, 14, 256)	0
conv2d_transpose_1 (Conv2DTr	(None, 28, 28, 256)	590080
conv2d_transpose_2 (Conv2DTr	(None, 56, 56, 128)	295040
conv2d_transpose_3 (Conv2DTr	(None, 112, 112, 128)	147584
conv2d_transpose_4 (Conv2DTr	(None, 224, 224, 64)	204864
conv2d_transpose_5 (Conv2DTr	(None, 224, 224, 1)	1601

Total params: 53,188,353
Trainable params: 53,188,353
Non-trainable params: 0

Encoder

Layer	Filter Size		Channel Size
1	CNN	5x5	64
	AvgPool	2x2	
2	CNN	3x3	128
	AvgPool	2x2	
3	CNN	3x3	128
	AvgPool	2x2	
4	CNN	3x3	256
	AvgPool	2x2	

Decoder

Layer	Filter Size		Channel Size
1	CNN Trans.	3x3	256
2	CNN Trans.	3x3	128
3	CNN Trans.	3x3	128
4	CNN Trans.	5x5	64
5	CNN Trans.	5x5	1

Fig. 8. Architecture of Auto-encoder

C. Single Disease Training

先前我們提到的方法，不論是[1]的做法、pseudo-label 、auto-encoder 最後都只訓練出一個模型架構，接著輸出 14 種疾病的機率。於此我們做一個大膽的嘗試，我們將 14 種疾病以同一個模型架構下去做訓練，並得到 14 個針對不同疾病所訓練出的模型架構，我們期望一個模型架構可以更專精於某一種疾病的特徵學習，最後將這 14 個模型架構所輸出的答案連接在一起當作我們最後的答案。模型架構的更動如 Fig. 9，將模型架構中的最後一層長度為 14 的 dense layer 改為長度為 1 的輸出，但如上敘所述，此模型架構須對 14 種疾病個別的去做訓練。

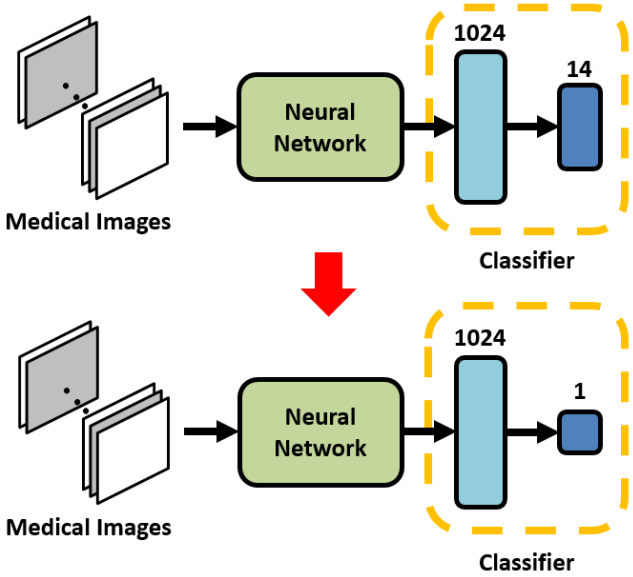


Fig. 9. Architecture of Single Disease Training

IV. Discussion

於第 III 部分我們嘗試以 pseudo-label 的方式來實作 semi-supervised 學習，然而我們得到的結果卻不盡理想。我們重新將訓練資料輸入模型架構當中，並依照與 pseudo-label 一樣的方式將所有訓練資料的影像依每種疾病個別去做機率由大到小的排序，接著分為 ground truth 為 1 或為 0 作圖，如 Fig. 10 與 Fig. 11，此舉可用來觀察此模型架構對於標籤 0 與標籤 1 的區分能力。

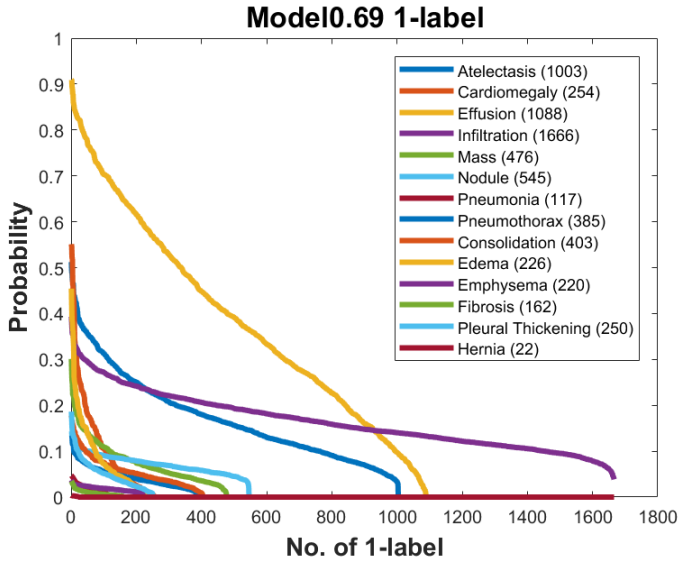


Fig. 10. Label 1 Distributions on Training Data

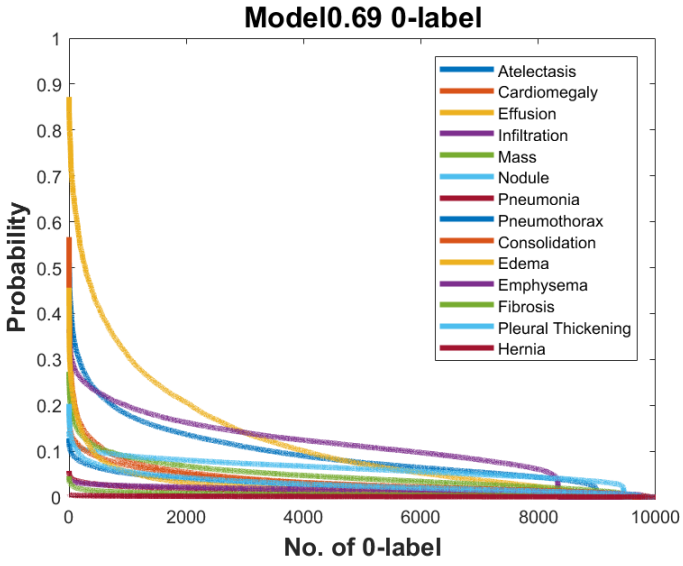


Fig. 11. Label 0 Distributions on Training Data

很意外的，我們發現兩者的分布狀況竟然是相似的，並沒有一個明顯的機率門檻去區分標籤 0 與標籤 1，如果以我們之前使用 Table 3 去取機率較大的部分當作 pseudo-label 的依據勢必會很大的誤差。以 Effusion 為例，Fig. 10 及 Fig. 11 最高的黃色線，我們發現在標籤為 0 的資料其部分的機率比部分標籤 1 的機率來的高，而同樣的道理，標籤為 1 的資料其部分的機率比部分標籤 0 的機率來的低，因此在做 semi-supervised pseudo-label 時，可能會有部分甚至更多的未標籤資料原本沒有疾病被標為 1 或是原本有疾病被標為 0，導致我們將這些誤判的資料加入至訓練資料當中。也由於這些 pseudo-label 的資訊均來自某個模型，最後所得到的最終結果的表現與原本只做 supervised 的 AUROC 差不多，甚至再低一點。

前面我們也實作了 auto-encoder，auto-encoder 的優點在於模型架構能夠訓練過所有的資料，資料量的增加在機器學習的領域中是避免過度擬合最簡單的作法，也因此我們可以透過 auto-encoder 實作 unsupervised 學習達到減少過度擬合的效果。Fig. 12 為四組還原前後的 X 光圖像，可以發現經由還原過後的圖片雖然有些模糊但大致上是可以與原圖對在一起的。

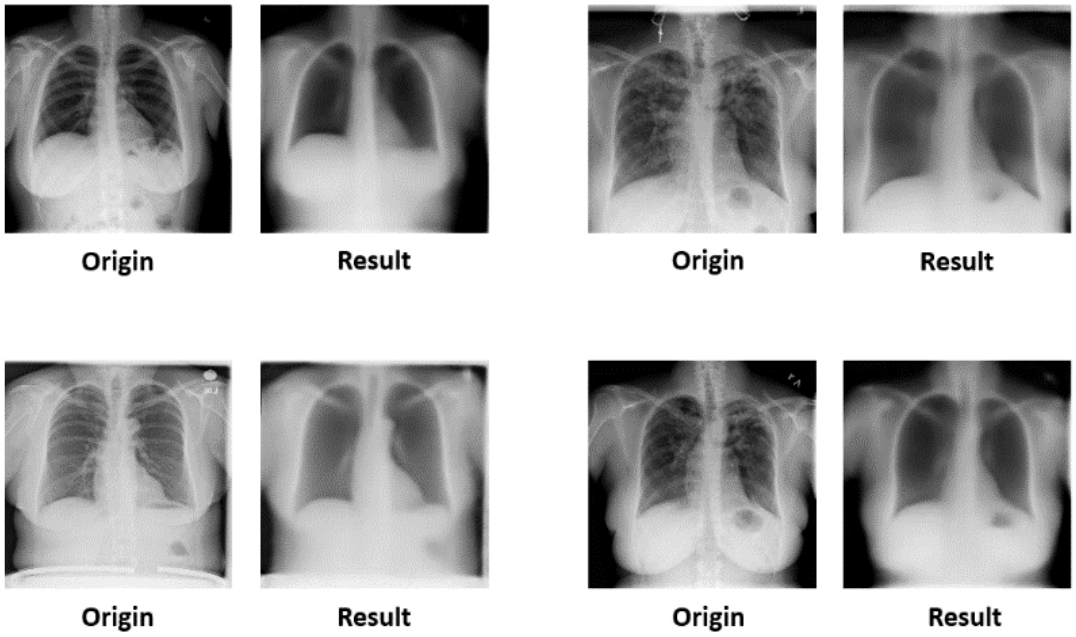


Fig. 12. Reconstructed Images of Auto-encoder

然而 auto-encoder 的最終結果並不如預期，可能的原因為當初在使用 autoencoder 時放入的圖片都非常的類似，autoencoder 不一定能學到 14 種疾病明顯的特徵或差異性，在 autoencoder 沒辦法明確分類目標的狀況下，很容易學習到依些較明顯的特徵，例如胸腔大小、骨骼形狀、影像灰階程度等等，而對於 14 種疾病的分類學習較淺。

我們在第 III 部分的最後一個嘗試的方法為 single disease training，其表現也與我們的預期分數有一段落差，甚至只需要單一模型架構就可以超越此方法的表現。會有這樣的結果主要是我們疏忽了疾病之間並非獨立的，如果以同一種模型架構但各個疾病分開來訓練，會造成我們人為的去阻止了模型對於疾病關聯性之間的學習，進而造成最後表現不佳的結果。Fig. 13 為[2]對於 14 種疾病中 8 種疾病針對疾病之間的關聯性作圖，由此可見疾病之間關連性的重要性。

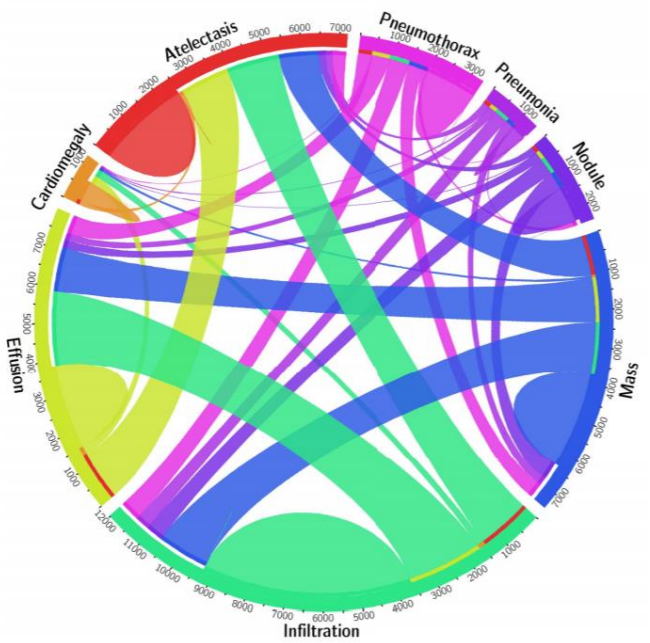


Fig. 13. The circular diagram shows the proportions of images with multi-labels in each of 8 pathology classes and the labels' co-occurrence statistics

V. Conclusion

這次期末專題的胸腔 X 光圖像疾病辨識，首先我們嘗試了最簡單的 transfer learning 方法，將 10,002 筆具標籤的資料訓練在後端有兩個長度分別為 1024 和 14 的 dense layers 當作 classifier 且具有 pre-trained weight 的 extractor 上，在此方法下可以發現過度擬合的情況極為嚴重，於此我們特別對訓練資料做一個深入的探究，包含 1) 具標籤資料中各個疾病標籤 1 和標籤 0 的個數與具標籤資料與未具標籤資料數量的差距、2) 訓練資料的多樣性、3) 過度擬合與 dropout 和 regularization 之間的關係、4) 14 種疾病之間的關聯性以及得病的比例。

由 1) 我們嘗試使用 semi-supervised learning 以及 unsupervised learning 來增加具標籤資料的資料量與增加對於訓練資料的特徵學習以降低過度擬合的情況。在 semi-supervised learning 中，我們統計 14 種疾病在具標籤資料中個別得病的比例，並將未標籤資料依照此比例標示為標籤 1 與標籤 0，再重新加入具標籤資料之中。在 unsupervised learning 中，我們利用 encoder 對訓練資料所有 224x224x1 的影像降維以提取影像的特徵，再藉由 decoder 將降維後 512 維的 code 還原成 224x224x1 原影像大小，最後再將 encoder 當作 pre-trained weight 的 extractor，後端再接上 DNN classifier。1) 的兩種做法都呈現了對於未具標籤資料的善加利用，能繼續改良的作法呈現於第 VI 部分。由 2) 我們套用不同 augmentation 的方法來模擬資料的多樣性並同時增加具標籤資料的資料量，採用的 augmentation 方法包含平移、旋轉、錯切、縮放、salt & pepper、coarse drop out。由 3) 我們挑選適當的 dropout rate，目的是在訓練過程中限縮模型架構對於訓練資料的預測的能力，一個好的 dropout rate 可以在 validation 以及測試資料的預測準確率上明顯的進步，而一個好的 regularization 參數除了可以達到適當的 weight decay 效果亦可以避免模型架構過度擬合於訓練資料，對於測試資料可以有更好的預測結果。由 4) 我們觀察到不同疾病標籤 1 的數量差距甚大，為了避免模型架構因為標籤 1 的多寡而對於不同的疾病預測效果的優劣，我們將每一種疾病個別去做訓練，希望可以更準確的提取出不同疾病的特徵，最後再將 14 個模型架構對於不同疾病所預測的結果連接起來當作最終的預測結果。在忽略疾病之間具有關聯性[2]的情況之下所得到的預測結果不如預期，因此我們利用 class weight 來強化模型架構對於發生率較低的疾病的學習以彌補標籤 1 分佈不平衡的問題，另外，我們觀察了各個疾病在每一個 epoch 中 AUROC 震盪的情形，將得病比例較少或是 AUROC 震盪劇烈的疾病在訓練過程中給予較大的 class weight，藉此穩定每一次 epoch 的梯度下降。

經由以上的方法、觀察與改進，我們訓練出的模型架構都能對於疾病的判別有很高的準確率。期望此方法能對於醫療領域與全人類的健康福祉帶來很大的效益。無論這些方法最終是否有被採用，我們仍會不斷的嘗試與貢獻。

VI. Future Work

在未來工作裡我們希望針對 semi-supervised 學習的方法去做更多的鑽研，其中 pseudo-label 的部分，未來我們希望對標籤 0 以及標籤 1 的各個疾病都設定一個適合的門檻值，例如規定機率小於 0.2 的部分以及機率大於 0.7 的部分才可以從未具標籤的資料裡重新標記成為標籤 0 和標籤 1，並把這些資料加入訓練資料中重新 fine tune 出一個新的模型架構，再餵入新的未具標籤的資料，其中部分的未具標籤資料又會被加到訓練資料中，直到大部分的未具標籤資料都放入訓練資料中為止。而兩個機率的門檻值則必須選的好，門檻值設定的鬆或緊都必須考量訓練的時間以及準確度。

另外在 auto-encoder 的部分，可能改進的作法有以下 3 種 1) 套用 randomcrop 的概念。在 224x224 的影像中每次隨機的位置擷取一個方框，此舉可縮小降維的目標範圍，使 auto-encoder 盡量不要學習到骨骼、性別、胖瘦等等過於明顯且大面積的影像特徵，而著重於小區域的疾病特徵。2) 考慮不同降維後 code 的大小。較大的維度可以記錄圖像中較多的特徵，而較小的維度亦可學得更精髓的特徵，如何針對這次判斷胸腔 X 光疾病選擇適合的 code 大小是非常關鍵的。3) 改用的 Imagenet 上的 pre-trained model 當作 encoder，利用 pre-trained weight 作為 encoder 的參數初始值，在這樣的加持下也許就能夠提取對於疾病的判別更具有鑑別度的影像特徵。

VII. Acknowledgements

[1]的 source code 是用 pytorch 所撰寫，我們所參考的部分為[4]的 Github code 內容，[4]為[1]的 Keras 版本，使用的部分包含 AUROC 的計算、Imagenet model 選用、Early stop 機制。另外，也非常感謝[5]對於 IAA 每個 augmentation 功能圖示上清楚的展示

REFERENCES

[1] P. Rajpurkar et al. (Dec. 2017). “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning.”

[2] W.Xiaosong et al. (Dec. 2017). “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.”

[3] Gao Huang et al. (Aug. 2016). “Densely Connected Convolutional Networks.”

[4] https://github.com/brucechou1983/CheXNet-Keras?fbclid=IwAR1FOXhlK2JbCQZFNfgfrNPBuRQqwu_MIcIOBcT4w_KH1vmATyzzhbsPkqk

[5] https://github.com/aleju/imgaug

Yi-Lin Lo was born in Taipei, Taiwan, in 1996. He is currently a graduate student in electronics engineering at National Taiwan University (NTU).

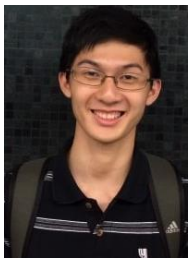
His research interests include energy-efficient VLSI design for signal processing with applications in clutter rejection filtering for color flow imaging and ultrasonic elastography imaging.

Jian-Chen Su was born in Changhua , Taiwan , in 1995 . He is currently a graduate student at National Taiwan University (NTU).

His research interests include algorithm design in deep image colorization and topics in Computer Vision.

Kai-Hsiang Liu was born in Taichung, Taiwan, in 1995. He is currently a graduate student in communication engineering at National Taiwan University (NTU).

His research interests include algorithm design in communication and computing resource allocation of mobile edge computing and small cell network in future 5G system.



Bing-Yuan Tzeng was born in Taoyuan, Taiwan, in 1996. He is currently a graduate student in computer science & information engineering at National Taiwan University (NTU).

His research interests include algorithm design in optical character recognition and computer vision.