Small Data Training for Medical Images

王柏翔 r07922018
r07922018@ntu.edu.tw
National Taiwan University

陳　毅 r07922059
r07922059@ntu.edu.tw
National Taiwan University

顏百謙 r07922135
r07922135@ntu.edu.tw
National Taiwan University

ABSTRACT

為了透過少量的資料，來判定NIH Chest X-ray資料集中相片罹患各項疾病的機率，本組使用機器學習的方式，在基於DenseNet架構的pretrain model下進行訓練，使用不同的training set以及DenseNet架構產生大量強項不同的model，並且使用PCA技術利用其他model對主要model進行微調，以產生更好的機率分佈。最終利用此方法將17個Kaggle Public最高分數為0.75372的model進行PCA合併之後Kaggle分數提升至0.78984。

KEYWORDS

datasets, neural networks, NIH Chest X-ray, Medical

ACM Reference Format:
王柏翔 r07922018, 陳　毅 r07922059, and 顏百謙 r07922135. 2019. Small Data Training for Medical Images. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

INTRODUCTION

在現代醫療領域中，為了觀察病人的病徵，醫生們會借助各種精密儀器來拍攝大量醫學影像，來判斷病人的病徵，例如：X光片、電腦斷層掃描（Computerized tomography, CT）、正子發射斷層掃描（PET, Positron emission tomography）等。

若能**透過機器學習以及這些大量的醫學影像，建立專門用於判斷病徵的模型**，就有機會可以節省醫生問診的時間（醫生不需要親自去解讀醫學影像，只需要知道病徵，然後給予處方即可）。

DATA PREPROCESSING / FEATURE ENGINEERING

資料描述

(1) Training data

共有78468筆資料，其格式如下表。

Image Index	Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position
00001522_000.png	0 0 0 1 0 ...	0	1522	50	M	PA
00001522_001.png	1 0 0 1 0 ...	1	1522	50	M	PA
00001522_002.png	1 0 0 0 0 ...	2	1522	50	M	AP
00001523_000.png		0	1523	51	F	PA
00001523_001.png		1	1523	51	F	PA

Figure 1: Training data前7個欄位

共有9個欄位，分別為

- Image Index：醫療影像的檔案名稱。
- Labels：病徵的標示，共有14個0/1以空白相隔，分別代表是否罹患某病徵，0代表無病徵，1代表有病徵。若該醫療影像未經過人工標示，則此欄位為空欄位。
- Follow-up #：追蹤次數。
- Patient ID：病人的編號。
- Patient Age：病人的年齡。
- Patient Gender：病人的性別，M代表男性，F代表女性。
- View Position：查看位置。
- OriginalImage[Width,Height]：原始影像的大小。
- OriginalImagePixelSpacing[x,y]：原始影像的像素間距。

Label index	病徵（英文）	病徵（中文）
0	Atelectasis	肺膨脹不全
1	Cardiomegaly	心臟肥大
2	Effusion	胸腔積液
3	Infiltration	肺浸潤
4	Mass	腫塊
5	Nodule	肺結節
6	Pneumonia	肺炎
7	Pneumothorax	氣胸
8	Consolidation	一處或多處的實變
9	Edema	水腫
10	Emphysema	肺氣腫
11	Fibrosis	纖維化
12	Pleural Thickening	胸膜增厚
13	Hernia	肺突出、肺疝氣

Table 1: 病徵對照表

Follow-up #	人數
0	21528
1	9317
2	6404
3	4955
4	4010
5	3363
6	2804
7	2377
8	2041
9	1766
10次以上	19903

Table 2: 追蹤次數人數對照表

(2) Testing data
共有33652筆資料，每一筆資料都僅有一個欄位：醫療影像的檔案名稱。

(3) Submission format
模型要根據輸入的醫療影像，預測該病人得到各種病徵的機率（共14種病徵）。

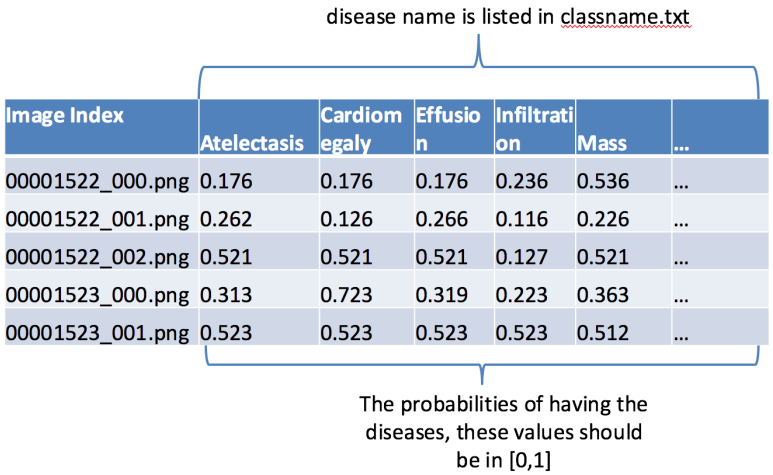

Figure 2: Submission format

資料分析

(1) Labels：78468筆資料中，共有68466筆未標籤的資料，10002筆已標籤的資料。其所表示的病徵如下表
Figure 3為他人製作的各病徵常見程度的統計表，其中，Infiltration、Effusion以及Atelectasis較為常見，而Hernia較為罕見。

(2) Follow-up #：統計不同追縱次數的人數，結果如Table 2。約略一半的人，最多只會追蹤3次以下。

(3) Patient Age：平均值46.77，標準差16.93。

(4) Patient Gender：男女比例約為14:11。

Preprocess method

(1) 把原始圖片從 1024*1024 降低解析度成 224*224 。

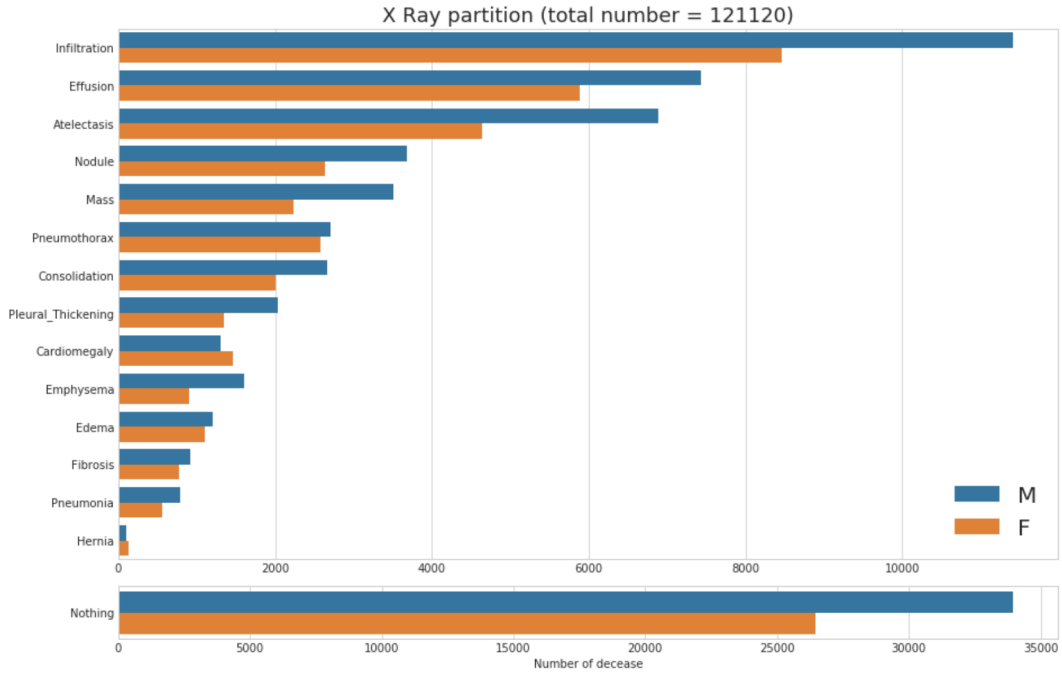

Figure 3: 病徵常見程度統計表

(2) 使用 ImageDataGenerator 產生單一批次用的訓練用資料集（batch size = {6, 8}）。
(3) 將產生的圖片透過 keras 提供的 preprocess_input 方法，產生適合 DenseNet pre-train model 的圖片。

Feature engineering

雖然原始資料中提供了大量的病人資料，但我們最終版本的 model 中，只用到了病人的肺部資料以及 其 label ，直接進行 supervised learning 。

MODEL DESCRIPTION

關於使用胸部X光進行機器學習的論文，在網路上可以找到一些已經被使用過的網路架構，如：CheXNet[7]、XNet[8]等，根據 ChestXRay [1] 的模型架構所示，進行肺部病徵的模型訓練，直接拿現有的 pre-trained models 來使用，就可以得到不錯的結果，因此我們使用了 keras 所提供的各式 pre-trained models 來進行預測。 以下為 keras 內，所提供的由 ImageNet 資料集訓練而成的 pre-trained models ：

- Xception
- VGG16
- VGG19
- ResNet50
- InceptionV3
- InceptionResNetV2
- MobileNet
- DenseNet
- NASNet
- MobileNetV2

其中，發現僅有 DenseNet121, DenseNet169, DenseNet201 共三個模型，較適合作為本次期末專案的模型。
而其實際組織起來的架構如 Figure 4。

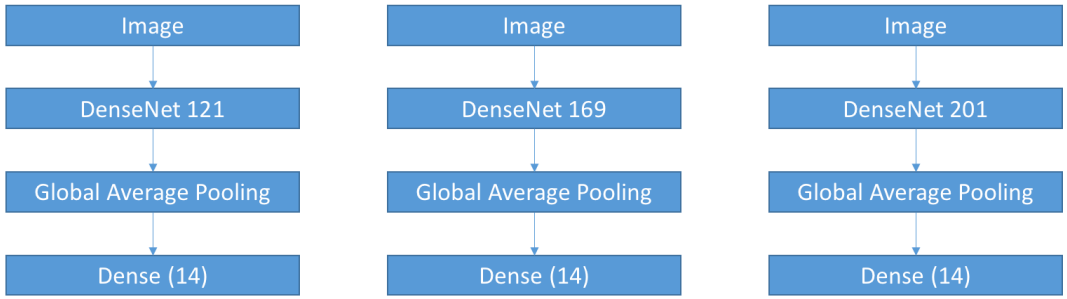

Figure 4: ModelArchitecture

EXPERIMENT AND DISCUSSION

在這次的期末專案中，一個模型的預測分數是有其上限的，因此我們要透過 ensemble 的方式來使分數得到提升，關於 multi-label 以及 probability 的問題，一般都是使用 stack-classifier [3, 12]來處理。但在這次的期末專案中，我們提出一個有趣的方法：用 PCA 來進行多模型預測結果的整合。

我們提出了 PCA unit ，如 Figure 5 所示， PCA 一般是用在資料特徵的降維工作，當我們

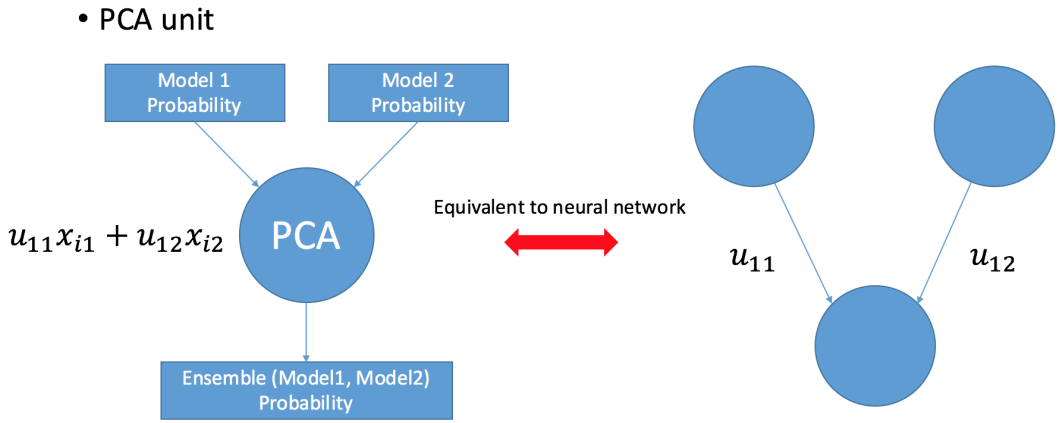

Figure 5: PCA Unit

將兩個 models 預測出的結果當作一筆資料的兩個特徵，那麼使用 PCA 可以幫助我們找到一個最適合的投影軸，使得所有資料的預測分佈是最為平均的，如 Figure 8。

假設今天有一個 model A 的預測結果，在機率為 0~0.3 和 0.7~1.0 的區間中，預測結果相當不錯，但在機率區間為 0.3~0.7 時，0和1的資料卻是無序的，也就是 0.3~0.7 之間，是沒辦法分清楚誰是0，誰是1。

此外，有另外一個 model B ，它給出了另外一個預測結果，但是他的分數比 model A 低，那我們會希望，能夠在 model A 的 0.3~0.7 的區間中，用 model B 的預測結果，給點加權，讓0的結果往0.3偏移，1的結果往0.7偏移。

因此，這時候，這 model A 與 model B 預測的結果希望能夠有一定的相關性，並希望圖中偏向(0,1)和(1,0)的資料盡量少。可是，考量到兩個 models 原先預測出來的結果分佈可能差異性過大，因此若直接平均加權，會導致原先分佈較不平均的 model 大幅影響整個結果（如Disease 10的 x model）。因此使用 PCA 能夠幫助我們較好的找到一個好的投影軸（綠

線）。

使用 PCA 進行 ensemble ，能夠大幅提高預測準確率。 Figure 9 與 Figure 10分別為兩個 models 進行 ensemble 和三個 models 進行 ensemble 得到的結果。

CONCLUSION

Figure 6為未做 ensemble 前的所有 models 的準確率。

model 1 (3)	model 2 (5)	model 3 (7)	model 4 (10)	model 5 (1)	model 6 (4)
DenseNet 169 1 epoch 0.72219 | 0.73173	DenseNet 169 2 epoch 0.74359 | 0.74140	DenseNet 169 1 epoch 0.71724 | 0.72046	DenseNet 169 2 epoch 0.74659 | 0.74897	DenseNet 169 2 epoch 0.72339 | 0.72643	DenseNet 169 1 epoch 0.74489 | 0.73860
model 7 (3)	model 8 (4_STB)	model 9 (4)	model 10 (6_STB)	model 11 (8)	model 12 (5)
DenseNet 121 2 epoch 0.75234 | 0.74607	DenseNet 121 2 epoch 0.75360 | 0.74985	DenseNet 121 2 epoch 0.74214 | 0.73271	DenseNet 121 1 epoch 0.72549 | 0.72118	DenseNet 121 2 epoch 0.73593 | 0.73182	DenseNet 121 1 epoch 0.73939 | 0.73795
model 13 (10)	model 14 (9)	model 15 (2)	model 16 (3)	model 17 (1)	
DenseNet 121 2 epoch 0.74832 | 0.74337	DenseNet 121 2 epoch 0.75089 | 0.75281	DenseNet 201 2 epoch **0.75372 | 0.75222**	DenseNet 201 2 epoch 0.73137 | 0.73622	DenseNet 201 2 epoch 0.73392 | 0.73728	(public, private)

Figure 6: PCA前各個model於Kaggle上的分數

使用Figure 11、Figure 12、Figure 13、Figure 14、Figure 7的結構，可以將分數提高到0.79附近。

ACKNOWLEDGMENTS

謝謝助教提供的種種協助，讓我們能夠找到人生未來的方向。
感謝 DeepQ 提供的 GPU ，讓我們擁有充分的運算能力可以去跑大量的 models 。
此外，我們也在網路上搜尋到他人所做的相關研究，讓我們的期末報告能夠順利進行。[1–12]

REFERENCES

[1] ChestX-ray8 2018. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. https://arxiv.org/abs/1705.02315.
[2] Classify with Convolutional Neural Network 2018. Classification of cardiomegaly using Convolutional Neural Network. https://www.linkedin.com/pulse/classification-cardiomegaly-using-convolutional-neural-anoop-singh.
[3] Ensemble modelling using model's probabilities 2018. StackExchange - Ensemble modelling using model's probabilities. https://datascience.stackexchange.com/questions/17837/ensemble-modelling-using-models-probabilities.

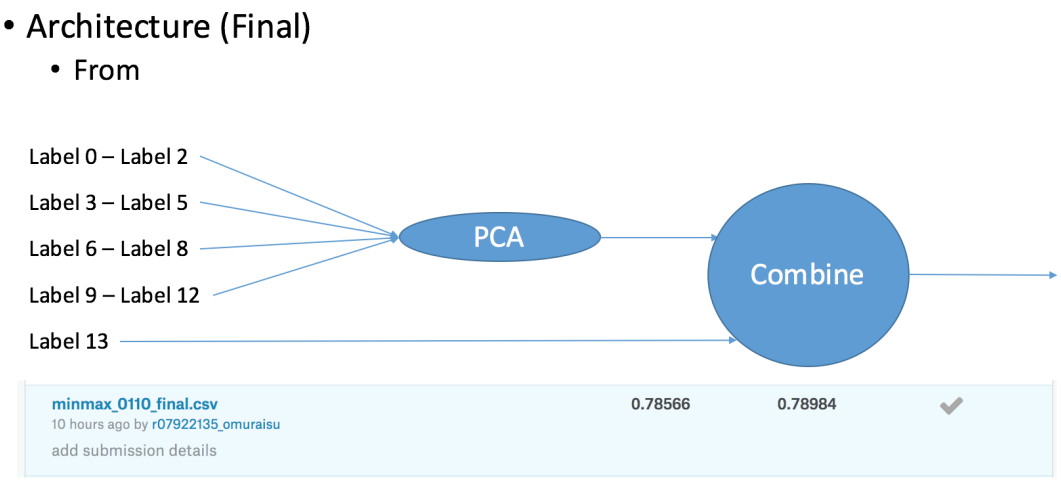

Figure 7: Final

[4] Ensembles through Metalearning 2018. AUC-Maximizing Ensembles through Metalearning. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4912128/.
[5] GitHub 2018. anoopsin/udacity-machine-learning-engineer-nanodegree/capstone_classifying_x_rays. https://github.com/anoopsin/udacity-machine-learning-engineer-nanodegree/tree/master/capstone_classifying_x_rays.
[6] GitHub 2018. gregwchase/nih-chest-xray. https://github.com/gregwchase/nih-chest-xray.
[7] GitHub 2018. uci-cbcl/ChestXRay. https://github.com/uci-cbcl/ChestXRay.
[8] GitHub 2018. zoogzog/chexnet. hthttps://github.com/zoogzog/chexnet.
[9] Kaggle (kernel) 2018. Cardiomegaly Pretrained-VGG16. https://www.kaggle.com/kmader/cardiomegaly-pretrained-vgg16.
[10] Lung cancer prediction 2018. Lung cancer prediction using machine learning and advanced imaging techniques. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6037965/.
[11] Mata Classifier 2018. A Meta Classifier by Clustering of Classifiers. https://link.springer.com/chapter/10.1007/978-3-319-13650-9_13.
[12] mlxtend 2018. StackingClassifier. http://rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/.

APPENDIX - FIGURE

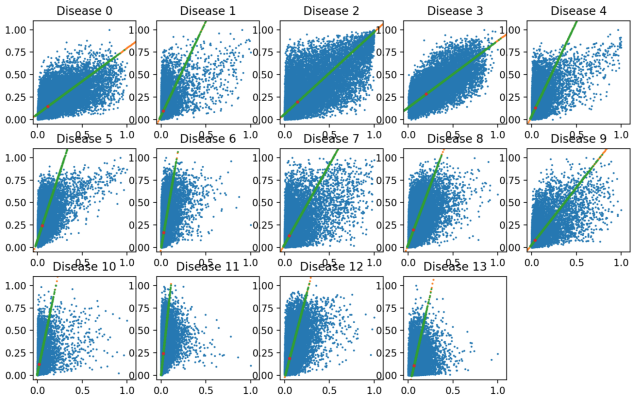

Figure 8: PCA (2D)

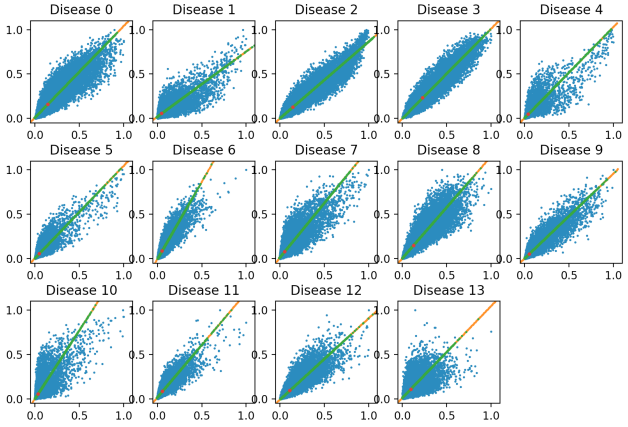

Figure 9: PCA (2D)

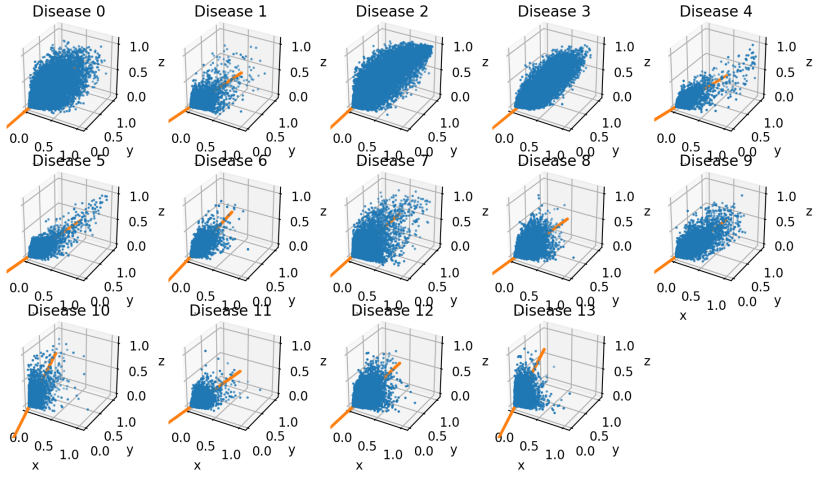

Figure 10: PCA (3D)

- Architecture (label 0 – label 2)

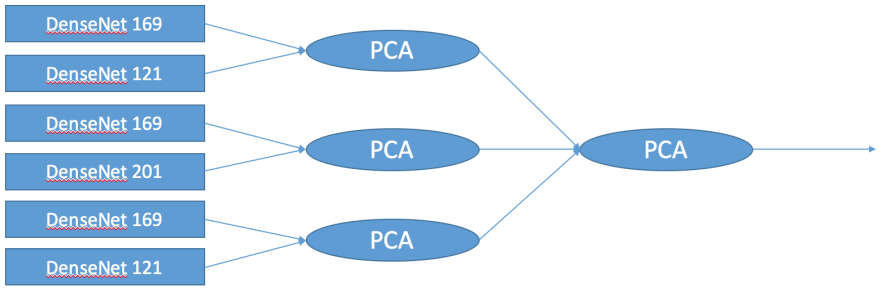

Figure 11: Label 0 - Label 2 (PCA)

• Architecture (label 3 – label 5)

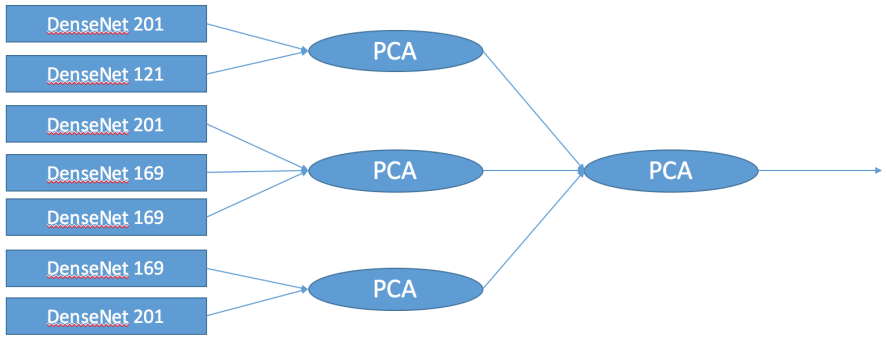

Figure 12: Label 3 - Label 5 (PCA)

• Architecture (label 6 – label 8)

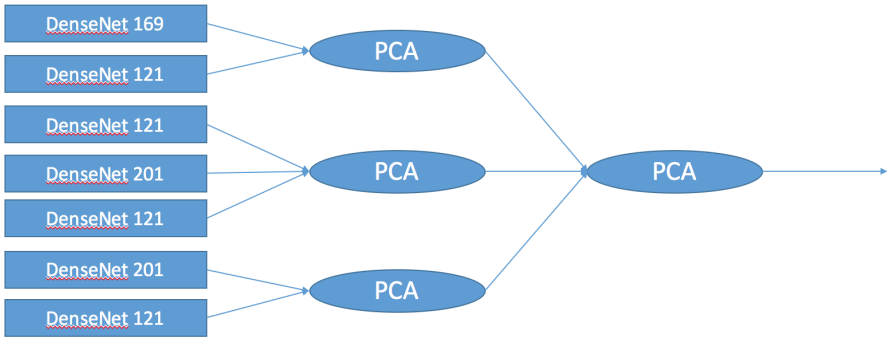

Figure 13: Label 6 - Label 8 (PCA)

- Architecture (label 13)

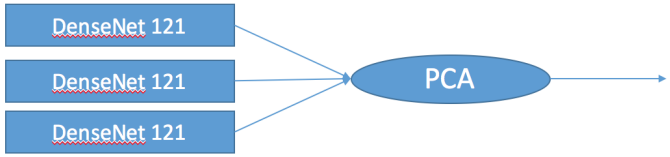

Figure 14: Label 13 (PCA)