Final Project - Small Data Training for Medical Images

R07922018 資工所碩一 王柏翔
R07922059 資工所碩一 陳 毅
R07922135 資工所碩一 顏百謙

December 13, 2018

隊名及隊員

隊名：NTU_r07922018_雷姆包伯歐姆蛋
隊員

1. 王柏翔：R07922018 資工所碩一
2. 陳 毅：R07922059 資工所碩一
3. 顏百謙：R07922135 資工所碩一

所選擇的題目

Small Data Training for Medical Images

Problem Study

問題描述

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）。

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

   (a) Image Index：醫療影像的檔案名稱。
   (b) 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。

(c) Follow-up #：追蹤次數。
(d) Patient ID：病人的編號。
(e) Patient Age：病人的年齡。
(f) Patient Gender：病人的性別，M代表男性，F代表女性。
(g) View Position：查看位置。
(h) OriginalImage[Width,Height]：原始影像的大小。
(i) OriginalImagePixelSpacing[x,y]：原始影像的像素間距。

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）。

disease name is listed in classname.txt

Image Index	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	...
00001522_000.png	0.176	0.176	0.176	0.236	0.536	...
00001522_001.png	0.262	0.126	0.266	0.116	0.226	...
00001522_002.png	0.521	0.521	0.521	0.127	0.521	...
00001523_000.png	0.313	0.723	0.319	0.223	0.363	...
00001523_001.png	0.523	0.523	0.523	0.523	0.512	...

The probabilities of having the diseases, these values should be in [0,1]

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

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

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。

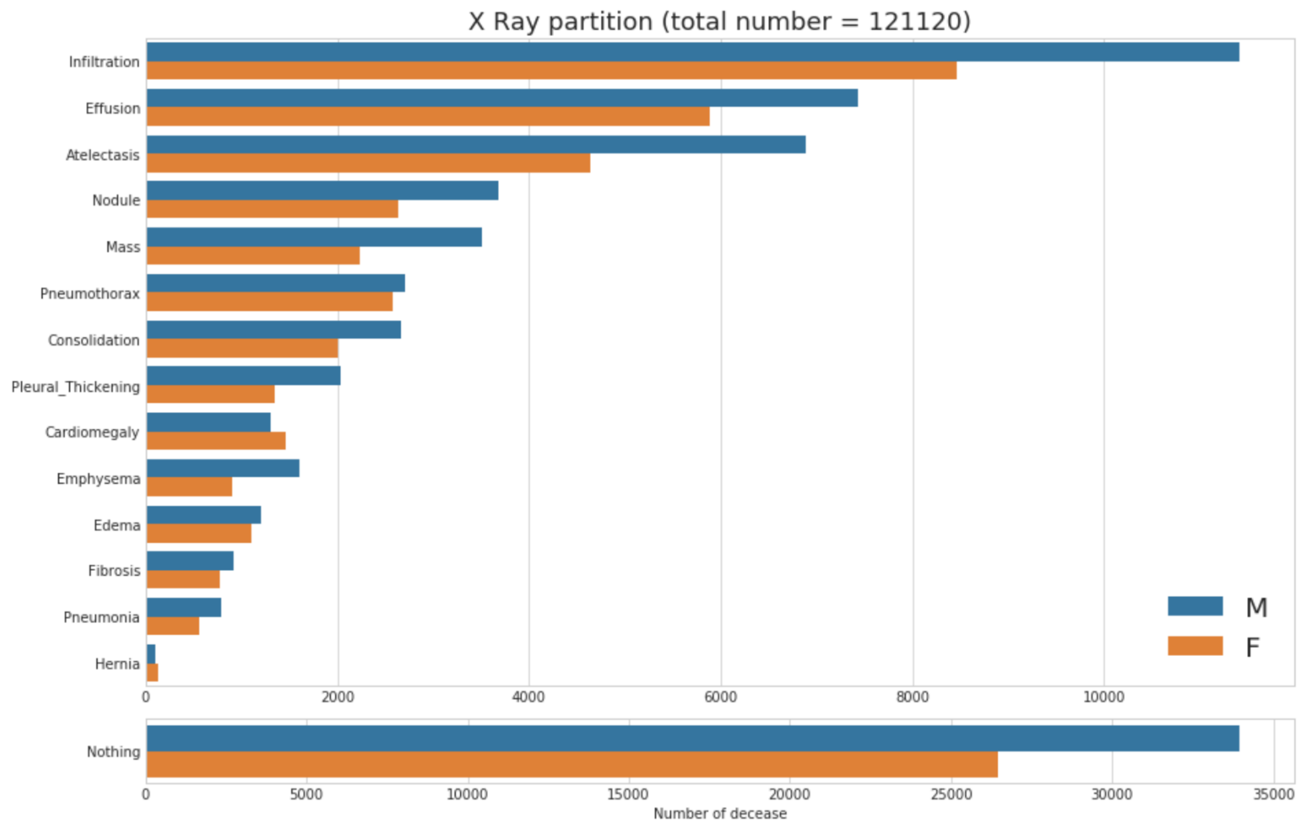

Figure 3: 病徵常見程度統計表

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。

待做分析

1. 病徵間相關性：併發症的可能性。
2. 病徵與追蹤次數的關聯性：不同病徵需要追蹤的次數可能不同。
3. 病徵與年齡的關聯性：不同年齡的人，所會罹患的病徵可能不同。

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

參考資料

1. **ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases**
   - Dataset：有8種症狀，共108948張胸腔X光圖，其中84312張無症狀，剩下24636張至少有一個症狀。每一張圖都有標明症狀(如果有)或是無症狀。其中有983張圖帶有標示症狀特張位置的bounding box，但是不用做training資料之用，僅用來確認ground truth。
   - Model：使用multi-class DCNN，其結構如下：
     (a) Pretrained model：使用ImageNet來pretrained過的model共四個，AlexNet、GoogLeNet、VGGNet-16以及 ResNet-50，移除各model中，Fully Connected層後的所有layers。其中，ResNet-50辨識率較高。
     (b) Transition layer: 統合各種pretrain model的convolution層為統一shape的層, shape=(S, S, D)，其中S=8, 16, 32. D為各pretrain model最後的層數。
     (c) Global pooling layer: 直接省略Fully-Connected，convlution的各層變為單一點，輸出shape=D. 使用LSE(Log-Sum-Exp), r=10時效果最佳。
     (d) Prediction layer: 將D維轉為C (the number of classes) 維，預測各class的症狀機率多寡。Transition layer的輸出和這一層的權重相乘會得到一組表示各症狀在data中各病症的likelihood heatmap，有peak的地方代表有較高的機率有症狀。
     (e) Loss layer: Cross Entropy Loss，但是由於資料的positive/negative不均，使用weighted CEL.

Pre-trained Models

Keras內建提供了許多由ImageNet資料集的pre-trained models，如下：

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

而關於使用胸部X光進行機器學習的論文，也提到一些可使用的網路架構，如：CheXNet、XNet等。

- CheXNet implementation in PyTorch
- X-Net: Classifying Chest X-Rays Using Deep Learning

Proposed Method

Preprocess

1. 年齡過高（大於145歲）或年齡過低（低於16歲）
2. 複診次數與年齡的關係不合常理，一個病患的複診次數在年齡增長後，反而變少

Model

- Supervised
- GAN Data Augmentation

以下為我們最初版實作的model，先將原始影像大小調整為224x224，再送進去四層的CNN model中，最後通過兩層NN，預測出14個病徵分別可能罹患的機率值。

Layer (type)	Output Shape	Param #
conv2d_5 (Conv2D)	(None, 111, 111, 64)	640
batch_normalization_4 (Batch	(None, 111, 111, 64)	256
leaky_re_lu_4 (LeakyReLU)	(None, 111, 111, 64)	0
max_pooling2d_4 (MaxPooling2	(None, 55, 55, 64)	0
dropout_4 (Dropout)	(None, 55, 55, 64)	0
conv2d_6 (Conv2D)	(None, 53, 53, 128)	73856
batch_normalization_5 (Batch	(None, 53, 53, 128)	512
leaky_re_lu_5 (LeakyReLU)	(None, 53, 53, 128)	0
max_pooling2d_5 (MaxPooling2	(None, 26, 26, 128)	0
dropout_5 (Dropout)	(None, 26, 26, 128)	0
conv2d_7 (Conv2D)	(None, 24, 24, 128)	147584
batch_normalization_6 (Batch	(None, 24, 24, 128)	512
leaky_re_lu_6 (LeakyReLU)	(None, 24, 24, 128)	0
max_pooling2d_6 (MaxPooling2	(None, 12, 12, 128)	0
dropout_6 (Dropout)	(None, 12, 12, 128)	0
conv2d_8 (Conv2D)	(None, 10, 10, 128)	147584
batch_normalization_7 (Batch	(None, 10, 10, 128)	512
leaky_re_lu_7 (LeakyReLU)	(None, 10, 10, 128)	0
max_pooling2d_7 (MaxPooling2	(None, 5, 5, 128)	0
dropout_7 (Dropout)	(None, 5, 5, 128)	0
flatten_1 (Flatten)	(None, 3200)	0
dense_1 (Dense)	(None, 96)	307296
batch_normalization_8 (Batch	(None, 96)	384
activation_1 (Activation)	(None, 96)	0
dropout_8 (Dropout)	(None, 96)	0
dense_2 (Dense)	(None, 14)	1358

Total params: 680,494
Trainable params: 679,406
Non-trainable params: 1,088

Figure 4: 模型架構