AIMAS Homework 2

F94081076 郭立晨 資訊113

**題目:** 利用資料集訓練模型區別健康與不健康的分類問題

**實作過程:**

1. 使用Keras、sklearn、pandas等套件作為工具。
2. 讀入Data\_Entry\_2017.csv檔案並利用dataframe儲存，內容包含病患的去識別化資料、疾病分類等重要標記。
3. 將每張X光影像的路徑也新增到dataframe新欄位中，方便後面做資料前處理。
4. **前處理:**

* 將原本有疾病的照片標記成Unhealthy，若是原本標記是”No Finding”則標記成healthy。
* dataframe新增欄位存放數字標記，若該圖片是Unhealthy則為1；healthy則為0。
* 一張含有 文字 的圖片

  自動產生的描述一張含有 文字 的圖片

  自動產生的描述去除不需要的欄位，以及去除部分照片來平衡健康、不健康的照片數量。
* 分割dataframe文字資料集為訓練集(0.85)及測試集(0.15)

train 10624 Validation 1875

1. 使用Keras ImageDataGenerator產生存放照片的「容器」，再用flow\_from\_dataframe函式，將照片與dataframe的文字標記組合成完整的訓練集、驗證集和測試集，作為模型的輸入。ImageDataGenerator也可以設定參數使圖片做Augmentation；我建立了有Augmentation以及沒有Augmentation的ImageDataGenerator作為對照。
2. **建立模型**，使用Keras已經建立好的VGG16模型作為基本架構，再加上全連接層、拋棄層，部分層設定為凍結層無法被訓練，只訓練高層的全連結層分類器。最後一層一個輸出使用sigmoid作為激活函式。
3. 為了加快訓練速度，使用Transfer Learning載入”imagenet”預訓練的權重；設定輸入層的維度為 (150,150,3)。
4. 使用” binary\_crossentropy”作為損失函式處理離散的分類問題，”Adam”作為optimizer，Learning Rate=0.00001；用準確率作為模型評斷標準。(Learning rate不能太大否則會破壞pre-trained weights)
5. 設定10個epoch，開始訓練模型。

**實作結果:**

**Model Structure**

**一張含有 文字 的圖片

自動產生的描述**

Augmentation

rescale=1./255, horizontal\_flip = True,

vertical\_flip = False, height\_shift\_range= 0.05,

width\_shift\_range=0.1, rotation\_range=5,

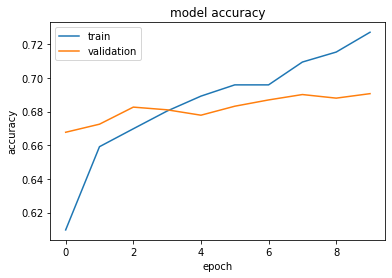
shear\_range = 0.1, fill\_mode = 'reflect',

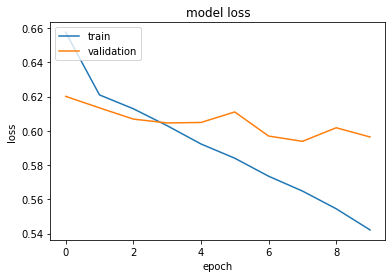
zoom\_range=0.15

**Without Augmentation**

accuracy: 0.6907

loss: 0.5964

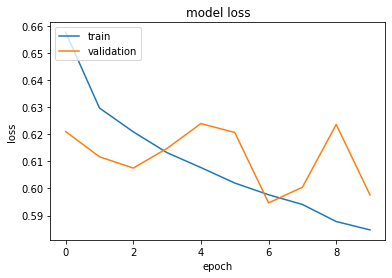
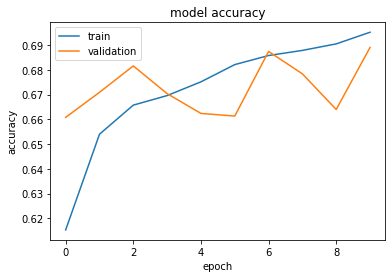
****

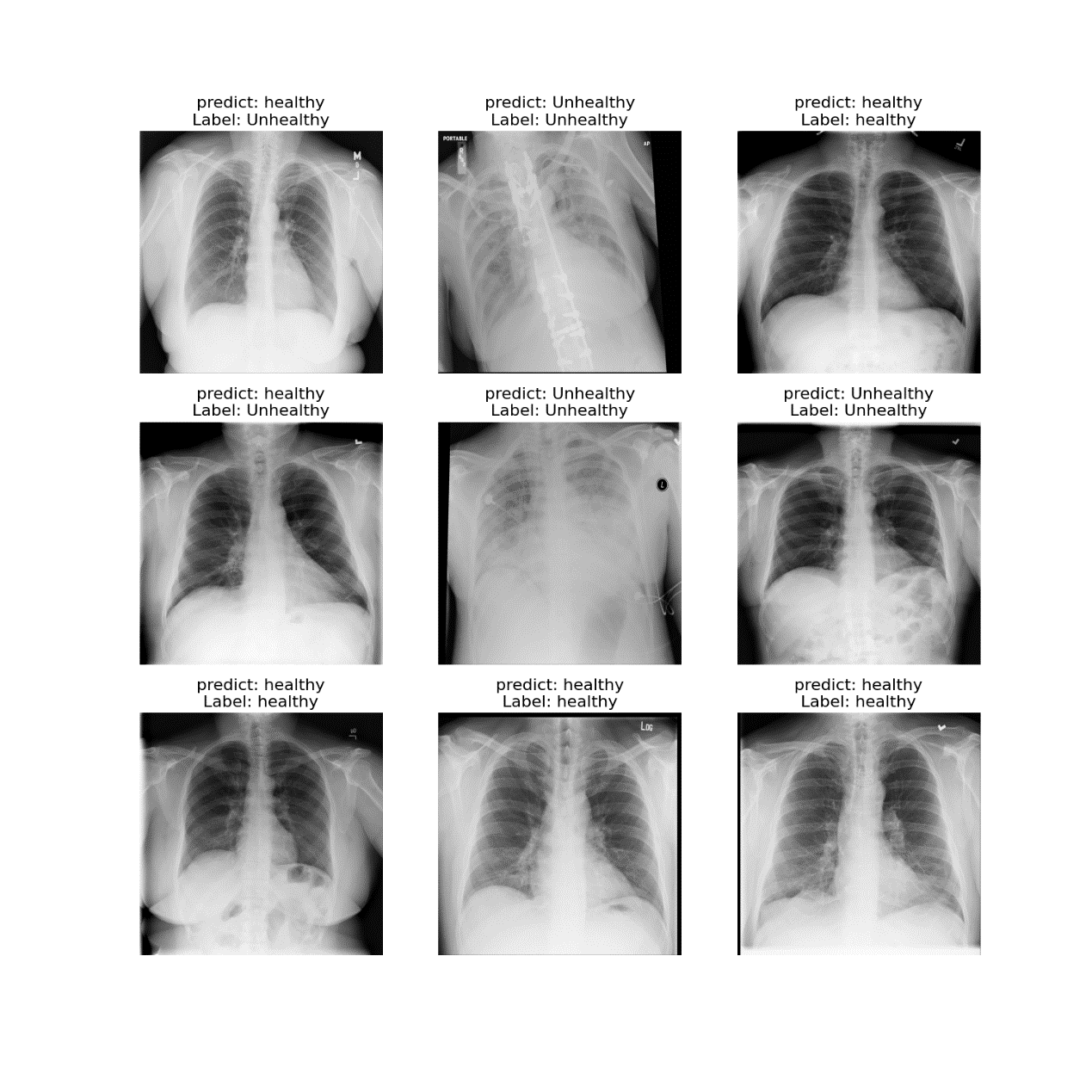
****

**Augmentation**

accuracy: 0.6837

loss: 0.5970

****

****

有無Augmentation的結果差異不大因此只印出其中一種的結果

**預測結果**

**心得感想:**

第一次實作Kaggle提供的題目才知道原來資料集都是這麼龐大，因為硬碟大小以及計算資源的限制所以只下載其中一部分來訓練。光是處理這一小部分資料集就花了我不少時間，也不能像對待其他較小的資料集一樣將照片一次全部讀入並存起來使用，不然會造成暫存空間不足的情況。因此我也看了Kaggle上其他網友的Notebook，最後使用了ImageDataGenerator作為引入照片資料的容器，也算是多學到了一個資料前處理的工具。

在訓練的過程中，我特別嘗試了Augmentation是否會影響訓練的準確度，經過實驗發現好像效果並不明顯，但也有可能是我使用的Augmentation不適合用來處理X光影像，這就衍生出了另一個課題，究竟如何找出最適合資料集的Augmentation? 希望在未來持續學習的過程裡可以學到這方面的知識。

另外，我發現在訓練的過程中準確率雖然有上升但是都上升的很緩慢，從訓練集的準確率以及測試集的準確率來看應該是有成功的訓練，但是，可能需要更多的epoch以及更多張訓練圖片才能得到更好的訓練結果。而當我調整輸入層圖片的大小時，所需的運算時間也大幅提升，礙於GPU使用時間的限制只好維持(150,150,3)的輸入大小，我認為準確率不夠高也有可能是我將圖片調整成太小造成關鍵的特徵被犧牲。總而言之，深度學習真的是一個需要運算資源的一項工具，只要資源夠豐富，深度學習相對的一定也能有更好的成效。

**模型下載連結:** <https://drive.google.com/drive/folders/14CQrsqHpCwzsxOdfeFj4DBPCjyGFialo?usp=sharing>