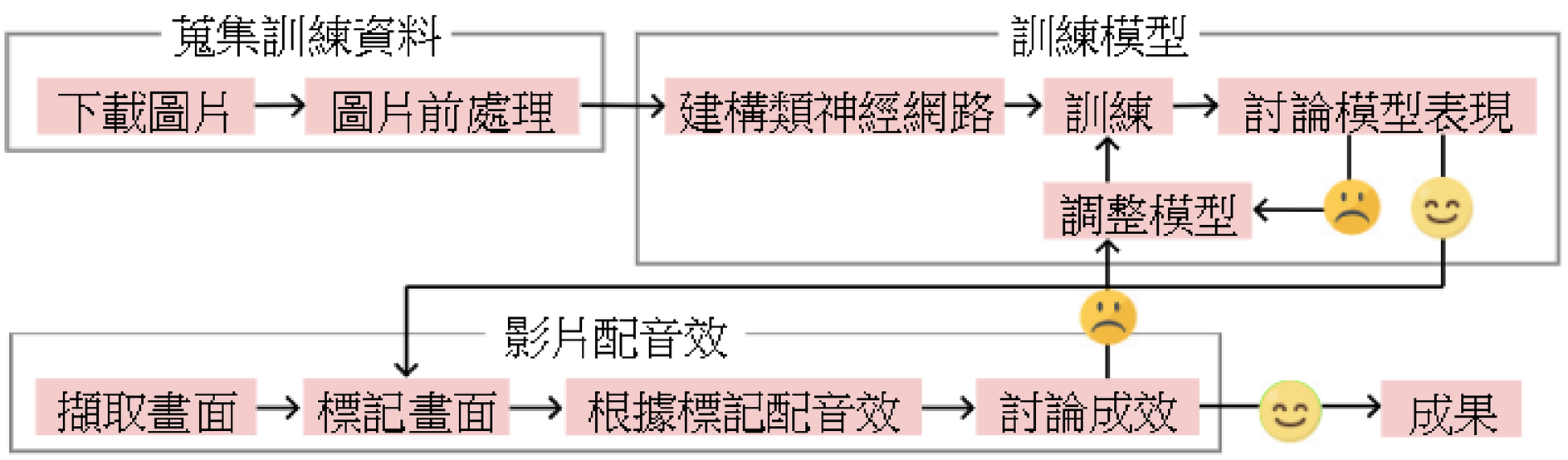
摘要

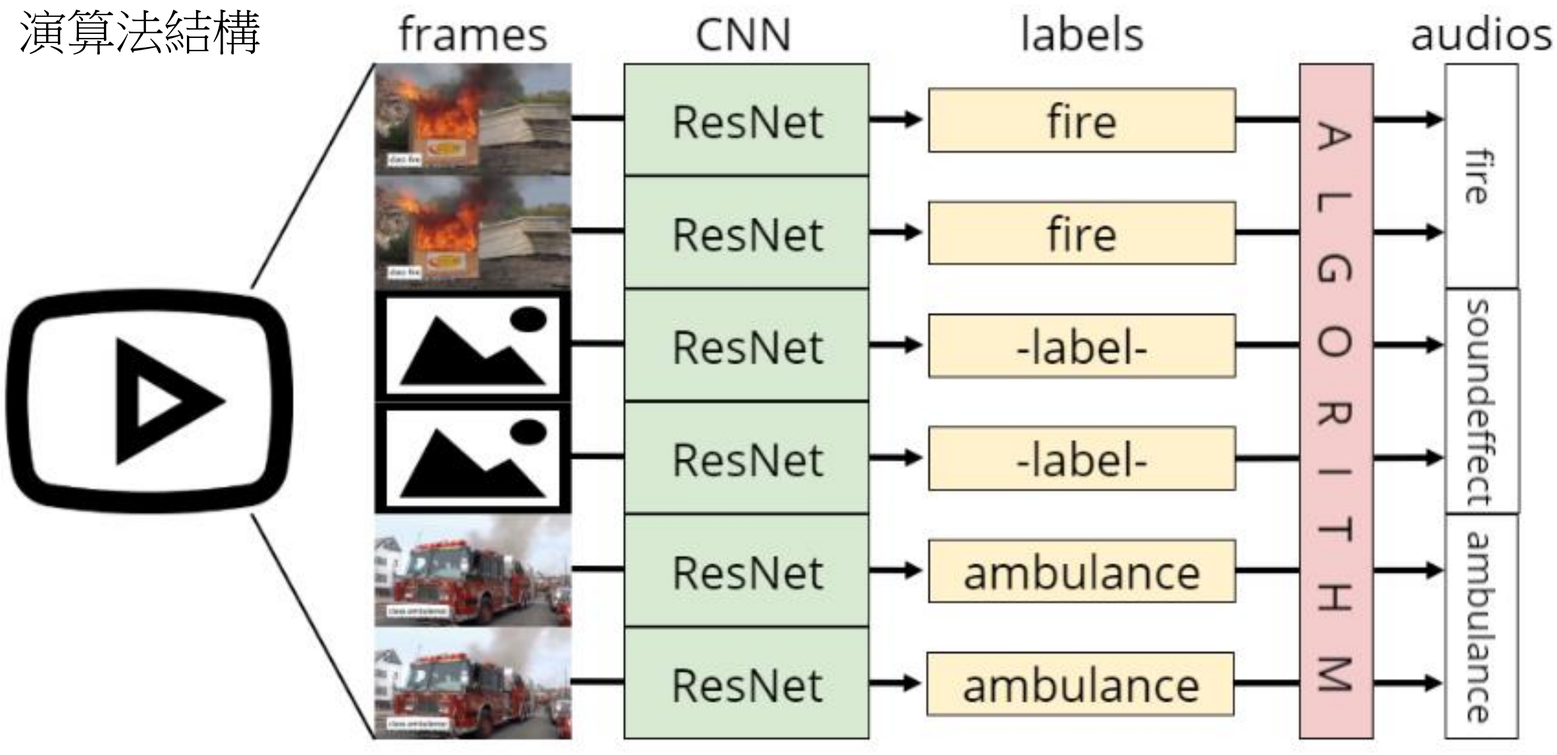
本研究旨在研究利用人工智慧深度網路技術訓練模型，並期望訓練後的模型能為當 下影片的內容配上適當的音效，使得幫影片配音的過程可以加速，並減少人力需求。本研究主要利用 Keras 這個 Python 套件創建模型，並參考 ResNet 這個著名的深度學習模型，透過大量 已標記圖片當作訓練資料，並進一步修改模型的架構，使準確率得到提升。研究結果符合預期， 能在適合的類別中為影片配上合理的音效。

壹、研究目的

本研究的目標為用機器學習的方法，訓練一個可自動幫影片配出符合當下影片內容的音效之自動配音人工智慧系統。

貳、研究過程





圖一

一、下載圖片資料庫

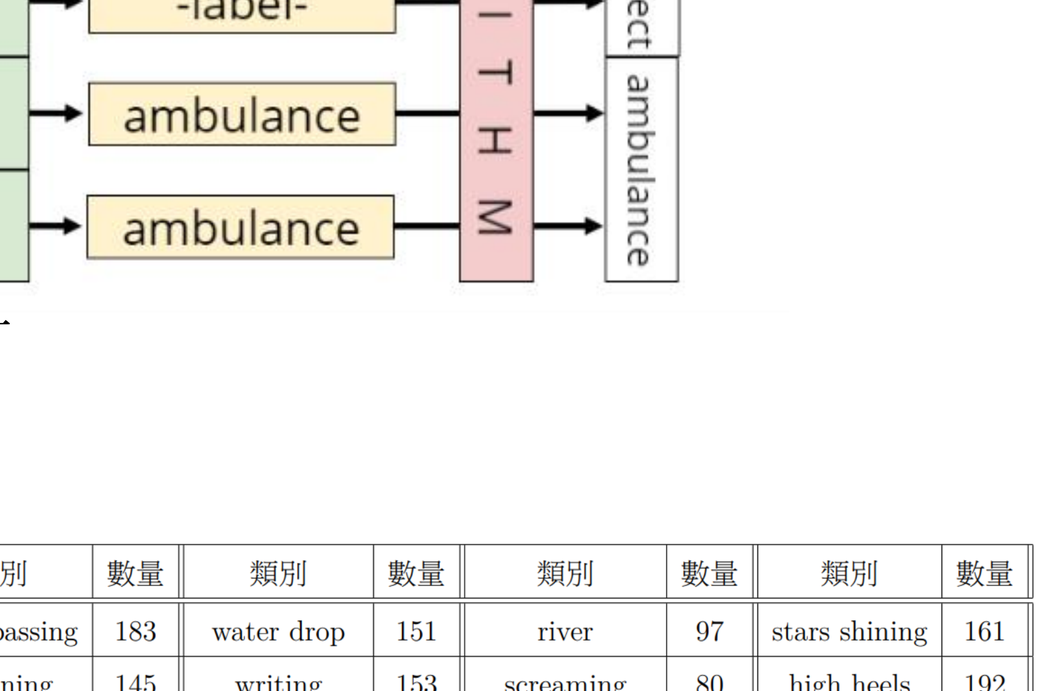
利用 google_images_download 這個 Python 套件中的指令大量下載 Google 搜尋引擎搜尋到的圖片。作為實驗，我們先選用 20 種音效相關圖片類別進行辨識。共蒐集 2804 張圖片，各類別的圖片數量參考圖二。

類別	數量	類別	數量	類別	數量	類別	數量
train passing	183	water drop	151	river	97	stars shining	161
lightning	145	writing	153	screaming	80	high heels	192
wave	196	typing	104	ambulance	74	dog barking	198
cars	187	door opening	133	glass breaking	120	cat meowing	184
cheer	183	fire	94	helicopter	96	blender	75

圖二

二、進行圖片前處理

為了方便處理，我們將圖片變成長寬相同的大小，再進行訓練。對應的，我們也找到專門處理圖片的 Python 套件 PIL，可以支援將圖片以重新取樣的方式縮放。而本次實驗我們統一將圖片轉換成 224 × 224 的大小。 圖三為資料庫的樣本。



圖三

三、建構深層神經網路

起初我們尋找最適合應用在圖片辨識的神經網路，有許多研究團隊開發出優化的捲積神經網路結構諸如 ResNet、Inception、NASNet 等等，最終我們以 Keras 提供的 ResNetV2 範例為架構，搭建 5 個區塊 (blocks) 的 ResNetV2 進行訓練。

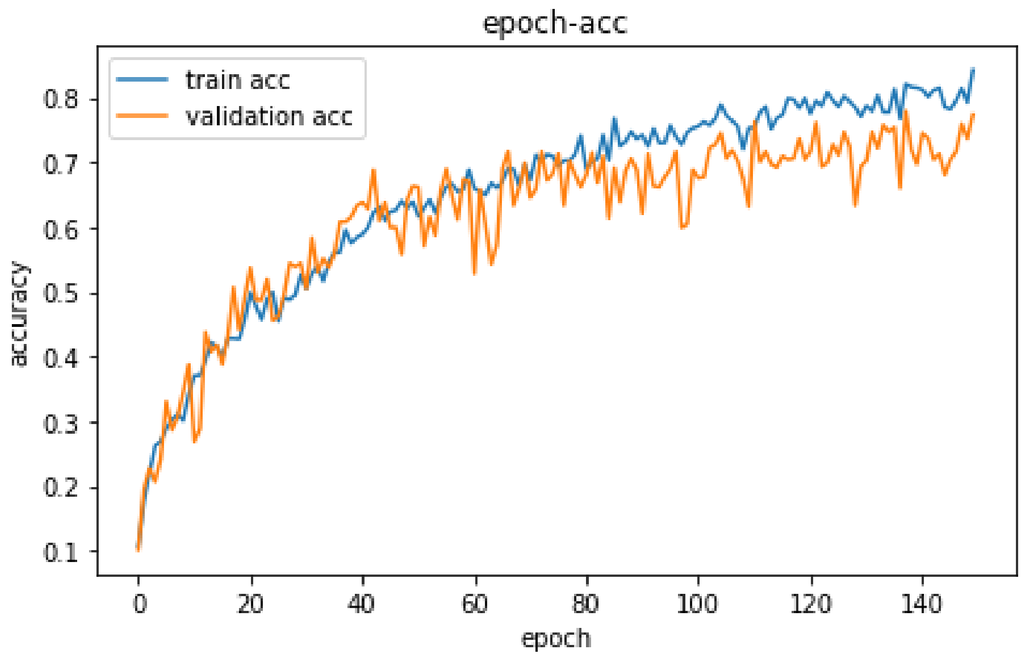
四、訓練、調整深層神經網路

針對圖片辨識的模型訓練時，常常使用資料擴增(data augmentation)的方式增加資料數量及多樣性，使相同的資料可被複製多次使用，更可以期望模型學習到稍微變化的圖片依然屬於同個類別。本次實驗詳細使用的資料擴增參數如圖四所列：

旋轉角度	15°	橫向平移	15%	垂直平移	15%	裁切	15%
縮放	15%	色調平移	10%	水平翻轉	True		

圖四

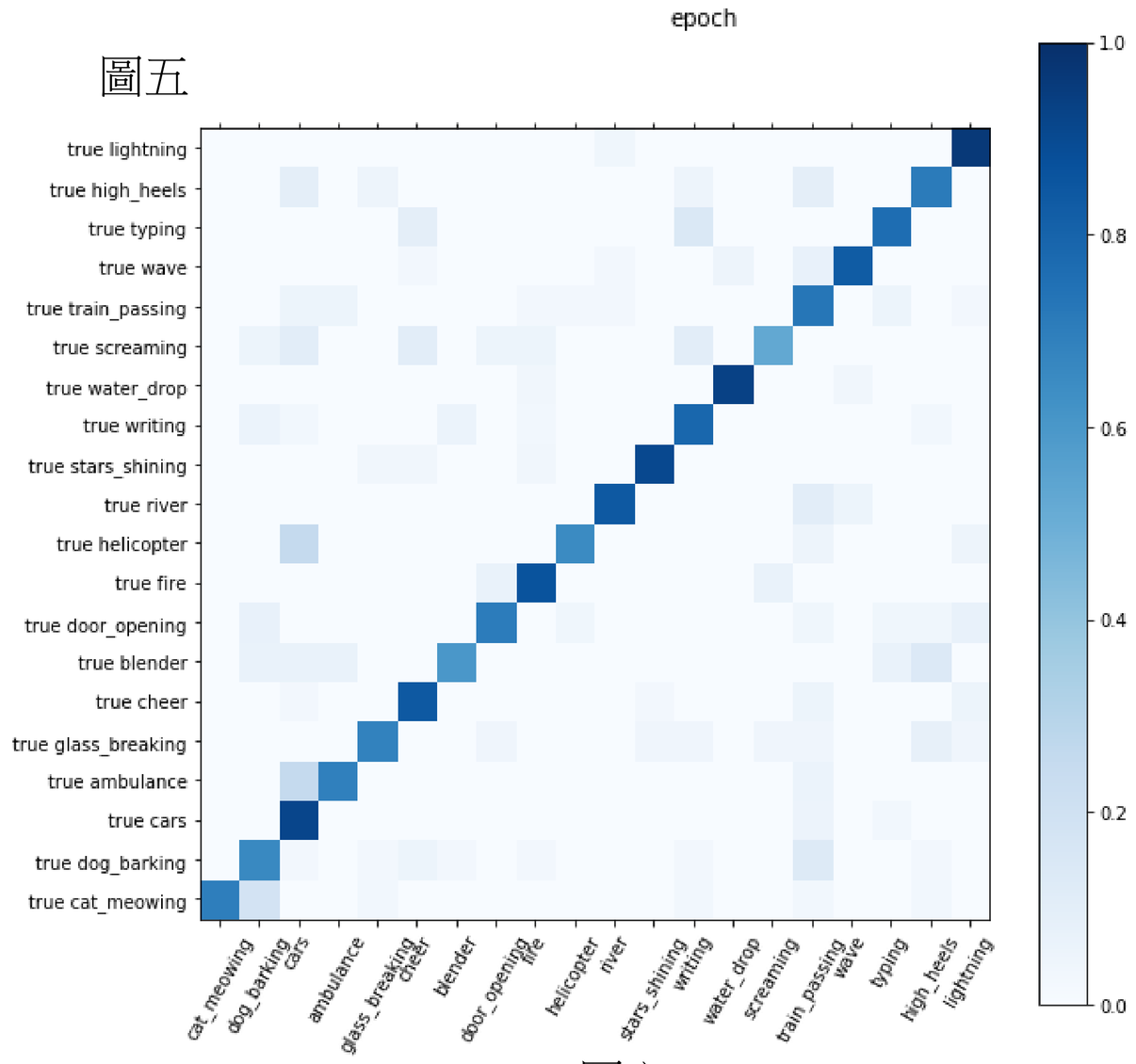
訓練時我們將 20% 的資料當成驗證資料（validation data）和測試資料（testing data），用以監督、客觀驗證模型的表現。在使用 Adam 優化下，我們跑了 150 次（epochs），訓練紀錄圖五所示，最終 20 類分類（測試資料）正確率約為 76%。



圖五

五、討論模型表現

混淆矩陣即統計每個類別的辨識情形統計在矩陣表格上，可明顯的看出類別間的互相混淆關係。圖六為根據訓練完的模型繪製混淆矩陣，顏色越深代表比例越高。



圖六

六、擷取影片中的圖片

利用 OpenCV 套件擷取出影片中的每個畫面擷取出來，並利用PIL套件將圖片縮放成固定 224 × 224 的大小，完成圖片前處理。

七、以模型為每個畫面標記類別

將每個畫面通過模型後得到每個畫面的標記，再以眾數濾波處理雜訊，解決模型預測不穩定之問題，並使標記盡量同類相連。

八、針對影片結果調整模型

從實驗結果的影片1,2,3中可以看出幾個問題：

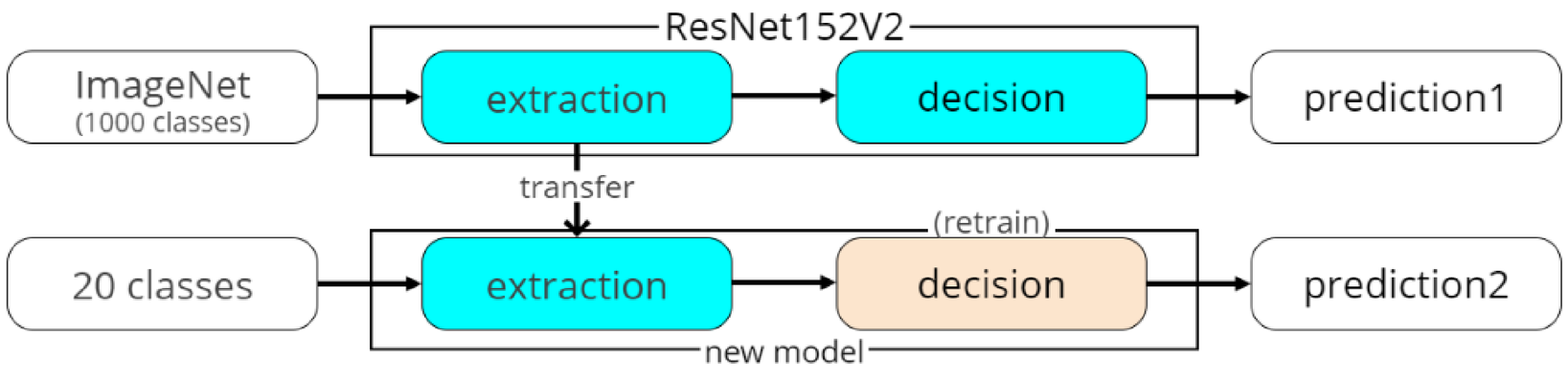
1. 有些片段不應該出現音效，但模型會將其分在最有可能的類別，代表每張圖片都會被配上音效，並不符合預期。
2. 模型無法非常正確地分出圖片的類別，推測是結構不夠完善抑或少量訓練資料無法應付多 變化的圖片。

針對結果，我們在加入加入第21類「其他類別」作為處理無音效的機制，並改以遷移學習的方式加強類神經網路表現。

九、遷移學習

利用遷移學習的方式，調用已在 ImageNet 上訓練好的 ResNet152V2，重新建構最後的全連結層並微調。藉此在少量的資料下調用經歷大量資料訓練的巨大模型，增進模型的表現。大致上如圖七所示：

建立完成後，在僅改變上述結構的情形下以相同參數訓練之。

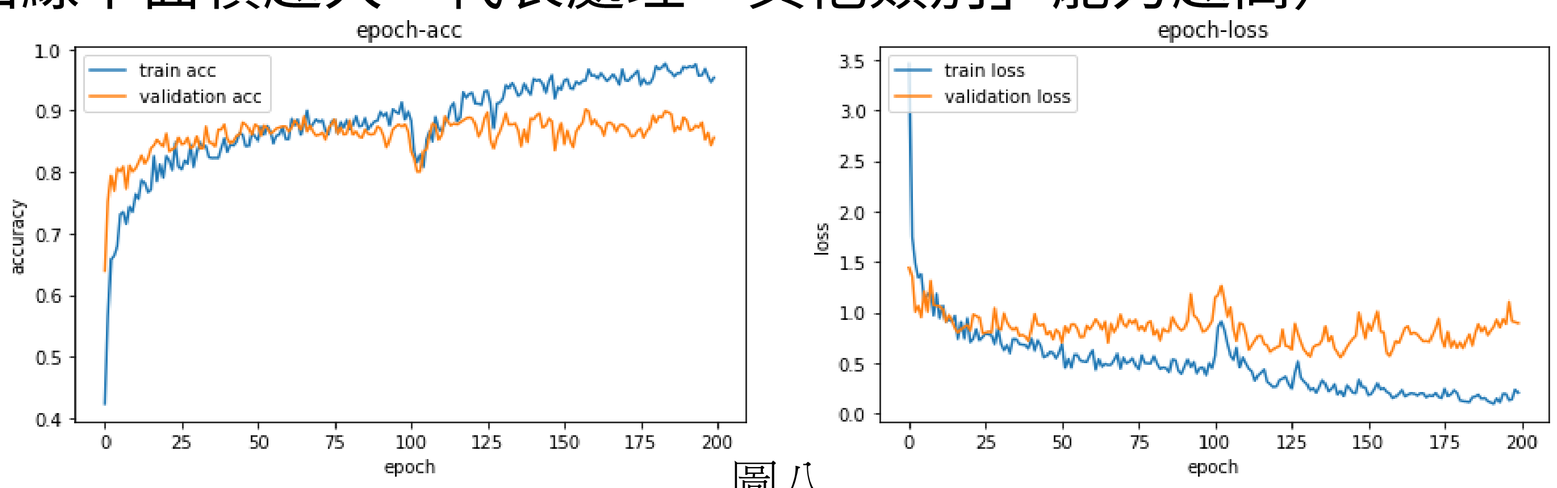


圖七

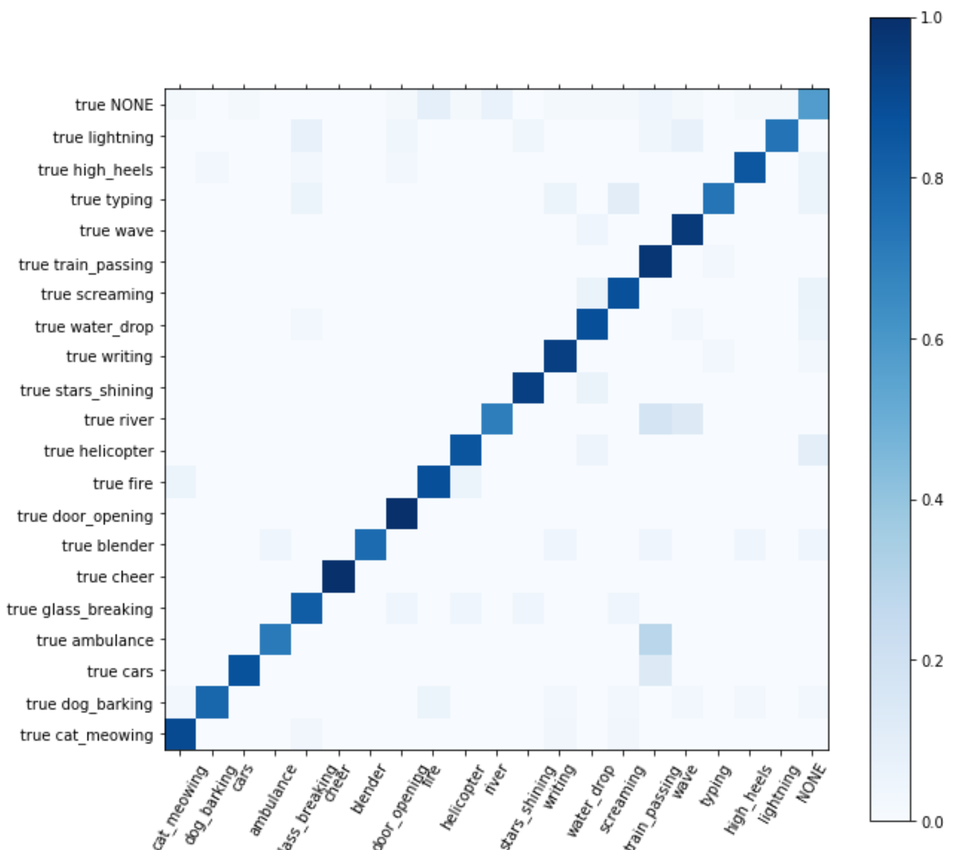
十、討論模型表現

我們畫出訓練的紀錄（圖八）以及混淆矩陣（圖九）如下，其中可以明顯看到第100次附近正確率有明顯的下降，系微調初期的影響。最終訓練達到86%，與先前的模型相比高出 10%。

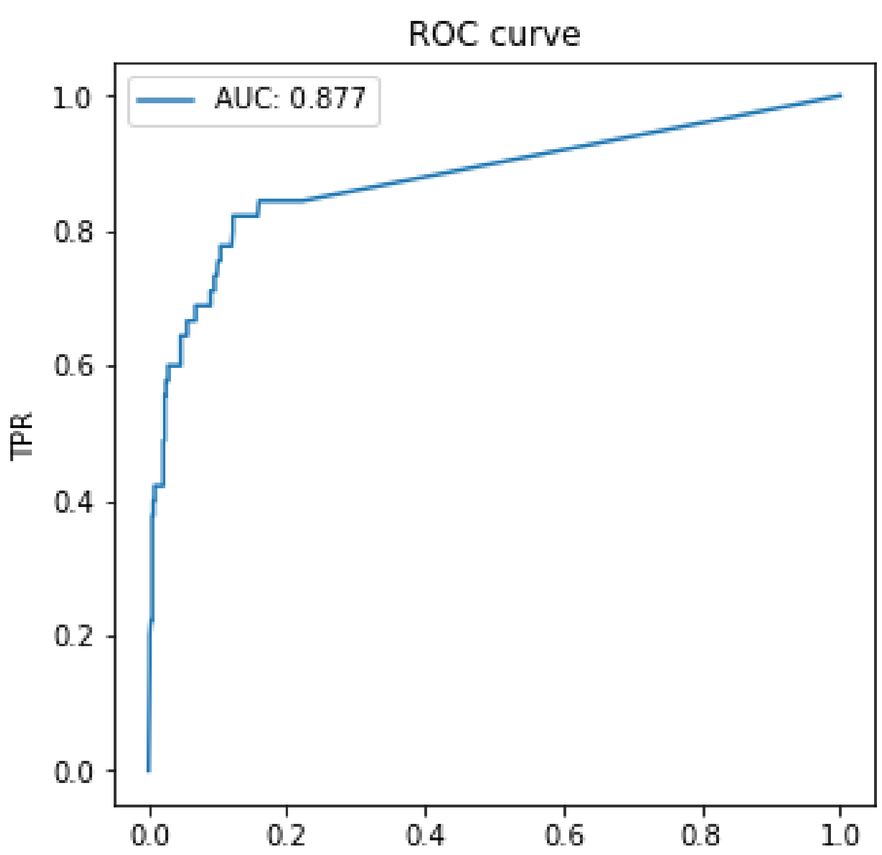
除上述兩項指標外，將「其他類別」視為陰性，其餘有音效類別是為陽性，則可根據「陽性樣本辨識率」和「陰性樣本誤判率」兩者隨這閾值變化的軌跡，繪製成接收者操作特徵曲線（圖十），進而以曲線下面積多寡對應其分類效能的好壞（曲線下面積越大，代表處理「其他類別」能力越高）。



圖八



圖九



圖十

參、研究結果

客觀數據而言，模型在分類上正確率從76%提升至86%，且在辨識「其他類別」的效能達到 Area Under ROC curve 為 0.877，已有極高的辨識率。而應用在影片的成效主觀而言，已將大部分的片段配上理想的音效，然而仍有不少的畫面無法成功辨識，仍有改進空間。

影片連結

肆、未來展望

- **增加音效類別**

  目前蒐集的音效遠不足應付平常影片可能需要的音效，因此蒐集更多類別進行訓練將是應用上重要的需求。

- **以循環神經網路作為決策網路**

  若考慮時序問題，每個畫面的類別確實和前後畫面相關，因此將本實驗的 ResNet 搭配 RNN，可期望加強辨識能力，提升生成音效的品質。

- **新增音效屬性**

  本實驗在將標記轉換成音效時粗淺的將同一音效片段重複播放，然而有些音效適合重複播放，有些重複卻不合理。因此在搭配音效時各類別的屬性也是必要的考量。

- **音效平行播放**

  預期中若畫面符合，同一畫面可能有不只一個音效出現。假如能針對每個音效分別訓練每個畫面「是否」有該音效的二元分類，即可處理平行播放的音效，同時也不需要「其他類別」的機制處理無音效片段。