機器學習期末報告

尋找花中君子

蘭花種類辨識及分類競賽

資工系學士班 3 年 A 班 1081434 林哲賢

資工系學士班 3 年 A 班 1081447 李明昕

生醫碩士班 1 年 A 班 1108302 余泰毅

資工系碩士班 1 年 A 班 1106006 林沛萱

一、開發動機與目標

目前進行對照品種選定工作，需耗費較多的時間與人力成本，台灣每年的蘭花新品種案件數量不斷增加的情況下，透過品種影像資料庫的建置，能應付龐大的蝴蝶蘭新品種辨別，提升蘭花的銷售與產值，且對蘭花此種高價值物種有更多認識。

蝴蝶蘭的形色優美，許多人都會購買收藏，但是卻不被民眾廣為認識，透過人眼去辨識，不僅耗時費力且容易品種辨別機率極高；市面上較少有蘭花品種辨識的軟體或技術，能精準辨識蘭花品種的工具，即使有，也是需要付費軟體。因此未來的教學、研發都能免費利用此資料庫。

透過這次競賽，相關民眾皆能夠透過辨識蝴蝶蘭提升辨識蝴蝶蘭的精準度，對於真心想了解其特性的人，有很大的幫助。相關產業人員與育種廠商皆能夠透過辨識蝴蝶蘭提升蝴蝶蘭知名度，增加銷售機會，還能夠幫助自己的國家提升知名度，是研究蘭花花部發育、花形、花種之重要題材。

而我們也希望用戶可以用資料庫與其圖片是否如同自己所想，呈現特徵明顯的蘭花品種而被正確辨識。怎麼分類透過蝴蝶蘭影像分類物種、依照外觀特徵辨識該物種是個很重要的課題。

- 以下使用的對象能提高辨識蝴蝶蘭的精準度:
  1. 一般百姓: 難以透過特徵推測蘭花品種，使用影像處理資料庫後，拍攝圖片即可辨識。
  2. 專業人士: 能輔助辨識蘭花的品種，減少不確定性。
  3. 育種廠商: 檢查蘭花品種生長，可確認是否有培育新的品種，降低人事檢測的成本。

二、參考文獻探討

我們做為借鏡的是 Dong Chen 等人所撰寫的“Performance Evaluation of Deep Transfer Learning on Multiclass Identification of Common Weed Species in Cotton Production Systems”。對 Deep Transfer Learning (DTL) 進行全面評估，以識別美國南部棉花生產系統特有的常見雜草。收集了 15 種雜草類別的 5187 張彩色圖像組成，通過遷移學習評估了 27 個最先進的深度學習模型，並為所考慮的雜草識別任務建立了廣泛的基準。

DTL 實現了 F1 - sorce 超過 95% 的高分類準確率，需要相當短的訓練時間（小於 2.5 小時）跨模型。ResNet101 取得了 99.1% 的最佳 F1 分數，而 27 個模型中有 14 個的 F1 - sorce 超過了 98.0%。這項研究為雜草分類任務的 DTL 模型的明智選擇提供了良好的基礎，並且可以有益於整個精準農業研究。

三、解決方案介紹

我們蒐集資料的方式，是對既有資料以水平翻轉、垂直翻轉、隨機旋轉的方式進行資料擴增將其擴增成十倍大，在將資料讀入前，資料前處理會考慮到 RAM 的大小，將圖片從原本的 640*480 調整為 160*160，使用深度學習在進行圖像分類或者對象檢測時候，首先需要對圖像做數據預處理，最常見的對圖像預處理方法是圖像的數值標準化/正規化處理。

運用多個模型預測圖片準確率，其中 ResNet101 model 的圖片預測準確度最為優秀。

1. CNN: 在影像識別方面的威力非常強大，許多影樣辨識的模型也都是以 CNN 的架構為基礎去做延伸。圖片經過各兩次的 Convolution, Pooling, Fully Connected 後，測出的準確度為 88%。
2. VGG16: 模型是由若干卷積層和池化層堆疊（stack）的方式構成，比較容易形成較深的網絡結構，但是需要訓練的參數十分龐大，導致其需要大量的計算資源。表現在訓練集上的準確度相較於淺層網絡不但沒有提高，反而會下降，而我們測出的準確度僅為 0.4%。
3. ResNet101: Resnet 提供了兩種選擇方式，也就是 identity mapping 和 residual mapping，如果網絡已經到達最優，繼續加深網絡，residual mapping 將被 push 為 0，只剩下 identity mapping，這樣理論上網絡一直處於最優狀態了，網絡的性能也就不會隨著深度增加而降低了。測出的準確度也是最好的，高達 98.6%。

四、模型預測結果

比較多個模型的預測效果:

Cnn:

```
>>> print('測試資料損失值:', cnn_model_loss)
測試資料損失值: 0.8463581204414368
>>> print('測試資料準確度:', cnn_model_val)
測試資料準確度: 0.880821943283081
```

VGG16:

```
>>> print('測試資料損失值:', vgg_model_loss)
測試資料損失值: 5.389562129974365
>>> print('測試資料準確度:', vgg_model_val)
測試資料準確度: 0.004921359941363335
```

```
>>> print('測試資料損失值:', resnet_model_loss)
測試資料損失值: 0.04596792161464691
>>> print('測試資料準確度:', resnet_model_val)
測試資料準確度: 0.9861999154090881
```

Resnet101:

我們的三個模型運算時間，都大致落在 1 小時左右，總計這三個圖片預測準度就花費 3 小時，而其他的前處理、後處理運算和資料全部加起來，總計有 6 小時去執行。

由於本身在執行全部程式碼，就需要龐大的時間去等待，資料前處理必須考慮到 RAM 的大小，於是將圖片從原本的 640*480 調整為 160*160，減少所需的記憶體，使工作進度更加快速。

五、開發最耗時的部份與原因

遇到的問題:

我們訓練模型的時候，使用資料擴增(Data Augmentation)，來達到更好的模型效能。使用資料擴增的原因有兩個：增加手上僅有資料的效率、以及增加模型的普適性。透過資料擴增，我們可以把一張影像擴展成很多不同的影像，並且使用這些擴展出來的影像來訓練模型，可以讓模型的辨識能力更好。

但是，資料擴增並不是隨便做，將影像上下顛倒後拿去訓練蝴蝶蘭模型，會導致模型硬學出一些其實不太有用的東西，反而讓模型效能降低。所以如何獲得最好的資料擴增，我們依照問題的屬性，來決定要使用哪一種資料擴增。

我們從頭到結束，程式執行的時間總共花費了 6 小時多，運算的時間會這麼久，是因為資料集的圖片過於大量，公開測資提供的圖片量就高達 2000 多張，以及資料擴增、三個模型的準確度預測。

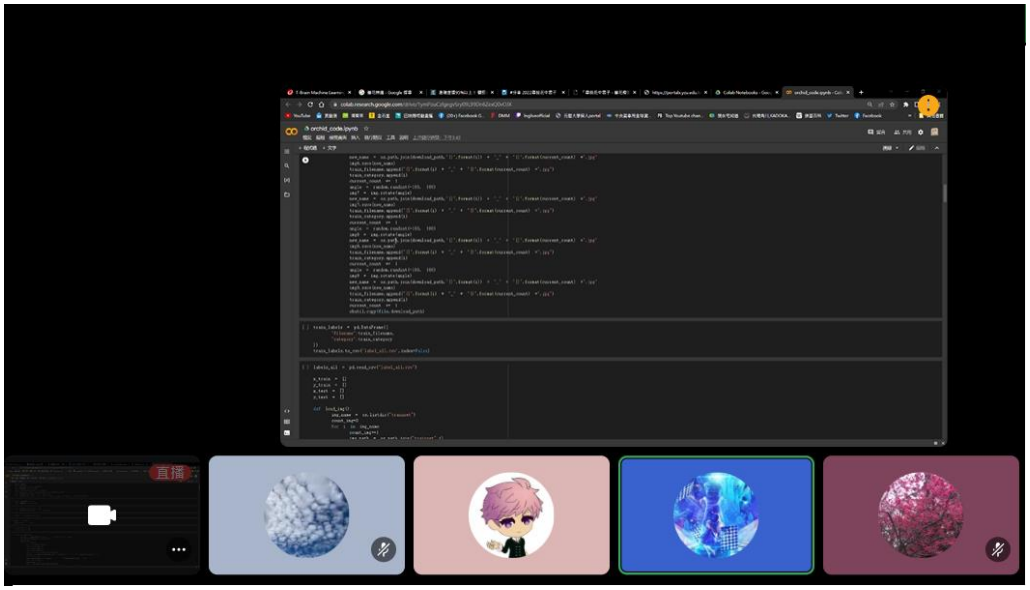
未來可改善的地方:

資料集不足，生成的模型通常無法正確訓練，如果成功的話也僅侷限在已知的樣本中，所以花的性徵去辨別種類就相當重要，可從花梗－花瓣－花心等去做更細緻的區分。再搭配使用不同的模型，去計算其準確率高，透過花朵分類結果及演算法架構，可建構更完善的蝴蝶蘭影像辨識系統。

然而人類的視覺系統，對於亮度比較敏感，而對於彩度比較不敏感。況且，三原色所構成的向量空間無法對影像強度(亮度)做處理，例如柔和化、銳利化等等。同時，由 RGB 構成的影像檔案也在傳輸時佔用較大頻寬、儲存時佔用較多的記憶體。

因此將原來 RGB 的格式由數學轉換成 YCbCr 的組合，其中 Y 是亮度(Luminance)，Cb、Cr 是色差(chrominance)。Y 值透過區域二元模式求得每個樣本的紋理特徵，並且加入 Cb 與 Cr 為色彩特徵做為輔助，將得到的紋理與色彩特徵串聯成所代表之特徵向量，未來就可採取以紋理形狀特徵及色彩的方式，實作於蝴蝶蘭品種辨識。

六、小組互動照片



七、專案開發心得

1081434-專題開發心得:

在獲得資料集時，第一個想法就是每個類別的資料量太少，需要將資料集擴大。一開始採取爬蟲的方法，結果發現爬出來的資料過於混亂，花了一些時間釐清原因之後，理解到是某些品種過於少見，使得在爬蟲時會爬到其他的品種。因此我們決定用旋轉和翻轉的方式來增加資料集。

在資料前處理時，由於資料使用原始大小會用盡記憶體，因此我們在資料大小和資料清晰度之間做了取捨。此次我們採用三個模型來訓練，並從中選取最好的模型來用。

雖然此次的成績並不理想，但是在不斷的 try and error 中也學到了一些意想不到的東西，如:爬蟲的方法與速度的調整、硬體限制下如何處理資料。這些經驗與老師提供的意見使我能用其他想法來看待往後會遇到的資料。

1081447-專題開發心得:

這麼龐大的資料集，原以為有足夠多的資料可進行學習，然而卻是超多的種類以及稀少的圖片，每個種類僅 10 張圖片在學習上可以知道這樣的數據量極為缺乏，最重要的環節變成如何進行資料增量，真的回憶起老師在上課所說，一個好的演算法若是沒有相應的數據，也無法比爛演算法搭配大量的數據來的準確，我們閱讀了一篇雜草分析的論文，雖然他提到可以使用 GANS 來用既有的圖片去生成，但該論文還是使用旋轉與翻轉進行資料增量，所以我們也採取這樣的方式進行。

在之後進行分析時，也遭遇了困難，面對如此龐大的運算，個人的電腦顯然是跑不動的，突然覺得只要是大型計劃，總是需要一個 server，才有辦法達到想要的成果，不然 colab 還沒跑完就斷線了，只能苦哈哈。

1108302-專題開發心得

當初下載資料集時嚇了一跳，各種奇怪的蝴蝶蘭映入眼簾。檔案最大的一張照片甚至看不到花，以為是不是給錯圖片了，才想起老師和指導教授提到的資料預處理。萬事起頭難，在機器學習上更是明顯，龐大的數據要整理出能用的部分，一張圖片也得掰成 10 份來用的感覺相當不輕鬆。

後來遇到算力不夠的問題，我們還向指導教授要了之前教授在算機器學習論文的 IP，這下才不會遇到被砍的問題。終於不會遇見學弟們之前跑 YOLO 不會被 google 砍，而我的一直被砍的夢魘。不過那個 IP 也只有 CPU，跑起來雖然保證不會斷，不過卻也算了好幾個小時，導致最後沒有剩餘多少時間能夠再做校驗，相當可惜。

這次順帶負責了簡報的美編，我明白資工美感好的不多，所以希望能夠呈現整體營造不差的簡報。預報是使用自己之前畫的簡報背景撐著用，一直覺得不符合蘭花的主題，於是結報改找模板使用。就像資料要預處理，拿到的模板也得校正那細微的差距看起來才會協調，希望最後看起來能令人滿意。只是看到成功率超級高的那一組的簡報卻是標準資工系會做出來的簡報時，心中真的大嘆，好的內容也要有好的主題來襯托才是應該。

1106006-專題開發心得:

在開始研究蘭花時，沒想到蘭花的品種如此之多，很多都有高相似度，要如何去分辨在顏色、外表相似的細小差異性，這是極為困難的課題，以及使用何種模型的準確度會是最好，我們將圖片的大小做調整，是為了電腦計算運行時間減少，而我們使用三種模型:CNN、VGG16、ResNet101，其中 ResNet101 準確度最高，能辨別蘭花的品種不同。

我們對資料集進行了分析和預處理，機器學習演算法需要單獨的訓練集、測驗和驗證來進行預測，資料集是訓練模型時不可或缺、十分重要的一環，在實驗結束後的隱藏測資，我們發現資料集不足時，極易導致過擬合，畢竟官方的公開圖片只有給予 2000 多張，隱藏測資去執行卻高達上萬張，這個也是訓練不足，影響其結果準確率。

在資料本身上也須慎選，就以我們最好的 ResNet101 而言，背景顏色、尺寸、旋轉角度等問題都可能導致辨識正確率偏低。圖片縮小模糊、未對焦、有陰影等也可能會導致和實際蘭花的差異，因此我們以後需要找到最不影響的驗證方式，以免影響準確度。

八、組員名單與分工

姓名	學號	工作內容	專題佔比
林哲賢	1081434	資料處理與建置模型	30%
李明昕	1081447	評估模型並修改參數	20%
余泰毅	1108302	蒐集相關文獻、簡報美編	25%
林沛萱	1106006	意見統整、製作簡報與報告書	25%