Machine Learning (2018, Spring)

Final project report

Humpback Whale Identification Challenge

Group：隊名產生口口犬口口
R06521605 許舜翔
R06521601 黃伯凱

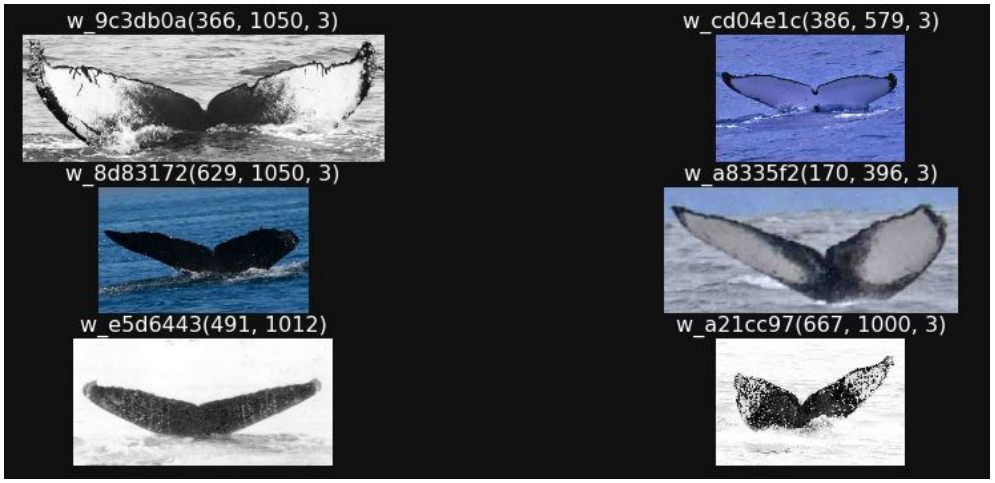
1. Introduction & Motivation

在禁止經歷數世紀的捕鯨行為後，鯨魚的復育仍面臨許多挑戰，除了需克服異常的氣候變遷之下，地球暖化造成的海溫上升，其食物來源也因為人類大量捕魚而受限，為了有效地進行物種保育，科學家長期對不同鯨魚個體進行追蹤，其係利用照片監控系統來掌握這些個體的動向，而這些觀測資料，已經蒐集將近四十年的時間，並由專業人員以尾鰭形狀及獨特的標記符號做分類依據，對這些照片進行標籤、紀錄。在這麼長時間之下，這些以人工密集的方式整理出來的資料，在過去受限於硬體設備之下，而未能得到有效地利用，隨著資料處理的技術提升，將能結合深度學習的演算法，重新進行加值應用，期未來將能取代或是作為輔助科學家進行標籤的工作。

使用的訓練資料集共有25,000張影像資料，而測試資料集則有15,116張影像資料，其中皆包含黑白及彩色照片，且像素大小不一。此分析困難的地方在於訓練資料集含括3000+個鯨魚個體，其中"new whale”又占了快1000張，而每個個體僅有數張照片，甚至只有一張等情況，造成資料分布嚴重失衡，加上格式不一的情形，皆是在處理資料中需要特別注意的問題。而這類個體判別的問題，也是目前機器學習領域發展的趨勢，如何利用少量的資料，就能達到不錯的精度，像是深度學習中的演算法 - 「one-shot learning」，本研究將嘗試不同的演算法及模型架構，配合不同的資料集進行訓練，再分析比較其結果，歸納收斂得出表現最佳的模型。

2. Data Preprocessing/Feature Engineering

2.1 分析資料



圖一、隨機作圖並標示Label及Shape

首先，參考Lex Toumbourou於Kernals上所作的資料分析流程 [1]，我們隨機從training set中挑6筆資料來作圖，如圖一，並且將其label以及image shape作為圖片標題，可以從中發現以下三點：

1. 這次比賽係藉由鯨魚的尾鰭達到辨識個體鯨魚之目的。
2. 訓練資料中包含黑白以及彩色照片，但值得注意的是，有些照片雖然從shape的角度看起來是彩色，但實際上是黑白照片。
3. 照片尺寸多為寬>長，大小分配不均，例如有的寬為1050，有的寬則為396，落差很大。

將所有出現過的label整理過後，藉以得知訓練資料一共包含了4251種鯨魚，此外，每種鯨魚約莫只有一到兩張照片，這和我們過去在做影像辨識相關的訓練非常地不同，過去我們的訓練資料大多都是每種品類皆有不少張照片，藉此達到較好的訓練結果。

表一、訓練資料前三多之鯨魚種類

種類	照片張數
new_whale	810
w_1287fbc	34
w_98baff9	27

如表一，”new_whale"代表的是從未出現過或是尚未被標示個體的鯨魚，光是此一label就有810張，整整佔了將近訓練資料中的百分之十。更進一步可以發現，有2220種鯨魚僅有一張訓練資料，另外有1034種鯨魚只有兩張訓練資料。一共也僅有4250種(不包含new_whale)鯨魚，僅有一到兩張照片的鯨魚種類就佔了3254種，這些特性都使得這次的訓練相當不容易。

2.2 資料處理

針對上節分析資料結果，可發現本次比賽目的為辨識個體鯨魚，而且每一個體的照片數並不多，除此之外，照片有分黑白及彩色，照片尺寸亦不統一。因此，本研究的資料處理方式將會針對尺寸、顏色以及鯨魚種類來做相對應地處理方式。

顏色部份本研究團隊構想為，將原先的訓練資料擴增為兩部分，第一部分為所有的黑白照片，再加上所有彩色照片轉換後的黑白照片，第二部分為和原先訓練資料相同。同樣地，test set也會預先做一樣的處裡。

尺寸部份本研究團隊統一將所有照片尺寸轉換為「350x200」，此尺寸能夠大略符合鯨魚尾鰭的特徵，也能夠降低訓練時電腦所需的效能。此部分test set也會做相同之處裡。

最後本研究團隊針對鯨魚種類訓練資料不均的問題做出兩種處理方式：

1. 鯨魚種類中，new_whale一共佔了810張，然而此種類是對訓練較沒意義的資料，因此不會納入訓練之中，本研究將採最後模型若分類的前五名把握度不高於某數值時，就將其擺在new_whale之後。

2. 因每一鯨魚種類訓練資料不足的問題，需實作data augmentation，針對單一照片，將其透過旋轉、水平偏移、垂直偏移、斜切角度、隨機縮放以及鏡射多項處理，藉此達到生成更多的照片來解決此一問題，實作的方法為透過keras套件來進行影像生成。

```
ImageDataGenerator( zca_whitening=False, rotation_range=8,
width_shift_range=0.2, height_shift_range=0.2, shear_range=0.2,
zoom_range=0.2, horizontal_flip=True, fill_mode='nearest')
```

2.3 資料集

承接上小節所敘，我們將原先格式不一的資料，轉成統一尺寸，為儘可能避免影像失真，考量我們所擁有的硬體設備能負荷的情況下，盡量保有最大的影像畫素，同時比較在維持影像相同比例，與改變比例會對結果造成什麼樣的影響（如判斷精度、收斂速度等），訓練模型用的資料集共有四種，如表二所示。

表二、本研究使用之資料集描述

名稱	描述	大小
資料集 A	去掉會影響平衡的”new_whale”，並保留鯨魚尾巴的寬比例，轉成200x350的黑白照片。	9040x200x350
資料集 B	去掉會影響平衡的新鯨魚，並保留鯨魚尾巴的寬比例，轉成200x350的彩色照片。	9040x200x350x3
資料集 C	去掉會影響平衡的新鯨魚，並將尺寸比例轉成一正方形，100x100的黑白照片。	9040x100x100

3. Model Description

3.1 CNN

卷積神經網絡(Convolutional neural networks, CNN)為目前處理影像深度學習的主要方式之一，在課堂中作業亦有實作CNN的案例，且CNN的表現相較於其他方式更顯卓越，因此很直覺地選擇CNN作為我們的主要模型。期望能透過CNN解決訓練資料不足以及個體辨識的困難。

本團隊模型架構參考於另一個在Kaggle平台上，個體鯨魚辨識競賽獲得第一名的組別 - Deepsense.io(2016)[2]，模型架構如圖二所示（以我們表現最好的模型為例）。主要透過多層卷基層、池化層以及一層平坦層，最後再接一般的影藏層，達到辨識4210隻個體鯨魚。

3.2 Siamese network

由於研究的題目資料集特性為，有多種類但各種類資料量卻遠低於現階段常用演算法所需資料量的情況，而為解決資料量不足的問題，one shot learning被視為未來相關技術發展的重要議題，而本團隊將嘗試利用該議題中其一的演算法 - 孿生網絡（siamese network），並參考sorenbouma, github source code[3]，架成模型如圖三所示。透過比對的方法，將資料形成pairs來進行訓練，並以歐式距離取作為兩張圖像間的差異，使得相同類別的圖像應差距越小，而不同類別則應越大，該方法期能透過比對其他種類，來達到增加資料量的效果，消彌部分類別資料量過少的缺點，並與前述方法所得之精度進行優劣比較。

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 200, 350, 64)	1792
conv2d_2 (Conv2D)	(None, 200, 350, 64)	36928
batch_normalization_1 (Batch	(None, 200, 350, 64)	256
max_pooling2d_1 (MaxPooling2	(None, 66, 116, 64)	0
dropout_1 (Dropout)	(None, 66, 116, 64)	0
conv2d_3 (Conv2D)	(None, 66, 116, 64)	36928
conv2d_4 (Conv2D)	(None, 66, 116, 128)	73856
conv2d_5 (Conv2D)	(None, 66, 116, 128)	147584
batch_normalization_2 (Batch	(None, 66, 116, 128)	512
conv2d_6 (Conv2D)	(None, 66, 116, 128)	147584
max_pooling2d_2 (MaxPooling2	(None, 33, 58, 128)	0
conv2d_7 (Conv2D)	(None, 33, 58, 256)	295168
conv2d_8 (Conv2D)	(None, 33, 58, 256)	590080
batch_normalization_3 (Batch	(None, 33, 58, 256)	1024
max_pooling2d_3 (MaxPooling2	(None, 11, 19, 256)	0
dropout_2 (Dropout)	(None, 11, 19, 256)	0
flatten_1 (Flatten)	(None, 53504)	0
dense_1 (Dense)	(None, 2048)	109578240
batch_normalization_4 (Batch	(None, 2048)	8192
dropout_3 (Dropout)	(None, 2048)	0
dense_2 (Dense)	(None, 4250)	8708250

Total params: 119,626,394
Trainable params: 119,621,402
Non-trainable params: 4,992

圖二、CNN模型架構圖

```
____________________________________________________________________________________________________
Layer (type)                     Output Shape          Param #     Connected to
====================================================================================================
input_5 (InputLayer)             (None, 100, 100, 1)   0
____________________________________________________________________________________________________
input_6 (InputLayer)             (None, 100, 100, 1)   0
____________________________________________________________________________________________________
sequential_3 (Sequential)        (None, 4096)          27413312    input_5[0][0]
                                                                   input_6[0][0]
____________________________________________________________________________________________________
merge_2 (Merge)                  (None, 4096)          0           sequential_3[1][0]
                                                                   sequential_3[2][0]
____________________________________________________________________________________________________
dense_5 (Dense)                  (None, 1)             4097        merge_2[0][0]
====================================================================================================
Total params: 27,417,409
Trainable params: 27,417,409
Non-trainable params: 0
____________________________________________________________________________________________________
```

圖三、孿生網絡模型架構圖

4. Experiment and Discussion

4.1 實驗結果

以前文所敘之模型結構，加上不同的訓練資料集，分成以下實驗方法，比較其上傳Kaggle的public score (如表三)，由於我們先前於準備資料集時便將'new_whale'類別的資料剔除，但觀察該競賽的sample submission上傳結果，該類別仍占相當的份量(public score = 0.38)，故仍需設定一權重來衡量要將其放入至預測值的位置，我們以試誤法得出當設定'new_whale'為0.97時，得到的驗證結果最好，依此設定權重，結合模型經過softmax函數輸出後各類別的機率，選定前五筆最有可能的類別來當作上船的預測值。

表三、本研究規劃的訓練敘述

模型	內容	訓練參數	精度
A	CNN + 資料集A	Adam(lr=0.00123) epochs = 400-600 batch size = 32	0.43734
B	CNN + 資料集B		0.40143
C	CNN + 資料集C		0.44047
D	Siamese network + 資料集C		0.39844

4.2 模型集成

我們將表現前三好的模型，實作ensemble方法，各模型分配的比重皆相同，僅單純結合最後softmax的結果，並依可能性大小進行排序，承接前小節對類別' new_whale'之權重為0.97的設定，選出前五名，觀察在結合不同模型的預測結果之下，對精度造成的影響(如表四)。

表四、ensemble結果

Ensemble	精度
模型A + 模型B	0.41897 (0.43734)
模型A + 模型C	0.42965 (0.44047)
模型B + 模型C	0.41237 (0.44047)
模型A + 模型B +模型C	0.40458 (0.44047)

4.3 誤差討論

對單純CNN架構的模型來說，仍無法解決部分種類資料量過少（一、兩張）的問題，導致預測的類別分佈與實際分佈相近，且預測的種類類別僅涵蓋1913種，而其中有六成就是原先training dataset中數量分布前1913個，可見此模型仍受到資料分布造成的誤差影響，表五列出兩份資料數量分布前十多的種類，前三多的種類是完全相合的。

表五、數量分布前十名

training data(9850)		testing data(15610)	
種類	數量	種類	數量
'new_whale'	810	'new_whale'	11837
'w_1287fbc'	34	'w_1287fbc'	38
'w_98baff9'	27	'w_98baff9'	20
'w_7554f44'	26	'w_eb0a6ed'	20
'w_1eafe46'	23	'w_fd1cb9d'	15

表五(續)、數量分布前十名

training data(9850)		testing data(15610)	
'w_fd1cb9d'	22	'w_73d5489'	13
'w_ab4cae2'	22	'w_ace8c54'	13
'w_693c9ee'	22	'w_c0d494d'	13
'w_987a36f'	21	'w_17ee910'	12
'w_43be268'	21	'w_654a5bb'	12

5. Conclusion

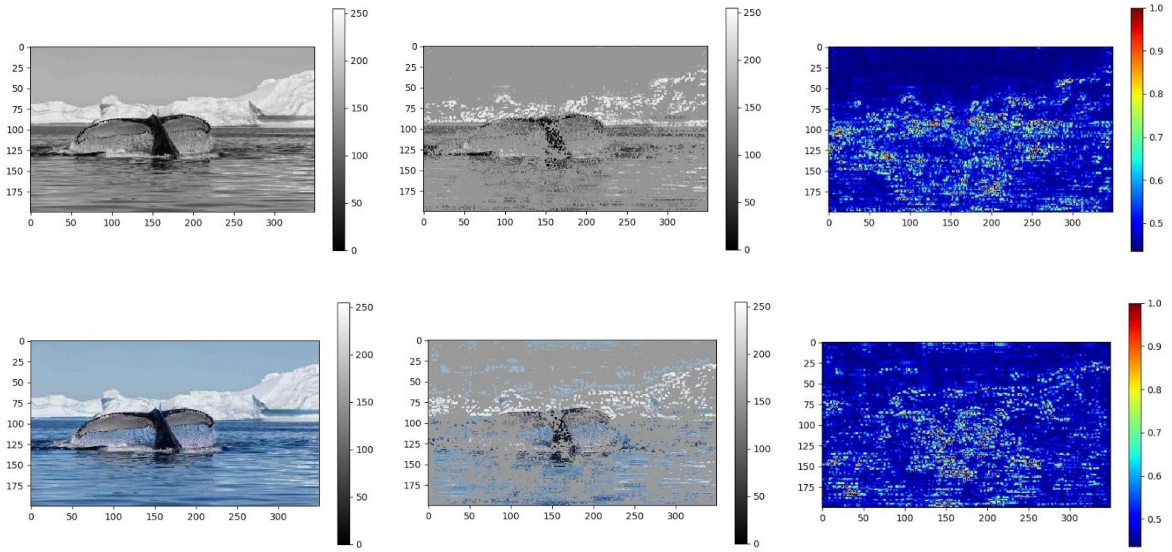
本團隊嘗試透過CNN以及siamese network搭配三種資料集，測試了四種實驗結果，最終以CNN+資料集C達到0.4407精度為最佳成果。本團隊發現，在相同的訓練參數下，CNN+資料集C比起CNN+資料集A更加準確、收斂速度更快，乃因資料集C之影像尺寸較資料集A小。另，siamese network最終沒有達到較佳的成果，收斂速度相當的緩慢，僅有0.39844之精度。

本團隊亦嘗試將四種模型做Ensemble之測試，最終發現並沒有帶來較佳之精度，推測應為難以去判斷各個模型該被分派之權重大小，原先較佳之預測反被其他較差之模型取代預測的結果。

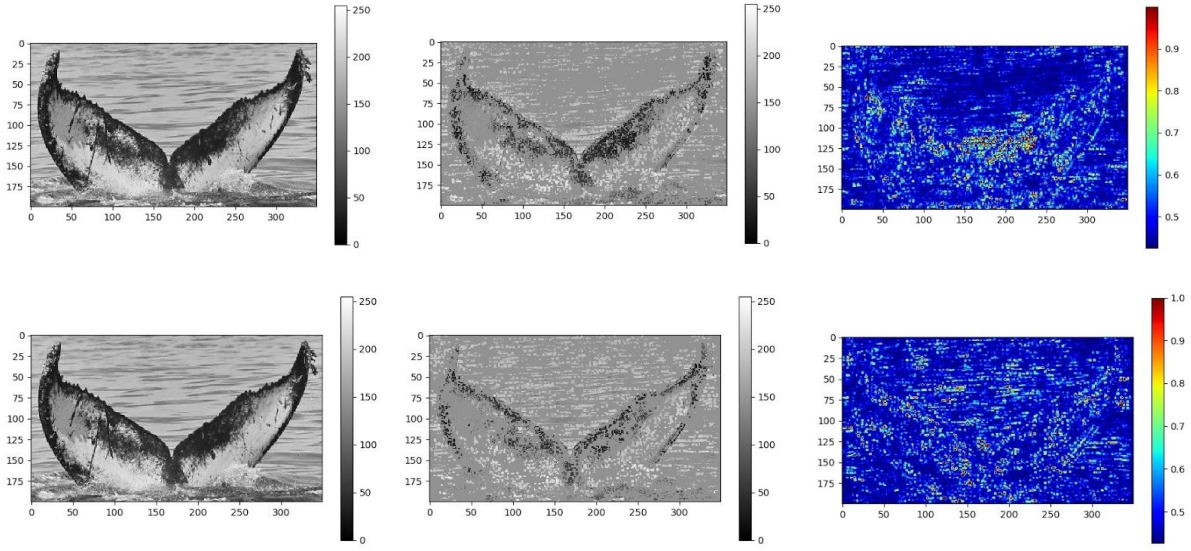
本團隊使用keras之套件，將影像透過旋轉、縮放等等之方式達到資料集增量，但此方法無法掌握其針對哪些照片做處理，故若能自行處理資料增量之問題，便能針對訓練資料較少之鯨魚種類做有效之處理。另，除了資料增量外，本團隊僅將照片做尺寸及顏色之修正，如圖四所示，由於我們是將整張照片做訓練之動作，發現其會勿將背景之雪山視為特徵，將會影響訓練之結果。若能針對影像中鯨魚之尾鰭部分擷取Bounding Box，將能使照片之重點能縮放在鯨魚尾鰭上，減少不必要之雜訊。

針對不同精度之模型製作saliency maps，如下圖四、五、六，上層為模型A，下層為模型B，其中模型A在相同的訓練參數下獲得較好之精度，且收斂速度快。從saliency maps可發現，模型A所擷取到的特徵皆比模型B更扼要明確，模型B雖為彩色資料集，但收斂速度不但緩慢，還容易擷取到不必要之雜訊，故最終模型A以叫好之

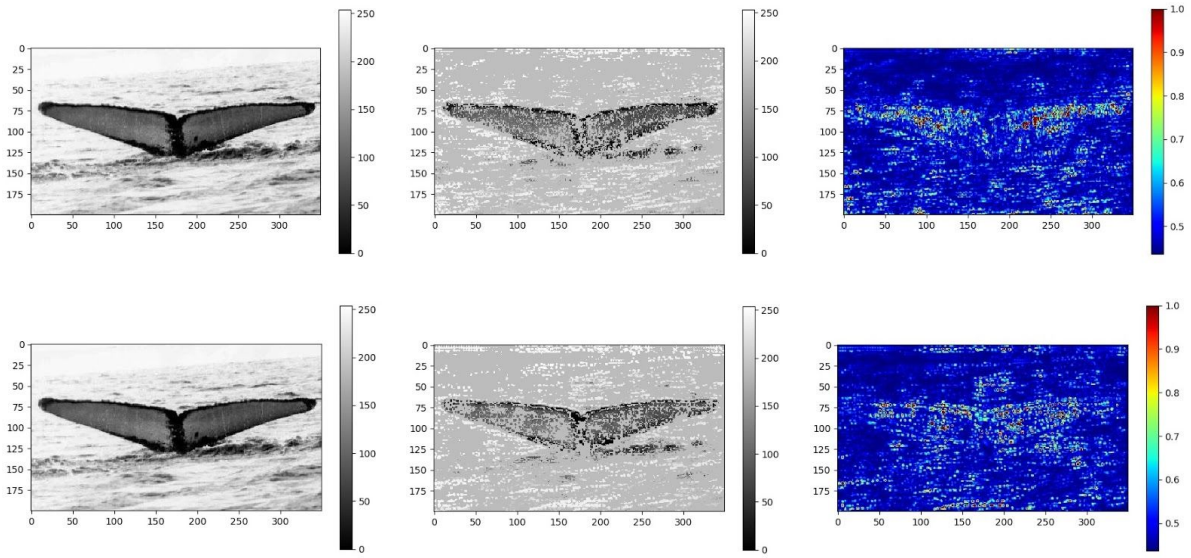
精度優於模型B。



圖四、saliency maps-1



圖五、saliency maps-2



圖六、saliency maps-3

Reference

[1] Lex Toumbourou, Humpback Whale ID: Data and Aug Exploration (2018, FEB). Retrieved from https://www.kaggle.com/lextoumbourou/humpback-whale-id-data-and-aug-exploration

[2] Robert Bogucki, Which whale is it, anyway? Face recognition for right whales using deep learning (2016, JAN). Retrieved from https://deepsense.ai/deep-learning-right-whale-recognition-kaggle/

[3] sorenbouma, keras-oneshot. Retrieved from https://github.com/sorenbouma/keras-oneshot