DSP 期末報告

資工碩一 R07922003 劉濬慶

資工碩一 R07922142 歐政鷹

Deep learning techniques for koala activity detection

- **前言**

此篇研究主要目的是透過偵測在大自然環境中的無尾熊(koala)的叫聲，以了解牠們的生態地區及數量，藉此幫助生態學家、動保團體和政府部門對無尾熊及其棲息地的保護工作。

此篇研究使用了卷積遞歸神經網絡架構（CNN + RNN），主要想法是使用 CNN 作為特徵提取器和 RNN 來模擬長期依賴關係。

- **Introduction**

為了監測野生動物，生態學家經常使用聲學傳感器作為收集數據的有效方法，記錄的聲學數據為生態學家提供了識別特定物種的手段，並根據動物交配時的叫聲可以進行棲息地中物種數量的調查。

隨著錄音裝置的普及，像 3C 產品上的麥克風，生態研究所產生的錄音數據數量遠遠超過可以手動分析的數量。現時生物聲學或生態聲學，已成為“大數據”研究領域之一。

在擁有大量數據的情況下，動物叫聲的自動識別系統已越趨重要，同時也是最近數十年來深入研究的主題之一。

此篇研究利用 Machine Learning 的技術，作出一個 Neural Network 來針對無尾熊的叫聲作自動識別。

此篇的 Detection 基礎技術基於：

(1) Energy threshold
(2) Spectrogram cross-correlation
(3) Hidden Markov Models (HMMS)

所以此篇還是前面是 HMM 處理後面才是把 CNN+RNN 結合在一起效果才會好，並不是全部都依靠 machine learning。

- **Datasets**

原始聲學數據(raw acoustic data)來自夜間時在澳洲新南威爾士州(NSW) Willi Willi 國家公園的 63 個站點的記錄數據，在每個站點中設置一個錄音裝置來記錄周邊至少 100M 範例內的叫聲。因為雄性無尾熊在繁殖季節的半夜會發出響亮的叫聲，所以我們將這些聲音紀錄當作數據，用來估計無尾熊的存在比例。

藉此收集了 2181 筆數據作為無尾熊叫聲的 Class，及 4337 筆數據作為非無尾熊叫聲的 Class，當中非無尾熊叫聲的數據來自其他聲音的錄音，例如：噪音、蟋蟀、青蛙和鳥類的叫聲、車輛駛過的聲音等。當中 80%數據用作 training set，20%數據用作 test

set。

研究中透過將 Time and frequency shifts 用作於 data augmentation 的方法來擴充 training dataset 以減少出現 overfitting 的情況。對於 time shift 上的處理，則是將 spectrogram 以隨機的時間切成 2 個部分，然後再將第 2 部分 spectrogram 排在第 1 部分 spectrogram 前面。

- **Feature extraction**

研究中主要用的方法是 **Constant Q Transform (CQT)** 來讓在時間與頻率訊號(time-frequency signal) 在 lower frequencies 上有更高的頻率分辨率，並在 higher frequencies 上有更高的時序分辨率。

步驟：

1. 先用 **anti-aliasing filter** 將訊號 downsampled 到原始採樣率(original sampling rate)的 $\frac{1}{12}$ 倍(1837Hz) ，並令分析的最高頻率f_{max}為新採樣率(new sampling rate)的一半(918Hz)，分析的最低頻率 $f_{min} = 124\ Hz$，中心頻率 $f_k = f_{min} * (2^{\frac{1}{b}})^k$
2. 然後使用 CQT 將 spectrogram 轉換為 log scale，公式如下：

$$Spectrogram(t,w)|_{dB} = 20\ log_{10}(\ |X^{CQ}(t,w)|\ )$$

經過轉換後，可得到 104 × 208 size 的 input features。

- **Models**

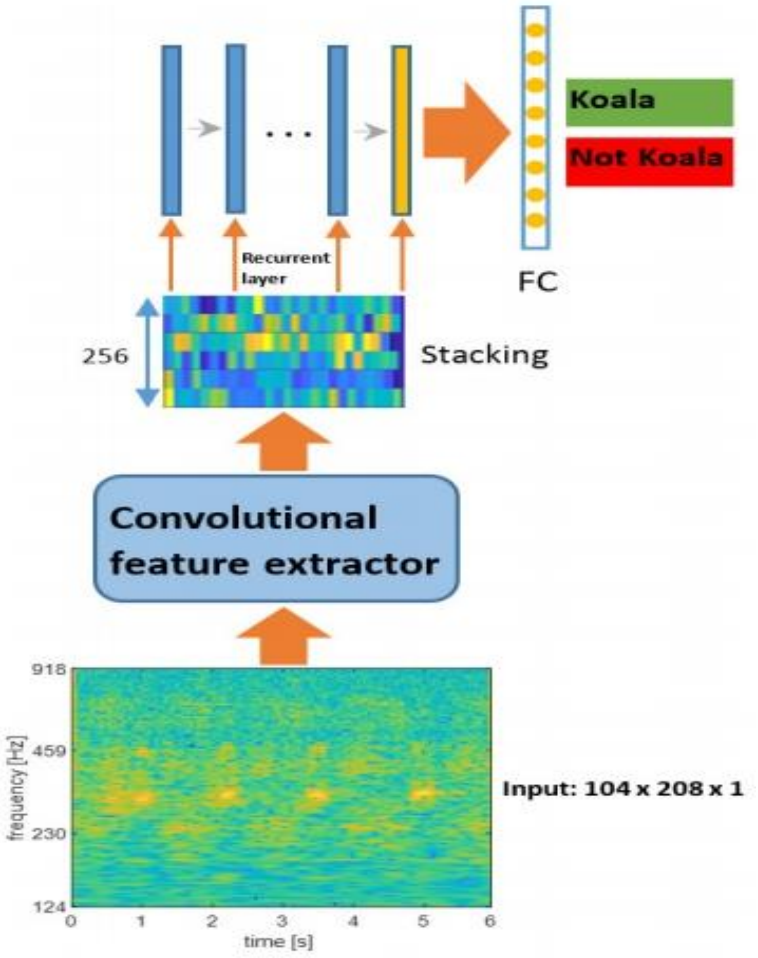

圖 1. 用於無尾熊偵測的 CNN + RNN 架構圖

研究中提出的整個 Netwok 架構如圖 1 所示，CNN 架構是由 3 個 convolutional layers 組成，以 Xavier initialization 作為參數初始化的方法，當中每一層使用了 3×3 的感知域(receptive field)，以 ReLU 作為 activation function，最後加入 4x2 的 max pooling 及 0.5 的 dropout rate。

最後一層 convolutional layer 的 output 會堆疊在頻率軸上

上再送到一個 LSTM layer 當中，LSTM 輸出後連接兩個 fully connected layers，最後再利用 softmax layer 來對結果進行分類。

整個 network 使用了 Adam 作為優化的 function，learning rate 為 0.001，保有 0.9 的動量(momentum)，並以 binary cross-entropy 作為 loss function。

表 1 展示出整個 Netwok Model 不同的資訊。

Type	Filter/Stride	Output	#Params
Conv1	3 x 3 / 1 x 1	104 x 208 x 32	320
MaxPool1	4 x 2 / 4 x 2	26 x 104 x 32	-
Conv2	3 x 3 / 1 x 1	26 x 104 x 64	18K
MaxPool2	4 x 2 / 4 x 2	7 x 52 x 64	-
Conv3	3 x 3 / 1 x 1	7 x 52 x 128	73K
MaxPool3	4 x 2 / 4 x 2	2 x 26 x 128	-
LSTM	-	64	82K
FC1	-	64	4K
FC2	-	2	130
Total			177K

表 1. 研究所提出的 CNN+RNN 架構

(The data shape indicates frequency × time × number of filters)

- **Experiments**

無尾熊叫聲辨識系統(koala call detection system)是透過曲線下面積(AUC)來對接收者操作特徵曲線(ROC 曲線)進行評估。

在 ROC 中 X 軸是 False Positives (FP)，Y 軸是 True Positives (TP)，曲線與 X 軸所圍成的面積就是 AUC。同時可以計算 ROC 的 average precision (AP)來作為另一個參考指標，$\mathrm{AP} = \frac{TP}{TP+FP}$。

此外，研究中提出了一個用作比較的 Baseline CNN 架構，與 CNN+RNN 的架構差別在於，最後一層 convolutional layer 輸出後是連接兩個 fully connected layers，最後再利用 softmax layer 來對結果進行分類，跳過了 LSTM 的處理，表 2 展示出 Baseline CNN Model 的架構及資訊。

Type	Filter/Stride	Output	#Params
Conv1	3 x 3 / 1 x 1	104 x 208 x 32	320
MaxPool1	4 x 2 / 4 x 2	26 x 104 x 32	-
Conv2	3 x 3 / 1 x 1	26 x 104 x 64	18K
MaxPool2	4 x 2 / 4 x 2	7 x 52 x 64	-
Conv3	3 x 3 / 1 x 1	7 x 52 x 128	73K
MaxPool3	4 x 2 / 4 x 2	2 x 26 x 128	-
FC1	-	1024	6.8M
FC2	-	2	2K
Total			6.9M

表 2. Baseline CNN 架構

(The data shape indicates frequency × time × number of filters)

- **Results**

Models	AUC	AP
CNN+RNN	0.9909	0.988
CNN	0.9908	0.988

表 3. AUC 及 AP 在 test data 上的分數

(The data shape indicates frequency × time × number of filters)

所有 models（CNN+RNN & baseline CNN）都是用 five-fold cross validation 來進行評估，每一個 fold 的 data 都有一次會被用作 test data，其他時候會被用作 training data 的一部份，並以 5 次 validation 結果取平均來取得最終的結果。

表 3 的結果是三次 separate cross-validation 後平均的測試分數。每次 cross-validation 用了不一樣的 random seed 來對 network 參數作 initialized。

為了更好地驗證 model 在不同情況下的準確性，研究還準備了另外一組從其他地方獲得的數據來對 model 進行測試(但原文中沒有提到數據來源站點是否原來的 63 個站點之一)，這一組 test data 裡有 10 隻無尾熊的叫聲出現，圖 2 及圖 3 分別顯示了 CNN + RNN 和 baseline CNN 模型對這一組 data 的分數(是無尾熊叫聲的機率)。從中可以看到除了 43 分~45 分那次的叫聲完全沒被辨識出，以及 CNN 在第一次

辨識分數較低外，其他情況都可以辨識出無尾熊叫聲，另外研究中有提到 43 分~45 分那次的叫聲可能是由於飛機發出的噪音掩蓋了無尾熊的叫聲。

可以看出 CNN+RNN model 對無尾熊叫聲的辨識有著不錯的效果，而研究中也指出他們再利用其他非原來站點地區所收集到的數據來進行測試也獲得了不錯的效果。

研究指出 CNN 模型很難獲取長時間依賴性(long temporal dependencies)的資訊，因此他們提出 CNN+RNN 架構來解決這個問題，使 model 能在不同環境中辨識無尾熊的叫聲。

研究最後指出，他們認為 CNN + RNN 框架是檢測野生無尾熊叫聲的首選方案，其次這個框架還可以用於檢測其他動物叫聲，例如與標準 CNN 相比具有優越性能的鳥聲檢測。

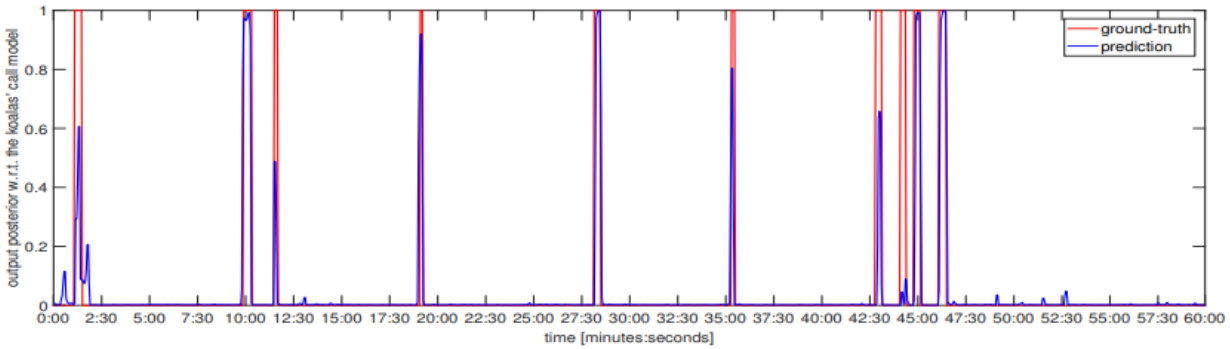

圖 2. CNN+RNN 在 1 小時的數據中對無尾熊叫聲的辨識結果

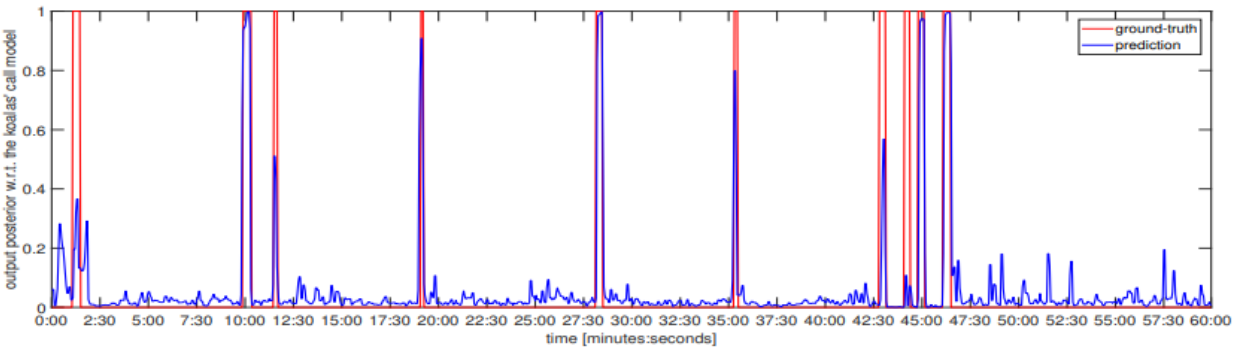

圖 3. CNN 在 1 小時的數據中對無尾熊叫聲的辨識結果

- **心得**

劉濬慶 - 心得：

由於自己很重視大自然的環境保護，也很喜歡動物，還有對 CNN 以及 RNN 這些 Deep learning 方法深感興趣，所以才選擇這篇論文。經過老師的教導之後本篇論文看得非常明瞭，幾乎 90%老師上課都有講過提到，所以讀起來很順利。能用語音相關的知識去幫助大自然的環境實在是一件太棒的事情了。

歐政鷹 - 心得：

去年我曾在一場演講聽到一個將聲音用於海底生物探測的方法，與這篇論文有異曲同工之妙。在以前我都未有想過語音訊號辨識的技術可以用來做這些方面的事情，而我認為這篇論文中所提出 network 架構很基本，只是 3 層 CNN+LSTM+2 層 FC，若他們使用較複雜 network，說不定 model 會有更好的結果，論文中未有深入提到 network 複雜度對 model 的影響，實在可惜。

- **參考資料**

i. Himawan, M. Towsey, B. Law, P. Roe “Deep Learning Techniques for Koala Activity Detection” : (https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1143.html)

ii. Hung-yi Lee, Machine Learning course “Convolutional Neural Network & Recurrent Neural Network” : (http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML17_2.html)

iii. Wikipedia “Constant-Q transform” : (https://en.wikipedia.org/wiki/Constant-Q_transform)

iv. Wikipedia “Receiver operating characteristic” : (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)