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 倍(1837Hz) ，並令分析的最高頻率為新採樣率(new sampling rate)的一半(918Hz)，分析的最低頻率，中心頻率
2. 然後使用CQT 將 spectrogram 轉換為log scale，公式如下：  
   經過轉換後，可得到104 × 208 size的input features。

* **Models**

|  |
| --- |
| C:\Users\劉濬慶\AppData\Local\Microsoft\Windows\INetCache\Content.Word\結構.jpg |
| 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。

|  |
| --- |
| 更像的 |
| 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，。

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。

|  |
| --- |
|  |
| 表2. Baseline CNN 架構  (The data shape indicates frequency × time × number of filters) |

* **Results**

|  |
| --- |
| 1547400054538 |
| 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**

1. 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>)
2. Hung-yi Lee, Machine Learning course“Convolutional Neural Network & Recurrent Neural Network”：  
   (<http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML17_2.html>)
3. Wikipedia “Constant-Q transform”：  
   (<https://en.wikipedia.org/wiki/Constant-Q_transform>)
4. Wikipedia “Receiver operating characteristic”：  
   (<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>)