數位語音處理概論 Final Project

R10922132 資訊工程研究所 碩一 吳峻銘

Topic : Spoken Keyword Spotting

簡介

現今各大科技公司都有推出語音助理的產品，如 Apple 的 siri、Amazon 的 Alexa 等，其中在喚起語音助理服務時我們可以透過語音輸入個關鍵詞來啟動語音助理，如我們要打開 siri 只要對裝置說”hey siri ”就可以開啟，這種技術叫做 keyword spotting，因為平常生活都會使用到這項功能，加上課堂上已學習語音辨識的基礎知識，最後覺得蠻有興趣的想要深入了解，所以計畫以[1]介紹的技術概況去深入探討 keyword spotting 的技術原理以及研究走向。

方法

Keyword spotting 的技術可以大概把它分為引入 deep learning 之前和引入 deep learning 之後，在引入 deep learning 之後模型可以學習到更多語音的特徵，也更具強健性，現今大多數商用的 keyword spotting 技術大多都有引入 deep learning，以下將大略簡介兩者比較著名的方法。

- **before deep learning**

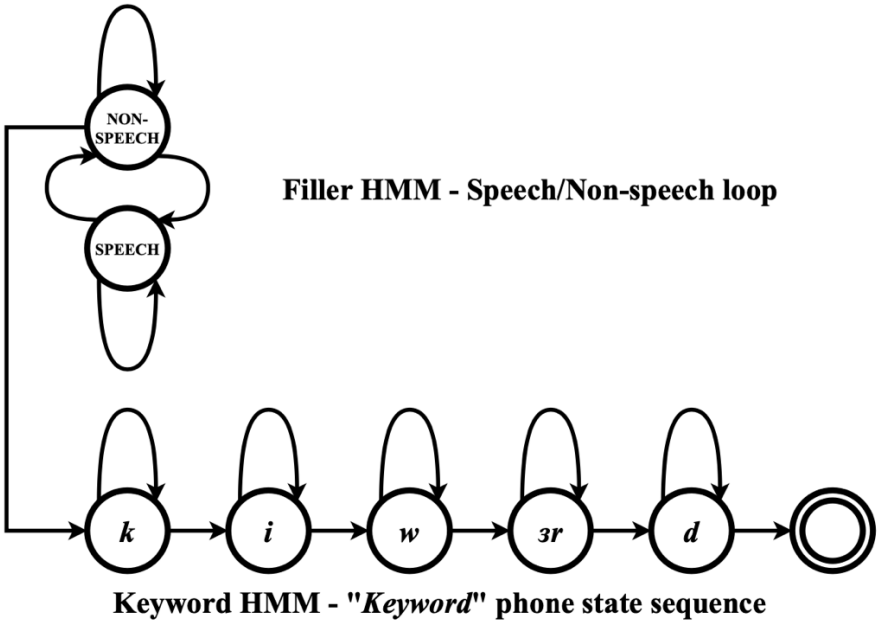
1. large-vocabulary continuous speech recognition [2][3][4]

此項技術就是把語音的訊號編碼成一組網格(lattice)，再從這組網格裡去收尋我們想要的關鍵字的路徑，這項技術的優點是切換關鍵字的彈性很大，因為主要是基於在網格裡搜尋，但是這一項技術的缺點是計算複雜度會很大，因為我們要產生一組網格，會花費很多時間，所以現今的 keyword spotting 大多不會使用此項方法。

2. keyword/filler HMM [5]

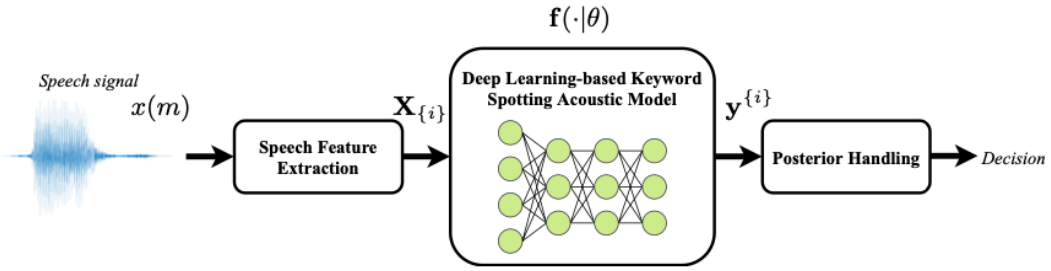
這方法就是我們課堂上所熟悉的 HMM，主要把聲音分成是 keyword 和不是 keyword 的聲音訓練出兩個 HMM，接著再去觀察語音訊號在兩個模型的可能性

(likelihood)分數，當 keyword 的分數高於不是 keyword 一個門檻就可以啟動語音助理的系統。



- **deep learning**

模型架構

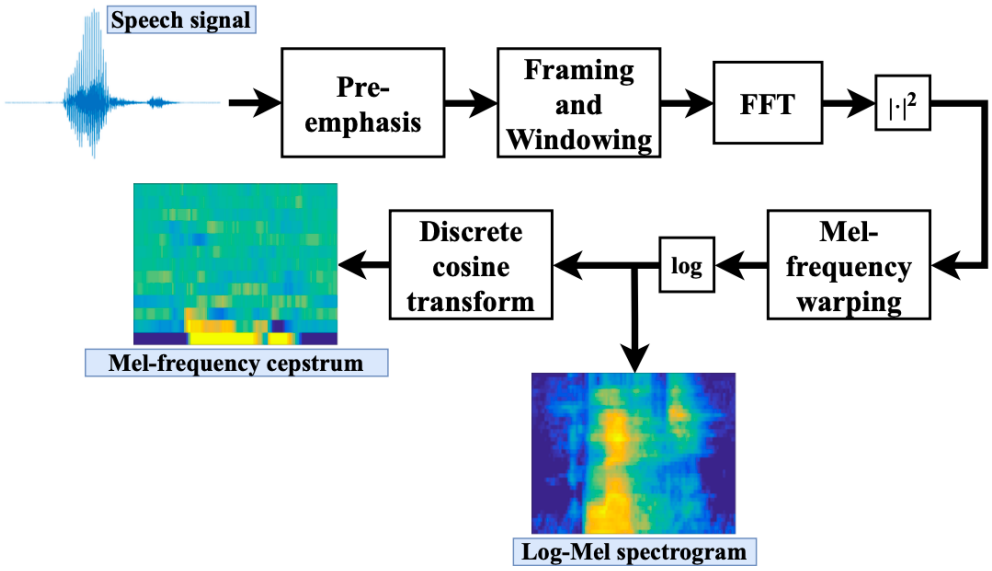


現今大多數的 deep learning keyword spotting 模型可以大概分成三個部分，首先是要從語音訊號中取出重要的資訊，再來就是要把這些重要的資訊丟到深度學習的網路裡面去預測每個時間段的聲音是屬於哪一種類別，再來最後一個步驟把前一步所預測出的類別機率去做處理進而得到最終的結果，其中每一個步驟都有很多不同的變化，以下將簡單列出幾項比較特別的。

1. Speech feature extraction

A. Mel-scale-related feature

這是課堂上有介紹的 MFCC 方法，也是最常用的取聲音訊號特徵的方法，聲音訊號藉由一系列的訊號處理最後可以變成一個同維度的向量。



B. RNN feature

遞迴神經網路是一個著名的神經網路架構可以用來處理時序序列的問題，他可以把不同長度的資料輸出成同維度的向量，因為聲音訊號是時序資料使用遞迴神經網路來萃取資料可以有效地抓出有用的特徵。

C. learnable filterbank feature

對於一個端對端深度學習系統，最好的特徵是透過學習而得來的，所以輸入原始的聲音訊號再透過預訓練的過程去取得聲音的特徵，這種發法可以有效最佳化端對端深度學習系統的效能。

2. acoustic model

A. FFNN

前饋神經網路是最簡單的神經網路架構，最早採用深度學習的 keyword spotting 模型[6]也是使用前饋神經網路，此架構即是把所有的特徵經過數個全連結層去判斷此聲音訊號的類別

B. CNN

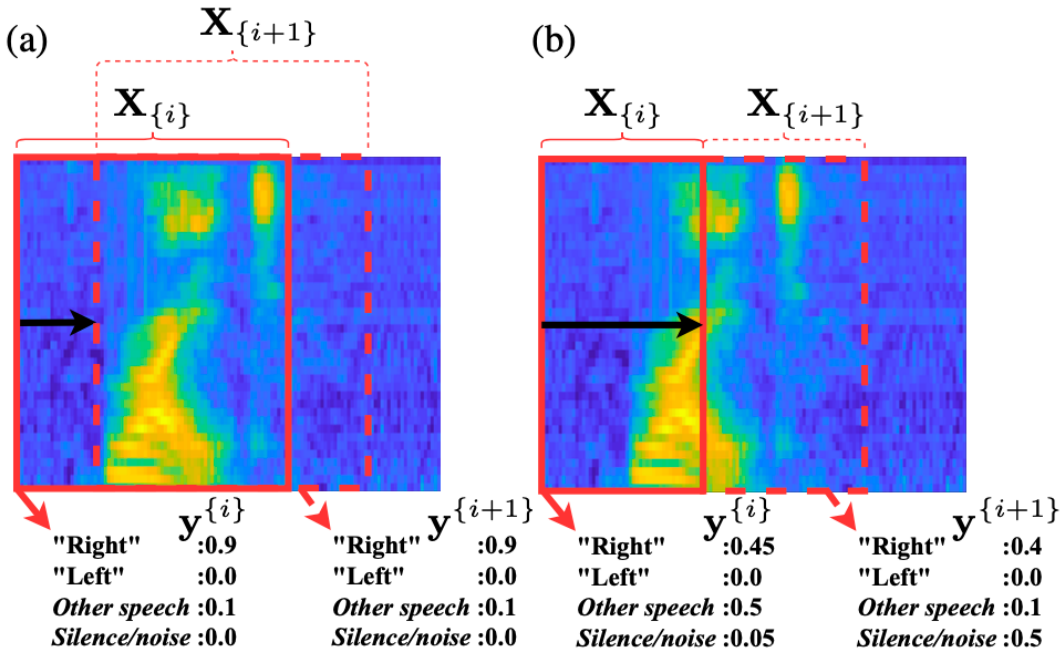
CNN 多數人是用來處理影像的問題，然而也可以把它用來處理聲音訊號的問題，透過聲音頻譜的輸入可以有效判斷出聲音頻譜圖的特徵，進而預測出最終的結果，此方法也有不錯的效果。

C. RNN

RNN 主要是用來處理時序的資料，所以在語音識別中也常常用 RNN 來預測語音訊號的類別，然而 RNN 的時序方向是由左至右會沒有考慮到一些前後文的語音資訊，所以近期引進的注意力機制(attention)[7]可以有效地處理前後文語音資訊。

3. Posterior handling

在前一個步驟所處理的類別預測機率不具強健性，如下圖所示在(a)部分中由於我們要保持聲音連續的資訊，我們每個偵測聲音訊號的頁面都會有一些重疊的部分(X_i ,X_{i+1})，所以當 X_i 有 keyword 時 X_{i+1} 有 keyword 的機率也會跟著提高，然而 X_{i+1} 不一定是具有 keyword 的音訊，如此就會造成錯誤的回報，此時就會想如果讓每個頁面不要重疊就好了，但是如果頁面不重疊就會造成聲音取樣的訊號太少(X_i 和 X_{i+1} 隔太遠)，如此也會造成兩者(X_i ,X_{i+1})都無法正確辨識出 keyword 訊號的問題，所以基於這個問替我們必須要對前面的機率向量做後處理。



資料集

在語音辨識領域中已有許多資料即可以供研究訓練，如 WSJ corpus[8]等，然而這些都不適合用在 keyword spotting 領域中，因為他們沒有特別設計關鍵字的語音資料集，其中 keyword spotting 中最有名的資料及是 google speech command dataset[9]，現在他有兩個版本(v1, v2)，第二個版本有更多關鍵字資料及也更大一

點，以下簡單列出兩個版本的關鍵字表，而在目前為止在這個資料集表現最好的模型是[10](第一版)[11](第二版)。

Version 1 (*v1*)	Version 2 (*v2*)	yes	no	up	down	left	KW
		right	on	off	stop	go	
		zero	one	two	three	four	Non-KW
		five	six	seven	eight	nine	
		bed	bird	cat	dog	happy	
		house	Marvin	Sheila	tree	wow	
		backward	forward	follow	learn	visual	

未來展望

經過探討 keyword spotting 的基本技術後已有了一些基本瞭解，接下來將簡單列出一些自己在了解這些技術後並結合一些已有的知識去發想 keyword spotting 未來有可能的研究發展。

知識蒸餾

在 keyword spotting 中我們已經可以在現有的資料及上訓練出不錯的結果，然而部署到商用產品中必須要考慮 keyword spotting 的耗能程度，因為我們總不希望多了這個功能卻因為耗能太大而導致裝置沒電，因此需要將現有的模型去淬取出重要的部分，同時也要保持良好的辨識表現，其中知識蒸餾就是一種技術可以從巨大的模型取出重要的知識，這對手持裝置會很有幫助。

語音-影像關鍵字偵測

因為現實環境中可能會有環境音過於吵雜的狀況發生，所以模型的表現就會大受影響，如過這時候我們可以結合影像辨識的技術一定會大幅解決這個問題，透過抽取出臉部的特徵以及在講話時嘴唇的運動模式，再透過有效的結合回有更強健性的表現。

不平衡分類問題

在這個任務中我們的標記是不平衡的，因為關鍵字總是會比非關鍵字少，所以初裡不平衡分類問題就顯得很重要了，常見的方法是調整目標函式的權重和把多數

的類別物件減少一些和把少的類別做資料增強，甚至還可以用半監督學習去從後面改善這問題，其中還有很多方法是可以解決這問題的，如果能解決不平衡資料的問題模型的強健度勢必會有所提升。

Reference

[1] López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep Spoken Keyword Spotting: An Overview. *arXiv preprint arXiv:2111.10592*.

[2] Chen, G., Yilmaz, O., Trmal, J., Povey, D., & Khudanpur, S. (2013, December). Using proxies for OOV keywords in the keyword search task. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 416-421). IEEE.

[3] Miller, D. R., Kleber, M., Kao, C. L., Kimball, O., Colthurst, T., Lowe, S. A., ... & Gish, H. (2007). Rapid and accurate spoken term detection. In *Eighth Annual Conference of the international speech communication association*.

[4] Weintraub, M. (1993, April). Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 463-466). IEEE.

[5] Sun, M., Snyder, D., Gao, Y., Nagaraja, V. K., Rodehorst, M., Panchapagesan, S., ... & Vitaladevuni, S. (2017, August). Compressed Time Delay Neural Network for Small-Footprint Keyword Spotting. In *Interspeech* (pp. 3607-3611).

[6] Chen, G., Parada, C., & Heigold, G. (2014, May). Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4087-4091). IEEE.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

[8] Paul, D. B., & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

[9] P. Warden. (2017) Launching the Speech Commands Dataset. [Online]. Available: https://ai.googleblog.com/2017/08/ launching- speech- commands- dataset.html

[10] Kim, B., Chang, S., Lee, J., & Sung, D. (2021). Broadcasted Residual Learning for Efficient Keyword Spotting. *arXiv preprint arXiv:2106.04140*.

[11] Banbury, C., Zhou, C., Fedorov, I., Matas, R., Thakker, U., Gope, D., ... & Whatmough, P. (2021). Micronets: Neural network architectures for deploying tinyml applications on commodity microcontrollers. *Proceedings of Machine Learning and Systems*, *3*.