AMD PYNQ 人工智慧終端節點運算創意競賽

作品名稱: 基於 ShuffleNet 之道路事件聲音識別輕量化邊緣智慧系統

學校:_____國立中央大學_____ 系所:_____電機工程學系_____

指導老師:____陳聿廣___

參賽隊員: 鄭凱方、戴仕庭、吳育丞、呂翊銓

E-mail Address: kevin71xb37@g.ncu.edu.tw, taishihting@g.ncu.edu.tw, mineeric@g.ncu.edu.tw, yichuan921030@g.ncu.edu.tw

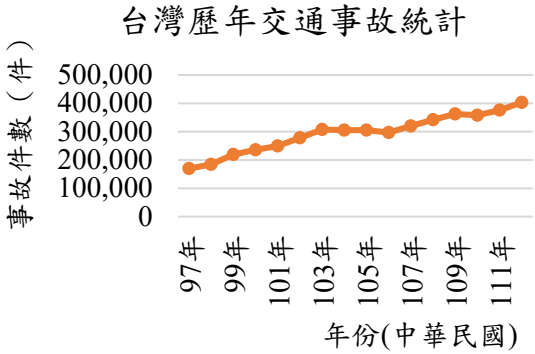
一、摘要

在現代社會中，隨著交通流量急速增加，道路事故發生頻率也相應上升。並且，當今事故通報流程仍十分繁瑣，需要人為介入才能進行，若是意外發生在比較偏僻的地區，沒有目擊者幫助傷者報案，就有可能導致延誤送醫而發生憾事。然而，現有監視系統存在視覺死角，後端影像辨識難以感知監視器鏡頭以外的事故發生，這對即時識別和救援效率提出了挑戰。另外，透過雲端運算搭建的偵測系統需要持續回傳資料到數據中心進行運算，過程中會耗費大量的傳輸成本，但是對於緊急狀況來說，即時提供反應十分重要。

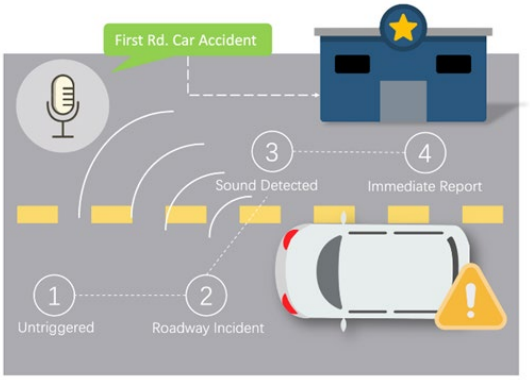
為解決上述問題，我們提出—基於 ShuffleNet 之道路事件聲音識別輕量化邊緣智慧系統，在軟體端進行參數調整與模型訓練，並在 PYNQZ2 FPGA 開發版建立音訊預處理與神經網路加速器，最後匯入軟體端訓練後的參數，完成軟硬體整合。在車禍發生時進行實時通報，簡化通報流程，並透過搭建邊緣裝置直接取得偵測結果，減少網路傳輸需求，降低整體的能源消耗，亦利用聲音傳播特性補足視野死角問題，形成交通事故監測之完整解決方案，我們期望此作品能藉軟硬體整合開發達成高度輕量化與高準確度的車禍聲音識別，實際使用於人煙稀少或易發生道路事件的路段，補足傳統監視器存在的視野死角問題。

二、動機及相關技術

根據交通部統計，全國交通事故總件數從民國 100 年的 235,776 件急遽增長至 111 年的 375,844 件，交通事故受傷的人數平均每年也增加約 16,834 人，趨勢如圖(一)。面對日益增長的交通事故，監控系統能否即時發現、通報並展開救援成為重要的道路安全議題。因此監控系統若能具有準確的事故偵測與自動通報能力，將有望提升救援團隊應對事故的效率。



圖(一) 台灣歷年交通事故統計



圖(二) 專題成品概念圖

目前的道路監控系統已可在遠端監控中心進行影像辨識，有效辨識車禍事故的發生並迅速展開後續的救援行動，縮減對人工通報的依賴，提高交通事故處理相關機構的反應速度。然而，傳統監視器卻存在「視覺死角」的限制，這

意味著監視器可能在特定情況下無法拍攝到完整的情況，顯示當今許多交通事故仍需要依賴目擊者的通報。因此本團隊提出道路事件聲音識別系統，充分利用聲音傳播「無死角」的特性偵測到以往無法觸及的死角，成品概念圖如上方圖(二)所示。另外，我們將在監視器端進行邊緣運算，即時偵測事件發生，避免等待訊號回傳到監控中心經過處理再進行辨識。如此可以節省寶貴的反應時間，有助於加速傷患的救援行動，提高即時救助的效果。

具體而言，本系統融合軟硬體整合開發。軟體端進行高準確率 ShuffleNet 神經網路參數訓練，其中包含資料集建立與預處理，須同時兼顧高準確率及神經網絡的運算複雜度，滿足邊緣運算需求。硬體端則需匯入軟體端訓練好的模型參數並建立高度輕量化的 ShuffleNet 神經網路模型架構，改善當前檯面上高成本的聲音辨識硬體架構。因此，本團隊之目標是在資源有限的 PYNQZ2 開發板上追求設計高效且輕量化的神經網路聲音辨識模型，以符合實際環境需求。

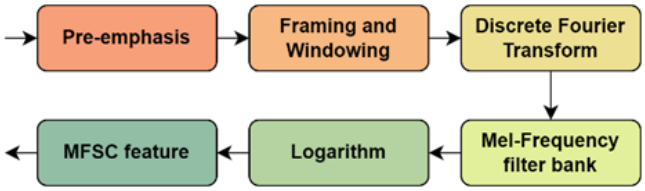
關於文獻部分，我們將針對欲使用之技術－(A)聲音預處理技術(MFSC)、(B)ShuffleNet 神經網絡以及(C)聲音辨識相關文獻進行技術方面的探討和比較。

A. 聲音預處理技術

聲音訊號主要可分為時域與頻域兩種。時域分析關注訊號在時間軸上的整體變化，將所有頻率成分綜合在一起以觀察訊號參數如何隨時間變化。另一方面，在頻域分析中，訊號被分解成一個或多個頻率分量，每個分量的功率參數都會被詳細展示，使得分析者能夠明確了解訊號中各個頻率成分的特性。這些特徵提取技術可以被分為兩類[1]，第一種是依「語音產生方式」，例如：線性預估編碼(LPC) [2]；第二種是依「語音感知」，例如：取梅爾頻率倒譜係數(MFCC)與本團隊欲採用之梅爾頻率譜係數(MFSC)技術。

- MFSC 介紹

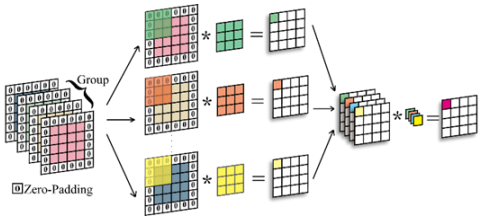
梅爾頻率譜係數(MFSC)流程如圖(三)，預強化(Pre-emphasis)補償訊號高頻損失，是整個音頻處理的首要步驟。接著將訊號拆分成幀(Framing)實現對訊號的穩定分析；隨後進行加窗(Windowing)操作，可依需求選擇漢寧窗或漢明窗，並搭配適當窗口大小和重疊時間，有助增強諧波並減少邊緣效應。透過離散傅立葉變換(DFT)，將訊號由時域轉為頻域，並計算訊號的頻譜功率分佈。Mel 帶通濾波器是一組基於音高知覺的濾波器，似人耳對語音的感知，旨在提取語音訊號的非線性表示，慣例 Mel 濾波器組由 40 個三角形濾波器構成[3]。最後，直接對過濾後頻譜資料取對數而不使用離散餘弦變換(DCT)，相較於 MFCC，這不僅簡化計算過程，同時也保留更多原始語音數據[2]，有效地保持頻譜能量的局部特性。MFSC 直接使用對數能量的方法為語音特徵提取提供一種有效的方案，在特定應用中可能更適用於保留訊號的本地特性。基於其對 MFCC 的運算簡化及特徵存取特性，因此我們採用 MFSC 作為本系統之聲音預處理工具。



圖(三) MFSC 流程圖

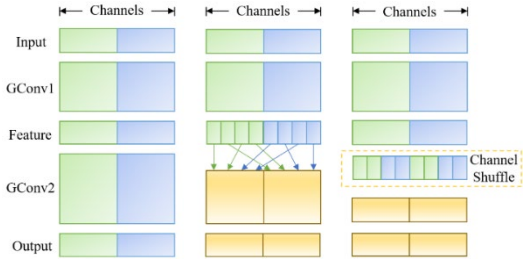
B. ShuffleNet 神經網絡

在硬體實時邊緣運算下，神經網絡的運算特性與參數數量多寡尤為重要，因此[4]提出 MobileNets，此網絡不僅擁有較輕便的模型參數數量，也兼顧較高的運算速度以追求能在行動裝置(Mobile)上實行，符合本作品之邊緣運算需求。其核心概念是將傳統卷積的提取特徵過程拆成 Pointwise Convolution 和 Depthwise Convolution 兩步驟，合稱深度可分離卷積，使模型能維持性能亦大幅降低計算複雜度和模型大小，適用於資源有限的設備，運算過程如圖(四)。



圖(四) Pointwise & Depthwise Convolution

ShuffleNet [5]是基於 MobileNets 的概念發展的高效神經網路架構，特別是針對邊緣裝置的需求設計。它結合了群組卷積 Group Convolution 和通道洗牌 Channel Shuffle 兩種技術來達到高效的模型輕量化，同時保持良好的模型表現，其概念實現方式如下圖(五)。



圖(五) Group Convolution & Channel Shuffle

群組卷積將傳統卷積拆分成兩個部分，首先對各個通道進行單通道的獨立二維卷積，取得區域的特徵；接著，使用 1×1 的卷積核對這些輸出進行卷積運算，強化特徵之間的關聯。最後，通過通道洗牌促使資料交換，從而使模型可以在大幅減少運算量和參數儲存量的前提下學習更複雜的特徵。

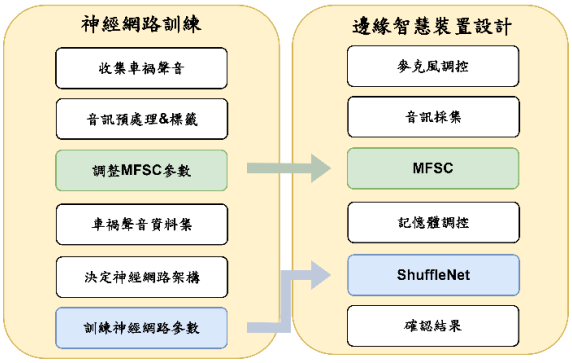
[6]在 FPGA 上成功開發 ShuffleNet 神經網絡，利用咳嗽聲音特徵進行疾病種類的辨識，最終可在 7.9K 個 LUT、12.9K 個 Flip-Flop 和 41 個 DSP 達到了 86.5% 的精確度，顯示出 ShuffleNet 在硬體端的潛力。

C. 聲音辨識文獻回顧

聲音作為自然界中常見的訊號蘊含豐富的資訊，協助人們發現潛在危險並適時做出反應，因此聲音識別技術被廣泛應用於各領域。其中語音視覺化[7]運用自然語言處理(NLP)將語音即時轉換成文字，並提高 67%正確率，幫助吶語症患者正常生活。聲音事件追蹤和定位[8]運用 T 型麥克風陣列在嘈雜環境中，仍有 93%精準率可定位聲音發生來源。另有融合聽覺與視覺的事件辨識[9]，藉由 Audio-Video Concurrence (AVC)與 K-nearest neighbors (KNN)應用，使影像結合聲音再進行辨識讓深度學習模型有更準確且更有意義的判定結果。由此可知，聲音辨識是極具潛力的道路車禍事件偵測方法，並且可克服影像辨識有視覺死角的弱點，可見此系統開發的必要性。

三、作品設計概念

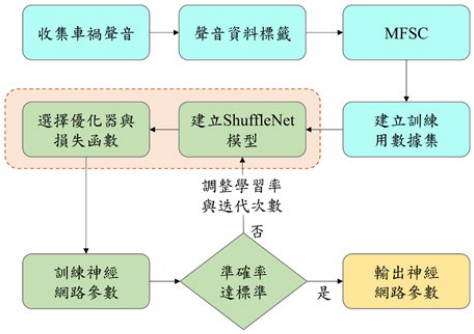
本系統結合軟硬體開發，架構如圖(六)。軟體部分涵蓋建立及預處理資料集並訓練出高辨識準確度的 ShuffleNet 神經網路，同時還需考慮運算複雜度以滿足邊緣運算的需求。硬體部分則將訓練好的模型參數整合至輕量化的 ShuffleNet 架構中，提升聲音辨識系統效能。



圖(六) 整體系統架構圖

A. 神經網路訓練

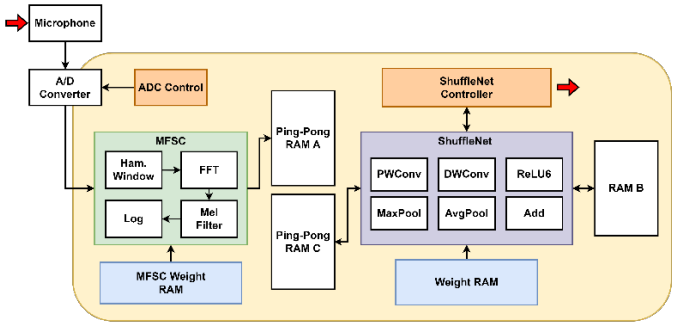
欲將聲音辨識的卷積神經網路實現於硬體，需在軟體端預先訓練卷積網路的所有參數，再將參數匯入硬體端的記憶體中。現今聲紋辨識的主要辨識資料為音訊頻譜，然而我們通常只能直接獲得一段時間區間中的聲音訊號，因此我們需要將時域的聲音訊號轉為頻域的頻譜分布，此步驟稱為「聲音預處理」。現今以 Mel-frequency spectral coefficient(以下稱 MFSC)聲音預處理流程最廣為使用，因此本系統透過 MFSC 將音訊資料轉換為頻譜，再將頻譜作為音訊資料集進行卷積神經網路模型的訓練與測試，神經網路訓練流程圖如圖(七)。



圖(七) 神經網路訓練流程圖

B. 邊緣智慧裝置設計

此部分通過外接麥克風實現音訊採集取得類比聲音訊號，再經由控制 FPGA 內配置的類比數位轉換器 Analog-to-Digital Converter (ADC)轉換為數位訊號；接著因頻譜訊號更利後續卷積神經網絡辨識，我們使用聲音預處理 MFSC 將時域訊號轉為頻域訊號；最後在狀態機與記憶體控制下，將即時的音頻特徵圖傳入 ShuffleNet 神經網路進行辨識，系統整體架構如圖(八)。



圖(八) 車禍聲音識別整體硬體系統架構

我們期望可以將本系統應用於許多不同場域，首先，將此系統安裝在偏遠地區或人煙稀少的路段，可以降低人為通報依賴。另一方面，在易發生道路事件的路段加裝此系統可加速通報流程，最後，在裝設傳統監視器的路段，透過運用音訊偵測的特性可消除監視器視線死角問題。我們期望在資源有限的 FPGA 開發板上設計出高效且輕量化的神經網路聲音辨識模型，簡化通報流程，利用聲音傳播特性補足視野死角問題，形成完整交通事故監測解決方案。

四、可行性分析

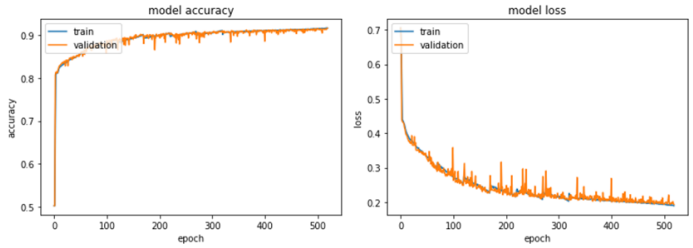
本系統運作流程如同上述所示，我們先在軟體端建立及預處理聲音資料集並訓練出高辨識準確度的 ShuffleNet 神經網路，同時盡量調整模型結構降低硬體實作神經網路架構的成本。另一方面，在硬體運作時，我們會先藉由外接的麥克風採集音訊，再聲音預處理 MFSC 流程將時域訊號轉為頻域訊號，最後在狀態機與記憶體控制下，將即時的音頻特徵圖傳入 ShuffleNet 神經網路進行辨識，最後透過 FPGA 版上的燈號顯示辨識結果。

我們使用 PYNQZ2 FPGA 開發板作為開發平台，此開發板搭載 Zynq-7000 SoC，有約 5 萬 LUT、10 萬 Flip-Flop 和 220 個 DSP 單元，並擁有 630 KB 的 BRAM 儲存空間，音效卡為 ADAU 1761 並且內建 3.5mm 音效接口與 ADC，擁

有音訊接收的能力及足夠的運算單元負荷音訊預處理及 ShuffleNet 的運算，和足夠的儲存空間儲存 ShuffleNet 模型參數，相當適合用於本系統的開發。

另外，論及實際應用場域方面，道路大多時間相當嘈雜，所以事件的聲音樣本中常帶有大量雜訊，這會使辨識的準確度大為降低，故[10]提出 Unusual Occurrences in Audio Forensics Database (UOAFDB)資料集，並為各種特殊事件聲音加上 15 種背景環境音效，運用 MFCC 與 CNN，辨識正確率可達到 81.50%。不同於上述論文僅為軟體層面實作，[11]致力於聲音辨識的邊緣化運算，優化 MFSC 與 CNN 硬體設計，成功減少 65.63%的 CNN 參數資料儲存空間，實現可在 FPGA 運行的輕量化運算。

我們在軟體部分也進行神經網路模型初步訓練，我們蒐集專業音訊資料庫 Soundsnap 及開源資料庫 ESC-50 車禍事件資料進行訓練，並採用 k-fold 訓練方式，測試後發現在 k=5 時訓練效果最佳，準確率高達 92.8%，loss 為 0.166，顯示模型收斂至最佳解，訓練成果如圖(九)，顯現聲音辨識車禍事件的可行性。



圖(九) 神經網路模型初步訓練結果

上述兩先前研究成果及我們在軟體的神經網路模型訓練成果顯示了 MFSC 聲音預處理方法適用於道路事件的聲音辨識，並且，研究中也佐證了用聲音辨識進行道路事件偵測的可行性，因此，本系統的開發具有嚴謹的證據支撐其可行性，由於可解決影像辨識的辨識死角問題，亦存在開發本系統的必要性。

五、系統實現之預期成果

如同上述所示，我們的系統計畫結合軟硬體開發，必須同時考慮神經網路模型訓練的準確度及硬體的輕量化和功耗大小。我們將系統實現方式大致分為神經網路訓練及邊緣智慧裝置設計進行描述。

A. 神經網路訓練

- MFSC

為了提升音訊資料集對模型訓練的成效，我們會先對資料集做 MFSC 音訊預處理，此處關鍵為梅爾過濾器，此濾波器是根據人耳聽覺特性設計，為非線性濾波，對於不同頻率的聲音敏感度不同。換句話說，梅爾濾波器由多個不同中心頻率的脈波重疊而非像大多數過濾器由覆蓋所有頻率的單一波形構成，有助於去除特定頻率範圍的資料，獲得更具代表性的頻譜特徵。梅爾濾波器組的公式如下所示，m 為濾波器編號，f_m 為第 m 個濾波器的中心頻率。從下方公式可看出在第 m 個濾波器中，若頻率 k 不在頻率 f_{m-1} 及 f_{m+1} 間，則完全過濾該頻率。反之，若頻率 k 介於頻率 f_{m-1} 及 f_m 間或 f_m 及 f_{m+1} 間，則根據公式計算該頻率組成振幅的放大倍率，並且，從公式可看出中心頻率的放大倍率為最大。

$$B_m[k]=\begin{cases} 0 & \text{for } k < f_{m-1} \text{ and } k > f_{m+1} \\ \dfrac{k-f_{m-1}}{f_m-f_{m-1}} & \text{for } f_{m-1} \le k \le f_m \\ \dfrac{f_{m+1}-k}{f_{m+1}-f_m} & \text{for } f_m \le k \le f_{m+1} \end{cases} \quad (1)$$

- ShuffleNet 卷積神經網路

由於 MFSC 後的頻譜資料為一維矩陣，不利於神經網路模型的卷積及池化。因此，本系統計畫將每 64 幀頻譜水平堆疊，形成方形頻譜圖，有助神經網路運作。再將這些頻譜圖作為資料集，分為訓練集和測試集，避免測試資料集遭到汙染，影響測試結果的可信度，將資料放入卷積神經網路訓練及測試。

由於並非所有神經網路架構都是用於辨識音訊，我們實作 ShuffleNet 卷積神經網路，其適用於較短音訊資料的辨識，並且其參數量相較於其他卷積神經網路更少，為輕量化模型。此模型大致可分為四階段，首先為一層全域卷積層，下一階段為兩次連續的深度可分離卷積，最後為一層全連階層兼輸出層，本系統具體之 ShuffleNet 模型架構如下表(一)所示：

Layer		Stride	Output Size	Output Channel
Input		-	64×64	1
3×3 Global Conv		2	32×32	16
MaxPool		2	16×16	16
Stage1	Block1	2	8×8	32
	Block2	1	8×8	32
Stage2	Block1	2	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
Global AvgPool		-	1×1	64
Fully Connected		-	2	1

表(一) 本系統之 ShuffleNet 模型架構

- 神經網路訓練資料集分布

訓練集與測試集資料總數如下表所示，我們將原先蒐集的音訊資料集進行 Data augmentation，提高模型辨識準確率。

音訊	車禍事件	非車禍事件	總數
訓練集	11443	11449	38875
驗證集	3814	3816	7630
測試集	3816	3817	7633

表(二) 資料集分布表

B. 邊緣智慧裝置設計

- 聲音訊號接收

本系統採 FPGA 板內建音效卡外接麥克風與訊號放大器接收音訊，並且藉由暫存區調整資料傳輸頻率來達成聲音訊號之即時傳輸。本系統所使用之 FPGA(PYNQ-Z2)自帶音效卡為 ADAU 1761，內建 3.5mm 音效接口與 ADC，可外接麥克風處理類比訊號。在錄製音效設備方面，我們考量了輕便性、指向性與頻率響應等因素，我們選擇 audio-technica AT9901 麥克風作為收音設備。聲音訊號接收後會被寫入於暫存區，供 MFSC 端進行讀取。

- 聲音預處理(MFSC)

MFSC 是根據人耳聽覺特性設計的非線性聲音預處理流程，其核心為梅爾過濾器。我們由(i)加窗、(ii)快速傅立葉轉換與複數取平方、(iii)梅爾濾波及取對數值構成。由於硬體設計須考慮記憶體儲存容量限制及需達到「即時」音訊處理，表示音訊處理速度必須快於音訊傳入速度，因此我們須簡化 MFSC 的流程以及梅爾濾波器組的參數精度，同時維持卷積神經網路的辨識準確度。

i. 加窗

此系統使用漢明窗，為最常見的窗函數之一，硬體部分我們將軟體端產生的漢明窗數值轉為 16 進位形式，透過自訂 ROM IP 的方式存入 ROM 中，在聲音預處理階段，硬體會直接從 ROM 提取漢明窗的參數並計算，提高運算效率。

ii. 快速傅立葉轉換與複數取平方

快速傅立葉轉換將時域音訊資料轉為頻譜以利卷積神經網路模型的音訊辨識。在硬體設計部分，快速傅立葉轉換在 Vivado 中有現有的 IP 可直接引用，因此，我們直接對 IP 做細部更改，改變輸入輸出小數位數及選擇硬體量最小的設計。接著，我們將 IP 輸出之複數的實部及虛部各自平方後相加，得到各頻率組成振幅的平方，加強頻譜的特徵，並削弱非特徵部分。

iii. 梅爾濾波及取對數值

梅爾濾波器是由多個不同中心頻率的脈波重疊而成，有助於去除特定頻率區域內的資料，從而獲得更具代表性的頻譜特徵。在設計梅爾過濾器時，為避免消耗過多硬體於梅爾濾波器組的初始化上，我們直接將軟體端產生的梅爾濾波器組先轉為 16 進位形式，再透過自訂 ROM IP 存入 ROM 中，提高硬體的計算效率。接著，在聲音預處理階段，硬體會直接呼叫 ROM 提取濾波器組的參數並計算。

最後，為了要讓頻譜特徵更為明顯，我們對頻譜取其對數值，使特徵部分加強。此部分透過查表達成。

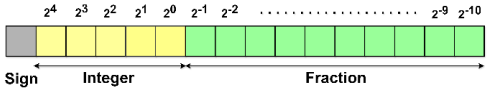
- ShuffleNet 神經網路與狀態機

我們的系統採用小型的 ShuffleNet 神經網路，包含了全域卷積層、兩個深度可分離卷積層以及一個全連接層。透過事先在軟體端架設並訓練好神經網路，我們提取其參數載入到 FPGA 中。相較於傳統的卷積神經網路，ShuffleNet 在維持相同精確度的情況下擁有更少的計算量以及參數量，但相應代價為更加複雜的計算步驟，因此需要設計對應的計算單元，並使用狀態機來調控每一步驟。

i. 16 位元定點數量化

傳統神經網路通常使用 32 位元單精度浮點數運算，這代表參數需要的儲存空間很大，對於需要大量加法與乘法的神經網路來說，浮點數的緩慢計算速度也是不容忽視的缺點。因此我們需要在不減少過多精度的情況下，改變格式或是削減位數來提升運算效率及減少需要的儲存空間。我們採用 16 位元定點數運算，分為 1 位符號位(Sign)、5 位整數(Integer)及 10 位小數(Fraction)，如圖(十)。

ii. ShuffleNet 神經網路



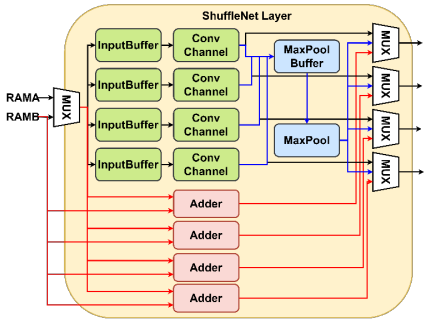
圖(十) 16 位元定點數分配

ShuffleNet 需要的計算步驟有很多種類，大致可以分為以下幾種，如表(三)：

步驟類型	說明	種類
DepthWise Conv1	步長 1 深度卷積	DW
DepthWise Conv2	步長 2 深度卷積	DW
PointWise Conv	逐點卷積	PW
MaxPool3*3	3*3 最大池化	-
AvgPool2*2	2*2 平均池化	PW
AvgPool4*4	4*4 平均池化	PW
Add	兩通道相加	ADD
Fully Connected	全連接層	PW

表(三) ShuffleNet 計算步驟與種類表

由於計算步驟種類繁多，為每一種步驟類型都各自設計一套電路顯然不利於輕量化。為此我們重新分析了各種計算步驟的性質，將逐點卷積、平均池化以及全連接層這些前後兩次計算不會出現輸入資料重疊的類型分為 PW；會出現資料重疊的深度卷積分為 DW；兩通道相加的部分因為每次只需一筆資料，因此單獨拉出來分為 ADD；而最大池化較為不同，由於初始的全域卷積輸出資料非常龐大，且與最大池化一起佔用了將近四分之一的時間，因此我們對其優化，做成流水線來減少時間與空間使用。以下是整體計算層的結構：

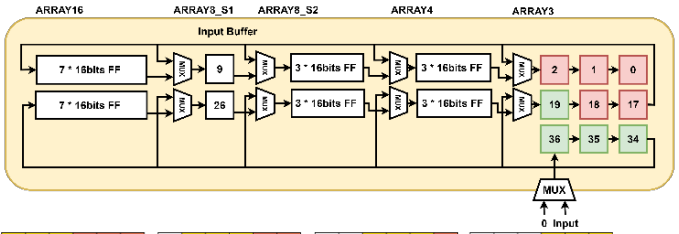


圖(十一) 計算層 (ShuffleNet Layer)結構圖

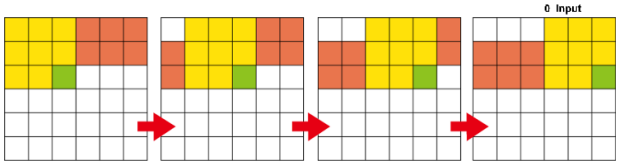
計算層將由四個平行的計算通道組成，這有助於提高運算速度。考慮到電路複用的情況，PW、DW 本質上都是相乘相加的卷積操作，其資料路徑為上圖中的綠色區塊，從輸入端進來依序是輸入緩衝、卷積通道；接在藍色部分則是最大池化的緩衝與計算單元；ADD 則是走紅色路徑使用加法器進行相加。

考量到 RAM 輸出位寬不足及需要處理 DW 這類輸入資料重疊的情況，我們使用移位暫存器做為緩衝。由於存在多種不同大小的特徵圖，有些還包含 Padding，因此我們也用 tapped delay line 的方式，容許緩衝以多條線路移位。

圖(十三)為 4*4 卷積(含 Padding)時，前四個計算周期使用 RAM 位置，黃色為緩衝當前輸出，綠色為緩衝當前輸入，橘色為暫存數據。

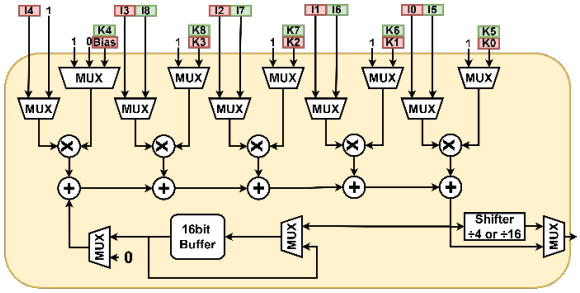


圖(十二) 輸入緩衝(InputBuffer)結構圖



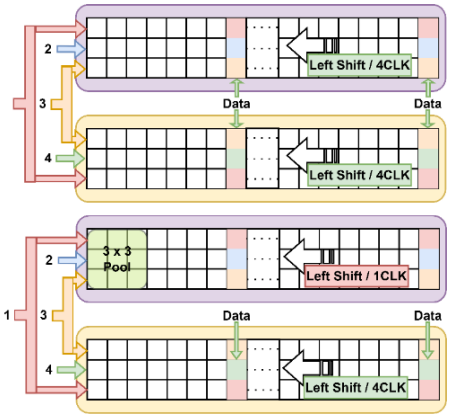
圖(十三) 輸入緩衝(InputBuffer)數據流圖

卷積通道由五組乘法-加法器、移位器、多工器與激活函數 ReLU6 組成，並包含 16bits 暫存器存放暫時資料。此設計容許電路在 2 個時脈訊號週期下計算 3*3 的卷積以及額外一位 Bias。透過輸入端的多工器轉換輸入類型，我們得以用卷積通道計算平均池化。同時，輸出端的移位器是用於處理平均池化的除法。為了減少資源使用量，特別使用 2*2 與 4*4 兩種平均池化而不是更常見的 3*3，這允許我們利用簡單的右移 2 位或 4 位來實現原本需要除法器的除以 4 或 16。透過狀態機調控多工器的輸出，我們得以將不同類型的操作都在卷積通道實現。



圖(十四) 卷積通道(ConvChannel)結構圖

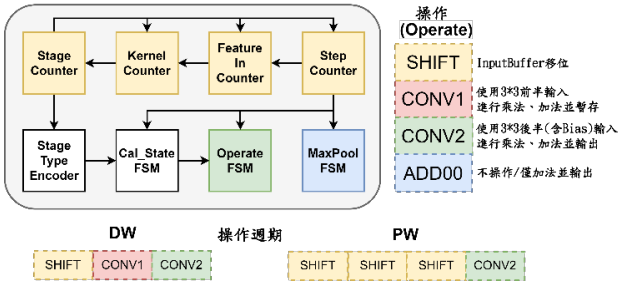
若要將全域卷積與最大池化做流水線，要考慮兩者存取不能衝突，因此設計成 Ping-Pong 形式並拆成四個狀態，狀態 1、3 會同時寫入兩組緩衝，圖(十五)上方表達的是狀態 3(橘色箭頭)；狀態 2、4 則是寫入一組、池化另一組，而下方表達狀態 4(綠色箭頭)，此時上半緩衝進行池化，而下半維持寫入緩衝。



圖(十五) 最大池化緩衝(MaxPool Buffer)結構圖

iii. ShuffleNet 狀態機

狀態機由 Stage、卷積核、特徵圖座標、Step 等四個計數器主導(黃色部分)，前一計數器達目標值後就會清零並驅動下一級計數器，如此遍歷所有步驟。綠色部分是操作狀態機，由各計數器數值可推得當下操作(Operate)種類，透過分析所有的計算都能拆為四種操作(圖十六右半)，能簡化控制訊號實作，例如圖下半 DW 與 PW 兩種操作週期。藍色部分則是控制最大池化流水線的狀態機。

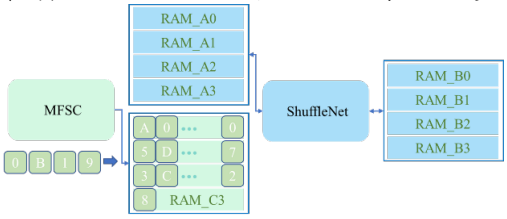


圖(十六) ShuffleNet 狀態機設計圖

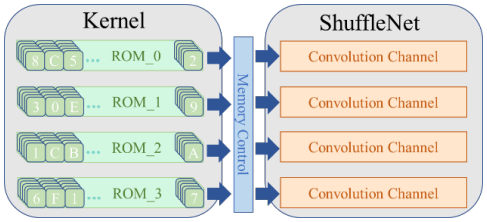
● 記憶體控制

為達成邊緣運算所需的輕便性，記憶體端輕量化分為兩部分，從模型設計角度上，ShuffleNet 使用 Pointwise、Depthwise 和 shuffle 等運算已大幅減少參數需求量；接著記憶體存取上我們也採用空間重複利用盡可能使用最少儲存空間。

記憶體控制端介於整體系統的聲音預處理 MFSC 模組與 ShuffleNet 運算模組之間。因此，硬體設計分為以下兩個課題：其一、即時聲音頻譜數據的暫存；此部分涉及 Ping-Pong RAM 設計，需動態切換與聲音預處理模組 MFSC 及 ShuffleNet 運算所連接之記憶體區塊，如圖(十七)。其二、ShuffleNet 運算所需之 kernel 參數存取；此部分將 ShuffleNet 訓練出的參數以 coe 檔格式匯入 FPGA 唯讀記憶體 ROM 中，並依序取出參數與頻譜數據運算，如圖(十八)。



圖(十七) 記憶體控制 RAM 概念圖



圖(十八) 記憶體控制 ROM 概念圖

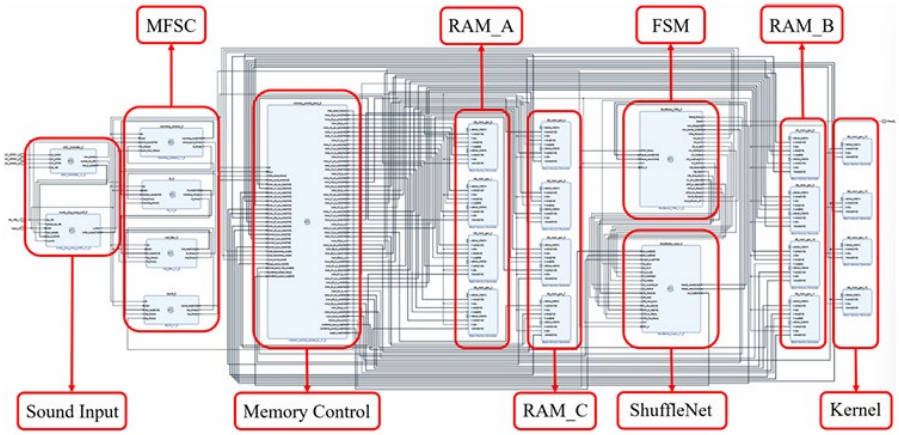
本系統有 4 通道 ShuffleNet 運算單元，因此記憶體單元 BRAM 也是以四個獨立 BRAM 合成一組。此部分使用 12 個獨立的 BRAM IP，位寬皆為 16bits。而在 kernel 儲存端亦是 4 個 ROM IP 與 4 通道進行對應，且因為 convolution unit 設計是一個 clock 執行 5 個運算，故 kernel 輸出位寬是 5 個參數(80bits)。

● 硬體實作結果

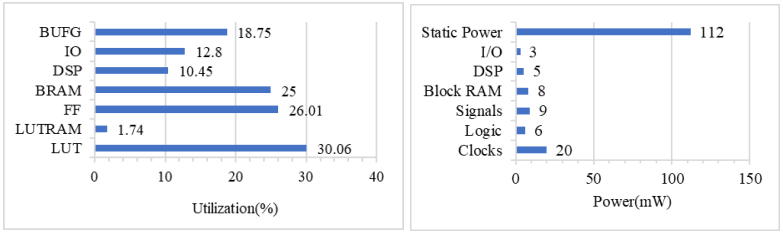
硬體端系統整體系統架構如圖(十九)。實作上，本研究採用 PYNQ-Z2 FPGA 開發板，以下表(四)為系統各部分硬體使用量，在滿足硬體資源輕量化的同時，也盡量追求低功耗的邊緣運算特性，而圖(二十)為總系統資源使用占比與整體系統的功率消耗。

	LUT	Flip-Flop	RAMB18	RAMB36	DSP
Sound Input ~MFSC	10935	23171	8	15	3
Memory Control ~ShuffleNet	5059	4504	8	12	20

表(四) 系統各部分硬體使用表



圖(十九) 硬體系統整體架構圖



圖(二十) 整體系統資源比例與功耗分佈圖

六、結論

本系統相對其他車禍偵測系統的不同之處在於我們採用音訊辨識而非影像辨識，從不同的角度切入此問題，並克服影像辨識的視覺死角問題。我們開發了一套道路聲音辨識的邊緣智慧系統，此系統結合軟硬體以達成高輕量化與高準確度的車禍聲音識別。開發成果可概述為以下 3 項：

1. 資料集建立：收集並整理訓練和驗證模型的聲音數據，並進行資料集平衡與增強，建立一個全面的車禍聲音資料庫。
2. ShuffleNet 神經網路模型訓練：對預處理聲音數據應用 ShuffleNet 進行模型訓練；在維持辨識高準確度的同時盡可能降低硬體模型的參數量與運算複雜度。
3. 智慧邊緣裝置設計和軟硬體整合：ShuffleNet 網絡架構透過平行運算、流水線以及硬體複用等技術，搭配狀態機與記憶體控制模組，實現輕量化的車禍聲音辨識系統。接著將軟體端的聲音預處理 MFSC 模組與 ShuffleNet 神經網路實作在 FPGA 上，不僅提高整體系統的運算速度和效能，也達成實時車禍聲音辨識的效果。

本團隊所提出之技術不僅有潛力彌補傳統攝影機的視覺盲點，也可顯著加快對緊急情況的反應速度，展現了監視器系統功能強化的潛力。同時能達到硬體資源輕量化的目標，實現一個既實用又可行的方案。我們希望能透過本系統降低交通事故通報額外產生的耗時，為緊急救援工作贏得關鍵的時間。

七、參考文獻

[1] 蕭依娜, "針對非特定語者語音辨識使用不同前處理技術之比較," 碩士論文, 電控工程研究所, 國立交通大學, 新竹, 2003, pp. 75.
[2] Kaur, S., "Mouse Movement using Speech and Non-Speech Characteristics of Human Voice," International Journal of Engineering and Advanced Technology, Jan. 2012.
[3] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," in IEEE Access, vol. 10, pp. 122136-122158, 2022
[4] Howard, A.G., et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in arXiv preprint arXiv:1704.04861, Apr. 2017.
[5] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6848-6856
[6] P. Peng et al., "Design of an Efficient CNN-Based Cough Detection System on Lightweight FPGA," in IEEE Transactions on Biomedical Circuits and Systems, vol. 17, no. 1, pp. 116-128, Feb. 2023
[7] S. R. Shahamiri, "Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 852-861, 2021
[8] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, London, UK, 2007, pp. 21-26
[9] M. Cristani, M. Bicego and V. Murino, "Audio-Visual Event Recognition in Surveillance Video Sequences," in IEEE Transactions on Multimedia, vol. 9, no. 2, pp. 257-267, Feb. 2007
[10] F. Iqbal, A. Abbasi, A. R. Javed, G. Srivastava, Z. Jalil and T. R. Gadekallu, "Identification and Categorization of Unusual Internet of Vehicles Events in Noisy Audio," 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), Florence, Italy, 2023, pp. 1-6
[11] Z. Wang, W. Zha, J. Chai, Y. Liu and Z. Xiao, "Lightweight Implementation of FPGA-Based Environmental Sound Recognition System," 2021 International Conference on UK-China Emerging Technologies (UCET), Chengdu, China, 2021, pp. 59-66