The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

結構性重參數化 VGG 架構之輕量化聲音事件偵測模型
Lightweight Sound Event Detection Model with RepVGG Architecture

劉家全 **Chia-Chuan Liu,** 黃頌仁 **Sung-Jen Huang,** 陳嘉平 **Chia-Ping Chen**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
{m103040063, m093040011}@student.nsysu.edu.tw,
cpchen@cse.nsysu.edu.tw
呂仲理 **Chung-Li Lu,** 詹博丞 **Bo-Cheng Chan,** 鄭羽涵 **Yu-Han Cheng,**
莊向峰 **Hsiang-Feng Chuang,** 陳威妤 **Wei-Yu Chen**
中華電信研究院
Chunghwa Telecom Laboratories
{chungli, cbc, henacheng, gotop, weiweichen}@cht.com.tw

摘要

本文中提出以模型輕量化爲目標的聲音事件偵測 RepVGGRNN 模型。其於卷積層使用 RepVGG 卷積塊，透過殘差連接的網路結構使模型達到良好的效能，並於模型訓練完畢後透過結構重參數化使得卷積參數得以縮減。此外，其於訓練階段合併使用知識蒸餾及均值教師模型之訓練方法進一步提昇輕量化模型之預測準確度。RepVGGRNN 在 DCASE 2022 Task4 驗證集中，PSDS(Polyphonic sound event detection score)-scenario 1, 2 分別以 40.8%, 67.7% 優於官方 baseline 系統所達到的 34.4%, 57.2%，並在模型參數量上，RepVGGRNN 使用的參數量約爲 49.6 萬，僅 baseline 系統之 44.6%。

Abstract

In this paper, we proposed RepVGGRNN, which is a light weight sound event detection model. We use RepVGG convolution blocks in the convolution part to improve performance, and re-parameterize the RepVGG blocks after the model is trained to reduce the parameters of the convolution layers. To further improve the accuracy of the model, we incorporated both the mean teacher method and knowledge distillation to train the lightweight model. The proposed system achieves PSDS (Polyphonic sound event detection score)-scenario 1, 2 of 40.8% and 67.7% outperforms the baseline system of 34.4% and 57.2% on the DCASE 2022 Task4 validation dataset. The quantity of the parameters in the proposed system is about 49.6K, only 44.6% of the baseline system.

關鍵字：聲音事件偵測、輕量化模型、知識蒸餾

Keywords: Sound event detection, Light weight model, Knowledge distillation

1 緒論

聲音事件偵測主要是利用機器來辨識聲音訊號中是否存在特定事件，而機器除了辨識音訊中的事件類別外，亦需標註事件發生的起始時間與終止時間，隨著 DCASE challenge Task4 競賽的舉行，此研究項目亦成爲了熱門的音訊處理研究主題之一，許多企業如三星 (Chen et al., 2022), LG(Kim and Yang, 2022) 亦一同參與競賽。在應用上，隨著移動端裝置及物聯網裝置的興起，結合聲音事件偵測的應用如 smart home, smart speaker 也隨之提出，爲日常生活帶來諸多便利性，但受限於硬體上的限制，使得需要高度運算資源的高複雜網路模型不利於佈署在這些裝置上，像是中國語音技術團隊 (Zheng et al., 2021) 在卷積層採用多分支卷積注意力機制，使得模型推論時占用較大的記憶體空間與計算量，而日本名古屋大學團隊 (Miyazaki et al., 2020) 使用 CNN 與 Transformer 結構，使得模型推論一筆音檔時需要較高的運算成本，這些網路模型雖然在事件偵測上有著高度的準確度，但其高

度運算需求之特性亦可能影響使用者的體驗。在本篇論文中，我們以模型輕量化爲目標提出 RepVGGRNN 模型，此模型在架構上採用近年來 DCASE challenge Task4 參賽隊伍主流採用的 CRNN 結構，並在 CNN 的部分參考 RepVGG (Ding et al., 2021) 卷積塊，於訓練時使用多分支殘差連接 (He et al., 2016) 協助訓練卷積層參數，並在訓練完畢後透過結構重參數化將 RepVGG 精簡爲 VGG (Simonyan and Zisserman, 2014) 使得模型整體推論時僅使用單一分支 3×3 卷積層進行運算，相較於 MobileNet (Howard et al., 2017) 透過深度可分離式卷積 (Depthwise separable convolution) 減少參數量，RepVGG 則是應用結構重參數化將原先複雜的多分支卷積簡化爲單一分支卷積達到模型參數與運算時間上的縮減。除了模型結構上的精進，我們於訓練階段中合併使用知識蒸餾 (Hinton et al., 2015) 與均值教師模型 (Tarvainen and Valpola, 2017) 之訓練方法來改善輕量化模型不易於訓練的問題，使整體模型兼具輕量及高準確性等特色。本文中其餘的章節將如下編排：章節二：研究方法，描述了模型架構與訓練方式；章節三：實驗設置，描述網路參數之設置以及評估指標；章節四：實驗結果，對比 RepVGGRNN 與 baseline 模型、預訓練模型之間的差異；章節五：結論，總結了我們所提出系統的特色。

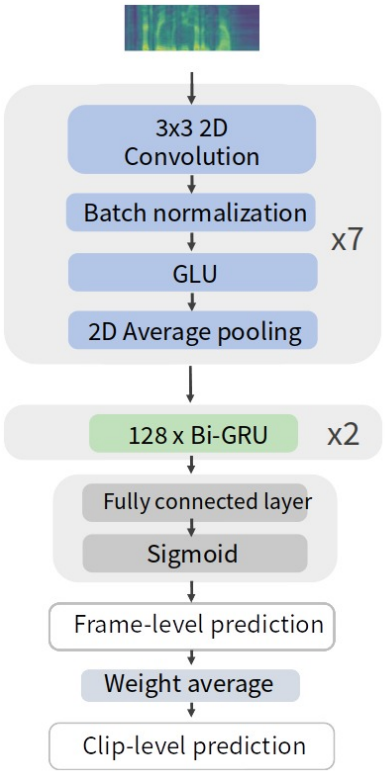

圖 1. Baseline 系統：整體爲 CRNN 模型，輸入特徵會先經 7 層卷積網路進行特徵擷取，並透過 2 層雙向門控循環單元與全連階層產生強預測結果與弱預測結果。

2 研究方法

本章節描述 baseline 與 RepVGGRNN 模型之間架構上的差異，後者於卷積層中使用 RepVGG 卷積塊使得模型在訓練階段透過不同分支卷積得以學習多尺度特徵擷取，除了模型本身的架構外，我們於訓練階段合併使用知識蒸餾及均值教師模型提高資料本身的使用度，進一步提升模型本身的效能，並且使用 mixup (Zhang et al., 2017) 資料增強來減緩模型過擬合等現象。

2.1 Baseline 模型

本文中以 DCASE Task4 官方所提供的 CRNN 模型 (Turpault et al., 2019) 作爲 baseline 系統，模型整體如圖 1 所示使用 7 層卷積網路層連接 2 層循環網路層，在卷積層的部分，每層使用 3×3 大小的卷積核，各層濾波器的數量分別爲 16, 32, 64, 128, 128, 128, 128 個，並以門控線性單元 (Gated Linear Unit,GLU) 作爲激勵函數, 以及在各層卷積層中使用批標準化 (Batch normalization) 與 dropout ，最後，每層卷積層運算完畢後會再進行平均池化，平均池化視窗的大小依序爲 2×2, 2×2, 1×2, 1×2, 1×2 與 1×2。在循環網路層的部分爲 2 層雙向門控循環單元 (Bidirectional Gated Recurrent Unit) ，各層具 128 個神經元，最後透過一層全連接層與 S 型函數 (Sigmoid function) 產生該筆輸入音檔的強標註預測 (Strong prediction)，此預測結果包含了預測的事件類別與該事件所發生的時間界線，接著將強標註預測依各類時間維度使用注意力池化 (Attention pooling) 取權重平均來產生該筆音檔的弱標註預測 (Weak prediction)，此預測結果相比於強標註預測，僅包含事件類別而無時間界線上的註記。

2.2 RepVGGRNN 模型

因應模型輕量化的目標，我們透過修改 baseline 系統的架構來達到減少參數量的效果，在卷積層中，層數由原先的 7 層縮減至 5 層，並在結構上參考 RepVGG 卷積塊，內部結構如圖 2 所示，單一卷積層於訓練時包含了 3 個分支，分別爲 3×3 卷積, 1×1 卷積與恆等層，並且各卷積皆連接批標準化層，與 residual network 殘差連接的網路設計相似，各分支的輸出會進行加總並在運用線性整流函數 (Rectified Linear Unit, ReLU) 後作爲 RepVGG 卷

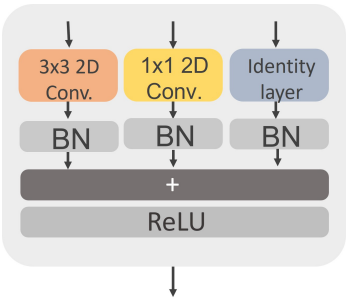

圖 2. RepVGG 卷積塊：訓練階段時共有三個分支卷積，依卷積核的大小分爲 3×3、1×1 與恆等層。

積塊的輸出。多分支卷積相比於單一分支在模型推論時通常佔用了較大的記憶體空間，因此 RepVGG 卷積塊在模型訓練完畢後再透過結構重參數化 (Structure re-parameterization) 將 1×1 卷積分支, 恆等分支合併至 3×3 卷積分支來縮小空間佔用率及整體模型的參數量，結構重參數化共分爲三個步驟 (1) 將各分支卷積所連接之批標準化層合併至卷積層中，以 RepVGG 中的 3×3 卷積分支爲例，令其輸入維度、輸出維度分別爲 c_{in}, c_{out} ，卷積核爲 $W \in R^{c_{out} \times c_{in} \times 3 \times 3}$，輸入特徵及輸出特徵分別爲 $F \in R^{N \times c_{in} \times T \times F}$ 與 $\hat{F} \in R^{N \times c_{out} \times T' \times F'}$ ，當中的 N 爲批次中的資料筆數且 T, F 爲時間維度與頻率維度大小，則該分支卷積輸出 $\hat{F}$ 各 *channel* 維度特徵 $\hat{F}_{:,i,:,:}$ $\forall i, 1 \leq i \leq c_{out}$ 爲第 i 個卷積核 $W_{i,:,:,:}$ 與輸入特徵 F 經卷積運算與標準化後的結果，如以下等式：

$$\hat{F}_{:,i,:,:} = \gamma_i \frac{(W_{i,:,:,:} * F) - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i \qquad (1)$$

，其中 γ_i , μ_i , σ_i^2 , β_i 表示第 i 個 *channel* 其批標準化層之參數，而 $*$ 爲卷積運算子，爲了將批標準化層之參數合併至卷積層中，令 $W'_{i,:,:,:}$ 與 β'_i 分別爲第 i 個卷積核在合併批標準化後的參數與偏差值爲以下等式：

$$W'_{i,:,:,:} = \frac{\gamma_i}{\sqrt{\sigma_i^2 + \epsilon}} W_{i,:,:,:} \qquad (2)$$

$$\beta'_i = \beta_i - \frac{\gamma_i \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \qquad (3)$$

，則此時可將公式 1 透過 W'_i, β'_i 簡化爲

$$\hat{F}_{:,i,:,:} = W'_{i,:,:,:} * F + \beta'_i \qquad (4)$$

完成批標準化參數的合併 (2) 將合併批標準化層之 1×1 卷積層、恆等層透過補 0 的方式將卷積核擴張爲 3×3 卷積 (3) 將擴張完成的卷積透過卷積運算之可加性將 1×1 卷積, 恆等層之參數合併至 3×3 卷積中以完成結構重參數化，經上述步驟後即可將原先多分支的卷積合併爲單一 3×3 卷積，整體流程可參考圖 3。RepVGG 原作者在卷積層之輸入特徵維度與輸出特徵維度不相同時並無使用恆等層，因此我們稍微了修改其設置，當輸入特徵圖的維度與輸出特徵圖的維度不相同時恆等層會以 3×3 卷積層做取代，使其整體網路保有三分支卷積的結構。RepVGGRNN 各層卷積的濾波器數量爲 16, 32, 64, 128, 128 個，並且在前三層中皆堆疊了兩層的 RepVGG 卷積塊，後兩層中各僅使用一層。循環網路層的部分我們將 baseline 中的 2 層雙向門控循環單元縮減至 1 層，而全連接層的部分與 baseline 系統相同，並以強標籤預測與弱標籤預測來作爲模型最終的輸出。

2.3 合併知識蒸餾與均值教師模型

有鑑於網路模型在縮減架構的同時也犧牲了精準度上的表現，因此我們透過合併使用知識蒸餾與均值教師模型來協助輕量化模型進行訓練，使得輕量化模型除了保有輕巧的特色外，亦能維持住相當水準的性能。

知識蒸餾是模型壓縮經常使用到的訓練方法，利用預訓練好的高複雜度模型 (我們稱爲預訓練教師模型) 來引導低複雜度模型 (我們稱爲學生模型) 進行訓練，使得預訓練教師模型具有的高精準度與泛化能力能夠遷移至學生模型上，而均值教師模型是近年來半監督式學習 (Semi-supervised learning) 中主流使用的訓練方法，與知識蒸餾間的相似之處在於，均值教師模型亦存在老師模型與學生模型的概念，不同的是在均值教師模型當中，兩者是完全相同的網路模型，而在訓練時，透過對於同一筆輸入資料之輸出，學生模型的預測除了需貼近眞實標注資料外，也必須與老師模型之預測結果相近，透過保持學生模型與老師模型間的一致性，也使得未具有標注的資料被充分的利用。在實做中，我們合併使用了知識蒸餾與均值教師模型，使用了三個模型分別爲預訓練教師模型與均值教師結構的 RepVGGRNN 模型，使得模型在訓練初期透過預訓練教師模型提供可靠的預測標籤供學生模型進行參考，而在均值教師結構的模型中透過移動平均更新的方式，使均值教師模型更近一步提升精準度。我們令均值教師中學生模型的預測結果與三者計算差值，分別爲 (1) 眞實標籤 (爲 Supervised loss) (2) 預訓練教師模型的預測結果 (爲 Knowledge distillation loss) (3) 均值教師模型中老師模型的預測結果 (爲

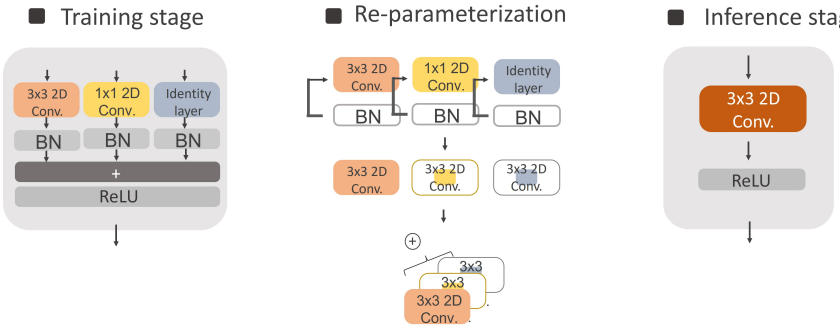

圖 3. RepVGG 重參數化：當模型訓練完畢後 RepVGG 透過結構重參數化將 1×1 卷積、恆等層之參數合併至 3×3 卷積層中，使其變爲單一分支卷積結構。

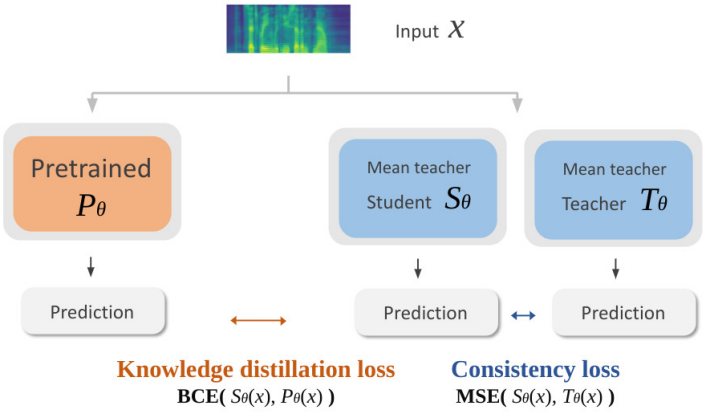

圖 4. 合併知識蒸餾與均值教師模型：學生模型在損失函數的計算上共有三個來源，分別是與 Ground-truth 之間的差值、預訓練模型之輸出間的差值以及均值教師模型中教師模型之間的差值。

Consistency loss) ，其損失函數如下:

$$L_{student} = L_{Supervised} + L_{KD} + L_{Consistency} * w \quad (5)$$

。由於均值教師模型於訓練初期有著較差的準確性，所以其權重 w 會先設爲 0，隨著訓練的進行再逐步調高其權重。在參數的更新中，僅有學生模型會參與反向傳播的更新，待更新完畢後再以指數移動平均 (Exponential moving average) 之方式利用學生模型之參數更新教師模型，如下公式:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \quad (6)$$

，其中，θ'、θ 分別代表教師模型與學生模型之參數，t 代表當前的訓練 step ，而 α 是介於 0 至 1 之間的權重，整體訓練流程可參考圖 4 。

2.4 預訓練教師模型

預訓練教師模型使用的是 VGGSKCCT 模型，該模型於卷積層使用殘差連接使得模型在有較深的卷積層數下也能減緩梯度消失的現象而有較佳的效能，並於卷積層間使用選擇性內核單元 (Selective kernel unit) (Li et al., 2019) ，透過不同大小的卷積核與注意力機制使得模型在不同事件的偵測準確率上能有效的提昇。整體模型之溫度參數 (Temperature parameter) 設置爲 2，並使用不同資料增強方式與 fusion 多個訓練結果來進一步提升模型效能，模型 fusion 的數量爲 3 個。

2.5 資料增強

模型訓練時對原始資料進行輕微的擾動可減緩過擬合的現象，我們參照了 DCASE Task4 所使用的資料增強方式，對同一批中的強標籤註記資料與弱標籤註記資料各自隨機成對進行 mixup，過程如下列公式所示:

$$x' = \lambda x_i + (1-\lambda)x_j \quad (7)$$

$$y' = \lambda y_i + (1-\lambda)y_j \quad (8)$$

，x_i 與 x_j 爲隨機兩筆同一批且具同性質標籤之資料，而 y_i 與 y_j 爲其各自的眞實標籤，經過介於 0 至 1 之間的權重係數 λ 進行線性組合後，所得 x' 與 y' 即爲經過 mixup 所得之資料與標籤。

3 實驗設置

本節將描述實驗所採用的相關設置，包括：模型所使用的訓練集與測試集、特徵前處理方式、整體訓練時中的學習率設置及模型所採用的評估指標等。

3.1 資料集

	資料筆數	類型	原始採樣率
強標籤訓練集	13470	眞實錄製或合成	44.1kHz/16kHz
弱標籤訓練集	1578	眞實錄製	44.1kHz
無標籤訓練集	10000	眞實錄製	44.1kHz
強標籤驗證集	1168	眞實錄製	44.1kHz
公開測試集	699	眞實錄製	44.1kHz

表 1. DESED 訓練集與測試集其資料筆數、類型與原始採樣率

資料集使用 DCASE 2022 Task4 提供的 DESED(Domestic Environment Sound Event Detection dataset) 資料集做爲模型的訓練與評估。每筆音檔的長度爲 10 秒，依標籤註記種類的不同分爲 (1) 強標籤資料: 音檔標籤包含了事件類別並註記事件的起始時間與終止時間 (2) 弱標籤資料: 音檔的標籤僅註記事件的類別 (3) 無標籤資料: 音檔沒有提供任何相關的標籤註記。各類音檔筆數分別爲 13470、1578、10000 筆。

測試集以 DESED 提供的驗證集 (Validation dataset) 與公開測試集 (Public evaluation dataset) 作爲模型的評估，各測試集中的資料筆數分別爲 1168 筆與 692 筆，並且每筆音檔皆具有強標籤的注記，詳細內容如表 1。

3.2 音訊特徵擷取

由於 DESED 資料集中的音檔存在採樣率、聲道不一致與音檔長度存在不一致的情形，因此我們使用 librosa 套件將所有音檔統一爲 16000 Hz、單聲道並透過補 0 之方式將各音檔長度填補至 12 秒，並將波型訊號 (Waveform) 轉換爲梅爾頻譜圖 (Mel-spectrogram) 並取 log 作爲網路模型的輸入，在參數設置上，我們以窗口大小 (Window size) 爲 2048、框擷取步伐 (Hop length) 爲 256 進行短時傅立葉變換，最後經由 128 個梅爾濾波器 (Mel-filter bank) 產生維度大小爲 751(時間維度)、128(頻率維度) 的梅爾頻譜圖。

3.3 參數設置

所有的實驗結果皆使用相同的參數設置，每個網路模型皆訓練 200 個 epoch ，在優化器的部分我們使用 ADAM 演算法，並於前 50 個 epoch 應用 exponential warm-up 策略，初始學習率會以趨近於 0 的極小值隨著訓練步伐的增加而遞增，至第 50 個 epoch 時學習率會遞增至最大值 0.001。

3.4 評估指標

Polyphonic sound event detection score (PSDS)(Bilen et al., 2020) 適用於評估模型於多類別聲音事件預測上的準確性，其在模型預測與眞實標籤之間依序透過 (1) 檢測容差標準 (Detection Tolerance Criterion): 事件預測標籤與眞實標籤間的交集是否超過 DTC 門檻 (2) 眞實標籤交集標準 (Ground Truth intersection Criterion): 眞實標籤是否存在 (通過 DTC 門檻的) 事件預測標籤與其交集超過 GTC 門檻 (3) 交叉觸發容差標準 (Cross-Trigger Tolerance Criterion): 事件預測標籤在時間上的預測正確但類別錯誤，分別計算眞陽性 (True positive): 通過 DTC 與 GTC 的事件、僞陽性 (False positive): 未通過 DTC 的事件、與跨類別觸發事件 (Cross-Trigger): 未通過 DTC 但通過 CTTC 的事件，接著再透過 TP, FP 與 CT 來計算最終的 PSDS。我們參考 DCASE 2022 task 4 的參數設置，使用兩個參數設置分別爲 PSDS-scenario1(簡稱 PSDS-1) 與 PSDS-scenario2(簡稱 PSDS-2) 來作爲指標，分別將三者門檻比率分別設置爲 0.7、0.7、0，該參數對於事件預測在時間區間的精準度上有著較高的要求，而後者則是設爲 0.1、0.1、0.3，著重於事件類別預測的正確性。

除了模型的效能外我們也評估了模型的資源使用量，分別統計了模型的參數量與浮點數運算次數 (Floating Point Operations, FLOPs)，前者分析模型於記憶體空間之佔用量，後者則是評估模型推論時之計算複雜度。使用的套件是 Pytorch 第三方函式庫 THop，此套件可用於統計 Pytorch 模型上的參數量及浮點數運算資訊，我們使用當中的 profile 函式統計模型之參數量與模型推論一筆 10 秒鐘音檔所需的 FLOPs 作爲實驗結果之數據。

4 實驗結果

我們比較了 RepVGGRNN 分別與 baseline、VGGSKCCT 及 DCASE 2022 Task4 競賽第一名模型於效能及資源使用量間的差異。Baseline 系統的實驗結果是由我們使用官方提供的程式碼重新訓練而取得，而 DCASE 2022 Task4 第一名模型是使用官方所公佈的數據結果。此外，若模型以均值教師模型訓練，則 PSDS 分數取學生模型與教師模型各別 (PSDS-1)+(PSDS-2) 較高者爲代表。

Model	Validation dataset		Public evaluation dataset	
	PSDS-1	PSDS-2	PSDS-1	PSDS-2
Baseline(Provided by DCASE Task4)	0.344	0.572	0.385	0.546
依不同訓練方式				
RepVGGRNN(均值教師模型)	0.370	0.620	0.421	0.660
RepVGGRNN(知識蒸餾)	0.388	0.654	0.441	0.687
RepVGGRNN(合併均值教師與知識蒸餾)	**0.408**	**0.677**	**0.447**	**0.688**
預訓練教師模型				
VGGSKCCT	0.426	0.670	0.489	0.712
DCASE 2022 Task 4 第一名模型				
Ebbers UPB task4_4	0.492	0.721	-	-

表 2. 模型效能比較：呈現 baseline、RepVGGRNN、VGGSKCCT 與 DCASE 2022 Task 4 第一名模型於驗證集與公開測試集下的 PSDS 結果。

模型	參數量	浮點運算次數
VGGRNN	$4.974*10^5$	$5.418*10^8$
RepVGGRNN(Training)	$6.283*10^5$	$7.515*10^8$
RepVGGRNN(Inference)	$4.965*10^5$	$5.279*10^8$

表 3. 重參數化之比較：RepVGGRNN 重參數化前、後與 VGGRNN 在參數量與浮點運算次數中的差異，當中的數值以科學記號來表示，並將實數部份之小數點第三位以下之數值進行無條件捨去。

Model	參數量	浮點運算次數
Baseline	$1.112*10^6$	$9.309*10^8$
RepVGGRNN	$4.965*10^5$	$5.279*10^8$
VGGSKCCT	$7.485*10^6$	$1.07*10^{10}$
Ebbers UPB task4 _ 4	$1.34*10^8$	-

表 4. 各類模型資源使用量的比較：呈現各類模型在參數量與浮點數運算次數的差異。

4.1 效能比較

表 2 爲 RepVGGRNN 依訓練方式之不同，分別使用 (1) 均值教師模型 (2) 知識蒸餾 (3) 合併使用均值教師模型與知識蒸餾之 PSDS。此外，Ebbers UPB task4_4 系統數據以 DCASE Task 4 官方公佈之結果作爲呈現，因爲提交系統並沒有上傳公開驗證集上的結果，因此沒有列出該數據。在驗證集與公開驗證集中，RepVGGRNN 以均值教師模型方式訓練下，其 PSDS 要高於 baseline 系統，若使用 VGGSKCCT 透過知識蒸餾方式訓練 RepVGGRNN ，其 PSDS-1、PSDS-2 皆有所成長，由此可知利用預訓練模型提供的高準確度預測相比於均值教師模型，可使學生模型有著較佳的訓練效果，若更進一步使用知識蒸餾並維持 RepVGGRNN 之均值教師模型訓練方式，在預測效能上相比於僅使用均值教師模型，在驗證集中 PSDS-1 由 0.370 提昇至 0.408，而 PSDS-2 亦由 0.620 提昇至 0.677 ，顯示了除了預訓練模型所提供的參考預測外，均值教師模型中的教師模型參數是以學生模型參數透過指數移動平均方式來更新，因此教師模型相比於學生模型既學習到了當前資料特徵之分佈，亦較大程度的保留過往資料所學習到的特徵，使得教師模型較不易受到離群資料的影響而降低在常態資料上的預測，進一步增進模型效能。最後，RepVGGRNN 與 DCASE 2022 Task 4 第一名的模型 Ebbers UPB task4_4 相比，雖然 RepVGGRNN 透過模型架構與訓練方式的改進來提升精準度，但受限於模型本身的規模與訓練資料的使用，在預測的準確度上仍有著較大的落差。

4.2 VGG 與 RepVGG 重參數化之差異

表 3 呈現了 RepVGGRNN 進行結構重參數化前、後與一般 VGGRNN 在資源使用量之差異，當中的 VGGRNN 是將 RepVGGRNN 中的 RepVGG 替換爲 VGG 而得，若卷積層中的 RepVGG 堆疊了兩層即以同樣堆疊兩層卷積之 VGG 替代，若僅有一層 RepVGG 則以單一層卷積層取代，使兩模型在卷積層的深度相同。在參數量上，由於合併了各 RepVGG 層中的 1×1 卷積、恆等層中的 3×3 卷積與批標準化層中的參數至 3×3 卷積後，RepVGGRNN 的參數量由原先的 62.8 萬縮減至 49.6 萬，減少的幅度約爲 20.9%，並且浮點運算次數亦由 7.515 億次降至 5.279 億次，幅度爲 29.8%，顯示模型整體透過合併卷積與批標準化的方式可達到參數量與運算量的縮減。而與一般 VGG 相比，重參數化後的 RepVGGRNN 因批標準化層皆融合進了卷積層，雖然在參數量上減少的幅度不大但在整體運算量有著相對明顯的降低，由 $5.418*10^8$ 降至 $5.279*10^8$，縮減的幅度約爲 2.7%。

4.3 資源使用量比較

表 4 列出四者模型於推論時之資源使用量，當中的數據以科學記號方式表達，同時對實數位小數點第三位以下的數進行無條件捨去，其中，VGGSKCCT 因爲使用 3 個架構相同的模型做 fusion ，因此其參數量與計算量以單一模型之數據的 3 倍作爲實驗結果數據。首先比較參數量，RepVGGRNN 與 baseline 、VGGSKCCT 與 Ebbers UPB task 4_4 相比皆爲其中最少者，總參數量約爲 49.6 萬個，僅使用 baseline 參數量約 111.2 萬之 44.6%，顯示了 RepVGGRNN 透過整體架構的縮減仍可以相對較少的參數量達到接近 baseline 系統的效能，與 VGGSKCCT 系統相比，僅約其 748.5 萬參數量之 6.6%，且是 Ebbers UPB task 4_4 系統 1.34 億參數量之 0.3% 。除了空間上的占用量外，運算量亦爲輕量化模型所需縮減的目標之一，RepVGGRNN 之運算量爲三個模型中最少者，其處理單一筆資料共需約 5.279 億次浮點運算，爲 baseline 9.309 億次運算之 56.7%，且爲 VGGSKCCT 所需 107 億次運算之 8.7%，可見 RepVGGRNN 模型在重參數化與縮減模型層數後，其在資源使用上具有相當的優勢。

5 結論

近年來隨著移動式裝置的普及，結合深度學習的移動端應用亦隨之而發展，除了網路模型本身的效能外，硬體資源使用的情形如記憶體使用量、續航力與運算需求亦是模型部屬所考量的方向，透過我們的實驗結果可見，RepVGGRNN 在驗證集中以 PSDS-1 , PSDS-2 分別爲 0.408% , 0.677% 皆高於 baseline 系統所達到的 0.344%, 0.572% ，且在資源使率中其參數量僅使用約 49.6 萬個，少於 baseline 系統所具有的 111.2 萬個參數，顯示了相比於 baseline 系統，其兼具了高準確性及輕量化的特色。在未來，希望能持續增進系統並在移動端裝置實踐聲音事件偵測之相關應用。

References

Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović. 2020. A framework for the robust evaluation of sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65. IEEE.

Minjun Chen, Tian Wang, Jun Shao, Yiqi Tang, Yangyang Liu, Bo Peng, Jie Chen, and Xi Shao. 2022. Dcase 2022 challenge task4 technical report. Technical report, DCASE2022 Challenge.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Changmin Kim and Siyoung Yang. 2022. Sound event detection system using fixmatch for dcase 2022 challenge task 4. Technical report, DCASE2022 Challenge.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519.

Koichi Miyazaki, Tatsuya Komatsu, Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, and Kazuya Takeda. 2020. Convolution-augmented transformer for semi-supervised sound event detection. Technical report, DCASE2020 Challenge.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah. 2019. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Xu Zheng, Han Chen, and Yan Song. 2021. Zheng ustc team’s submission for dcase2021 task4 — semi-supervised sound event detection. Technical report, DCASE2021 Challenge.