殘差模組結合擠壓與激發注意力機制改進少樣本道路警報偵測模型

Residual Modules Combined with Squeeze-and-Excitation Attention Mechanism for Improving Few-Shot Road Alert Detection Model

郭子豪 **Tzu-Hao Kuo**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
m113040053@student.nsysu.edu.tw
鄭羽涵 **Yu-Han Cheng**
中華電信研究院前瞻科技研究所
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
henacheng@cht.com.tw
陳嘉平 **Chia-Ping Chen,** 陳品錞 **Ping-Chun Chen**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
cpchen@cse.nsysu.edu.tw
pipichun17@gmail.com
呂仲理 **Chung-Li Lu,** 詹博丞 **Bo-Cheng Chan,** 莊向峰 **Hsiang-Feng Chuang,**
陳威妤 **Wei-Yu Chen**
中華電信研究院前瞻科技研究所
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
{chungli, cbc, gotop, weiweichen}@cht.com.tw

摘要

本文提出了一個少樣本分類模型 SE-RCNN。該模型使用殘差模組來提升特徵擷取能力，並利用 GELU 激勵函數改善 ReLU 資訊丟失的問題，同時通過擠壓與激發注意力機制強調關鍵特徵。在 5-way 1-shot 的少樣本情境下，該模型在 ESC-50 資料集上的準確率從 76.2% 提升至 82.1%。接著，我們以此模型爲雛型，在 4-way 的情境下利用自行蒐集的道路警報資料集進行調適。我們的模型在每個類別提供 15 個樣本的情形下各類別的 F1-score 均不低於 0.8。最後，我們以片段級預測的方式實做一個道路警報偵測模型。

Abstract

This paper proposes a few-shot classification model called SE-RCNN. The model uses residual modules to enhance feature extraction capabilities, GELU activation functions to mitigate information loss from ReLU, and Squeeze-and-Excitation attention mechanisms to emphasize key features. In a 5-way 1-shot few-shot learning scenario, the model's accuracy on the ESC-50 dataset improved from 76.2% to 82.1%. Subsequently, we used this model as a prototype and adjusted it using the self-collected road alert dataset in a 4-way scenario. Under the condition of providing 15 samples for each category, our model achieved an F1-score of no less than 0.8 for all categories. Finally, we implemented a road alert detection model using a segment-level prediction approach.

關鍵字：聲音事件偵測、少樣本學習、注意力機制

Keywords: sound event detection ,few-shot learning, attention mechanism

1 緒論

聲音事件偵測 (Sound Event Detection, SED) 是一種自動從聲音訊號中識別和分類各種聲音事件的技術。這些聲音事件涵蓋環境音效（如雷聲、狗吠）、人爲活動（如講話、拍手）以及機械聲（如救護車警笛、汽車引擎聲）等。SED 技術在智慧城市、監控系統、醫療監測和人機交互等領域具有廣泛的應用。例如，在工業環境中，異常聲音檢測系統能夠及早識別需要維護的潛在故障設備 (Dohi et al., 2022)。在家庭環境中，偵測系統可以即時偵測破窗聲

或火災警報等事件，從而實現即時警報和緊急應變。然而，在實際應用中，資料蒐集經常面臨挑戰。由於聲音事件的種類繁多，蒐集大量資料集涉及龐大的人力成本，某些特殊聲音事件甚至難以進行大規模蒐集。因此，我們嘗試結合少樣本學習 (Few-Shot Learning) 與聲音事件偵測，以解決這一問題。

少樣本學習是機器學習的一個特殊分支，旨在使模型在面對新任務時，能夠依賴少量資料進行訓練，並展現出良好的泛化能力。這一概念在資料蒐集困難且類型繁多的聲音事件偵測領域具有巨大的應用潛力。然而，目前的少樣本學習研究主要集中在圖像領域。例如，Vinyals et al. (2016) 與 Song et al. (2020) 分別使用圖像資料集進行訓練並應用於圖像識別任務。此外，大多數研究主要著眼於演算法性能的提升或公開資料集準確率的比較，卻忽略了少樣本學習在實際應用中的便利性和優勢。

因此，本論文提出了一個道路警報偵測模型，該系統能夠識別道路上需禮讓的緊急車輛聲音，如救護車、警車和消防車。在當今社會，隨著車輛隔音技術的進步，佩戴耳機的行人和騎士在道路上的比例逐漸上升，這使得該系統能夠有效提醒用路人對緊急事件車輛進行禮讓，從而提升整體的交通安全環境。我們的模型基於 MetaAudio (Heggan et al., 2022) 作爲核心分類模型，並通過引入殘差網路結構 (He et al., 2016)、高斯誤差線性單元 (Hendrycks and Gimpel, 2016) 以及擠壓與激發注意力機制 (Hu et al., 2018)，來增強基礎模型的特徵擷取能力。隨後，我們利用自行蒐集的道路警報資料集對模型進行微調。最終，我們的模型可以進行片段級別的預測。

本文的後續章節安排如下：第二章：研究方法，將說明模型架構與訓練方式；第三章：實驗設置，將介紹實驗的相關設置與資料集；第四章：實驗結果，將比較不同模型之間的差異並說明相關實驗數據；第五章：結論。

2 研究方法

在這個章節中，我們將詳細說明本次實驗所使用的各種方法，包括少樣本學習的方法、所使用的骨幹網路架構細節，以及使用的注意力機制。實驗的核心基礎主要基於 MetaAudio (Heggan et al., 2022)，並對其骨幹網路架構進行了改良。

2.1 模型架構

此部分將說明基礎模型與我們改良的模型，並進一步說明高斯誤差線性單元以及擠壓與激發注意力機制的相關內容。

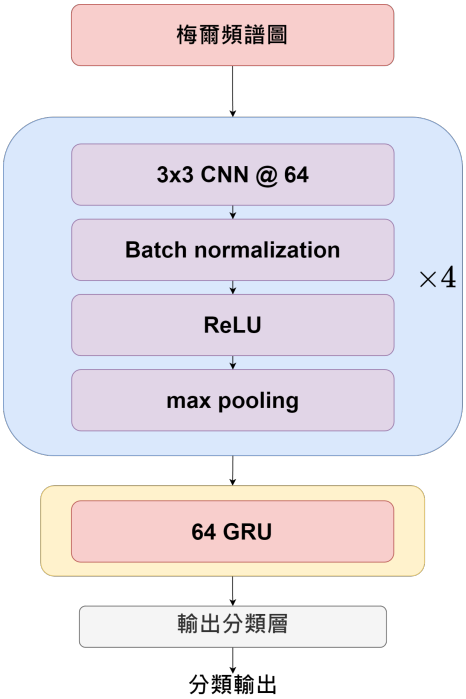

圖 1: CRNN 模型: 輸入頻譜圖首先經過四個堆疊的卷積模組，之後通過單層的 GRU，最後由輸出分類層輸出結果。

2.1.1 基礎模型

MetaAudio 使用兩種不同的基礎模型，分別是卷積神經網路 (Convolutional Neural Network, CNN) 和卷積遞歸神經網路 (Convolution-Recurrent Neural Network, CRNN)。CRNN 在 CNN 的基礎上進一步擴展，加入了一層遞歸神經網路（Recurrent Neural Network, RNN）。CNN 架構由四個卷積模組組成，每個模組均使用 3×3 的卷積層，隨後是批正規化層（Batch Normalization）、線性整流函數（Rectified Linear Unit, ReLU (Agarap, 2018) 以及最大池化層（Max Pooling）。其中，最大池化層的取值範圍皆爲 2×2。在 CRNN 中，所使用的 RNN 架構基於閘門遞歸單元（Gated Recurrent Unit, GRU） (Chung et al., 2014)。GRU 通過其門控機制，有效地捕捉序列資料中的時間依賴性，這使得 CRNN 能夠在處理時序音訊資料時更好地保留和利用上下文資訊。最後，無論是 CNN 還是 CRNN，兩者都連接至一個分類層。該分類層由一個具有 30% 丟棄率的丟棄層（Dropout）、一層批正規化層和一層全連接層（Fully Connected Layer）組成。CRNN 整體架構圖如圖 1。

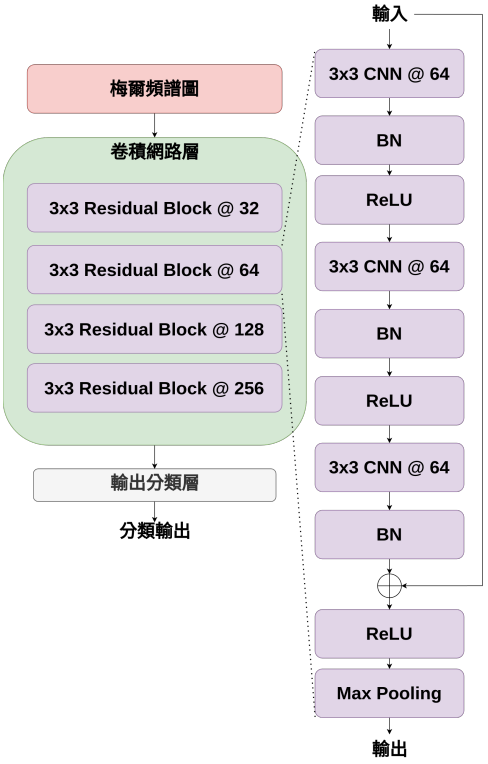

圖 2: RCNN 模型: 輸入頻譜圖首先經過四個堆疊的殘差模組，四個殘差模組的通道數依序爲 32、64、128 與 256，最後由輸出分類層輸出結果。

2.1.2 殘差卷積神經網路模型

我們選擇以 CNN 模型作爲改進的基礎，主要的改動想法是增加模型的整體深度，以便使模型能夠學習到更爲複雜的特徵表示。透過增加深度，模型可以捕捉到更細緻的資料特徵，從而提高其表現。爲了實現這一目標，我們引入了捷徑連接（shortcut connection）(He et al., 2016)，並結合多個卷積層構建殘差模組（residual module）。這種殘差結構使得資訊在模型中可以有效地流動，並減輕了隨著層數增加而可能出現的梯度消失問題。具體而言，捷徑連接允許原始輸入直接傳遞到更深的層，從而促進了更快的訓練和更好的收斂效果。我們將修改後的模型命爲 RCNN。RCNN 由四層殘差模組與一層全連接層組成。殘差模組中包含兩個分支，一個分支是恆等映射函數，另一個分支是多層結構。該結構依序由一個卷積層、批正規化層、ReLU 激勵函數、卷積層、批正規化層、ReLU 激勵函數、卷積層與批正規化層組成。這兩個分支的輸出相加後，再經過 ReLU 激勵函數和最大池化層進行處理。每個殘差模組的通道數分別爲 32、64、128 和 256。CRNN 整體架構圖如圖 2。

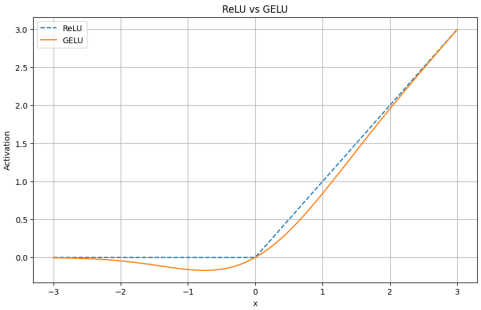

圖 3: 高斯誤差線性單元與線性整流函數圖：圖中橘色實線爲高斯誤差線性單元，藍色虛線爲線性整流函數。

2.1.3 高斯誤差線性單元

高斯誤差線性單元（Gaussian Error Linear Units, GELU (Hendrycks and Gimpel, 2016)）是一種激勵函數。與常見的 ReLU 激勵函數相比，GELU 在輸入值接近 0 的區域具有平滑的轉換，這有助於避免梯度的突然變化，從而提高模型訓練過程的穩定性。公式 1 爲 GELU 的數學表達式：

$$\mathrm{GELU}\,(x) = x \cdot \Phi(x) \tag{1}$$

$\Phi(x)$ 代表高斯分佈的累積分佈函數，可由公式 2 表示：

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{2}$$

erf 表示爲誤差函數 (error function)。爲了便於計算，GELU 通常會採用近似公式來簡化計算。公式 3 是常用的 GELU 近似公式：

$$\begin{aligned} &\mathrm{GELU}\,(x) \approx 0.5x(1+T) \\ &T = \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right] \end{aligned} \tag{3}$$

圖 3 爲 GELU 激勵函數與 ReLU 激勵函數的比較圖。

2.1.4 擠壓與激發注意力機制

此外，我們還嘗試引入擠壓與激發注意力機制（Squeeze-and-Excitation attention, SE attention）(Hu et al., 2018)，SE attention 的結構如圖 4。注意力機制能夠幫助模型自動聚焦於重要的特徵，進一步增強了模型的表示能力。SE attention 它通過擠壓特徵圖的通道資訊，生成一個權重向量，然後根據這些權重自適應地調整各通道的特徵強度。具體而言，

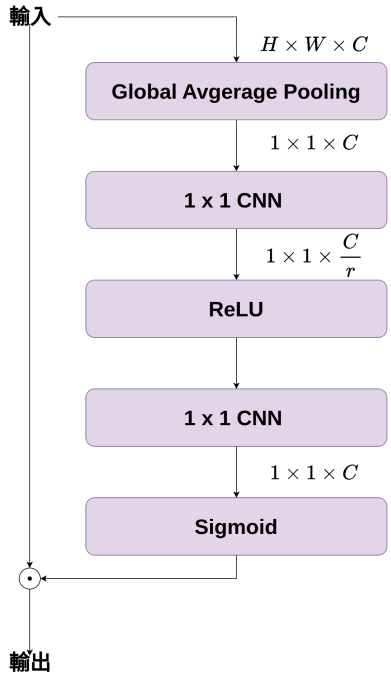

圖 4: SE attention 結構圖：C 代表特徵通道數，r 代表擠壓率，$\odot$ 代表元素對應相乘。輸入先經由全局池化層將每個通道的空間維度進行平均，產生一個 $1 \times 1 \times$ C 的向量，之後經由兩個卷積操作擠壓與激發向量。最後經過乙狀函數 (Sigmoid) 產生通道權重向量。

SE attention 包括兩個步驟：首先對特徵圖進行全局平均池化以獲取通道向量，然後通過全連接層或 1×1 卷積生成權重向量。這種方式能夠強調重要特徵，抑制不重要特徵，提升模型表現。

2.2 少樣本學習

少樣本學習的訓練過程與一般深度學習有所不同，其目標是能夠快速適應「新」任務，因此在少樣本學習中，每一筆訓練資料的單位是一個「任務」。任務會被劃分爲支持集（support set）與查詢集（query set），其中支持集與查詢集彼此互斥。支持集的功能類似於訓練集，而查詢集則接近於測試集的概念。任務通常會設定一個情境，說明需要分類的類別數量以及每個類別提供的帶標籤資料數量，並以 N-way K-shot 來表示。N-way 說明任務中有 N 個不同的目標類別，K-shot 則表示每個目標類別有 K 個樣本。例如，在一個 5-way 5-shot 的任務中，支持集包含 5 個類別，每個類別有 5 筆樣本，而查詢集通常包含 5 個類別，每個類別有 1 個樣本。在少樣本學習的訓練過程中，模型會接觸到很多個任務，每個任務都是從訓練集中抽樣。通過這樣的方式，模型不僅能夠學習到如何快速適應新任務，還能夠捕捉到不同任務之間的關聯性。這種任務間的相互作用有助於模型更好地理解特徵的共享性，從而提升其在新任務上的性能。

2.2.1 元曲率

元曲率 (Meta-Curvature) (Park and Oliva, 2019) 是我們主要使用的少樣本學習方法，旨在訓練時學習通用的模型參數與更新的曲率，以提高模型對於新任務的是適應能力。爲了實現這一目標，訓練過程會將「任務」是爲訓練資料的單位，模型反覆在不同的任務上進行訓練，以便訓練後的模型能夠使用少量的樣本或訓練迭代適應新任務。

首先，元曲率定義了三個元曲率矩陣，$M_o \in \mathbb{R}^{C_{out} \times C_{out}}$、$M_i \in \mathbb{R}^{C_{in} \times C_{in}}$ 和 $M_f \in \mathbb{R}^{d \times d}$，$C_{out}$、$C_{in}$ 和 d 分別表示輸出通道數、輸入通道數和濾波器大小。在卷積層中，d 表示高度 $\times$ 寬度，而在全連接層中，d 表示 1。元曲率的函數定義如公式 4 所述：

$$MC(\mathrm{G}) = G \times M_f \times M_i \times M_o \tag{4}$$

G 代表損失函數的梯度，且 M_f、M_i 與 M_o 在初始化時爲單位矩陣。

假設 f_ϕ 代表一個參數爲 ϕ 的模型。從訓練資料集隨機抽樣 k 個任務，k 是一個可自訂的超參數，以 $\mathcal{T}_i$ 表示這 k 個任務中的第 i 個任務。f_ϕ 以任務的支持集作爲訓練集訓練參數，更新後的參數爲 θ_i。如公式 5 所述：

$$\theta_i = \phi - \alpha MC(\nabla \mathcal{L}_{T_i}(\phi)) \tag{5}$$

θ_i 是經過 $\mathcal{T}_i$ 支持集訓練後的參數，而 α 爲學習率，並將此參數更新稱作內循環 (inner loop)。之後，每個任務使用各自的查詢集再次計算損失，根據這些損失更新模型參數 ϕ 與矩陣參數 M_f、M_i 與 M_o。如公式 6 所述：

$$\begin{aligned}
\phi &\leftarrow \mathrm{ADAM}(\phi, \beta, \nabla_\phi \sum_{T_i} \mathcal{L}_{T_i}(\theta_i)) \\
M_o &\leftarrow \mathrm{ADAM}(M_o, \beta, \nabla_{M_o} \sum_{T_i} \mathcal{L}_{T_i}(\theta_i)) \\
M_i &\leftarrow \mathrm{ADAM}(M_i, \beta, \nabla_{M_i} \sum_{T_i} \mathcal{L}_{T_i}(\theta_i)) \\
M_f &\leftarrow \mathrm{ADAM}(M_f, \beta, \nabla_{M_f} \sum_{T_i} \mathcal{L}_{T_i}(\theta_i))
\end{aligned} \tag{6}$$

β 代表學習率，ADAM 爲我們使用的優化器 (optimizer)，並將此參數更新稱作外循環 (outer loop)。以上爲元曲率的一次迭代訓練，流程圖如圖 5。

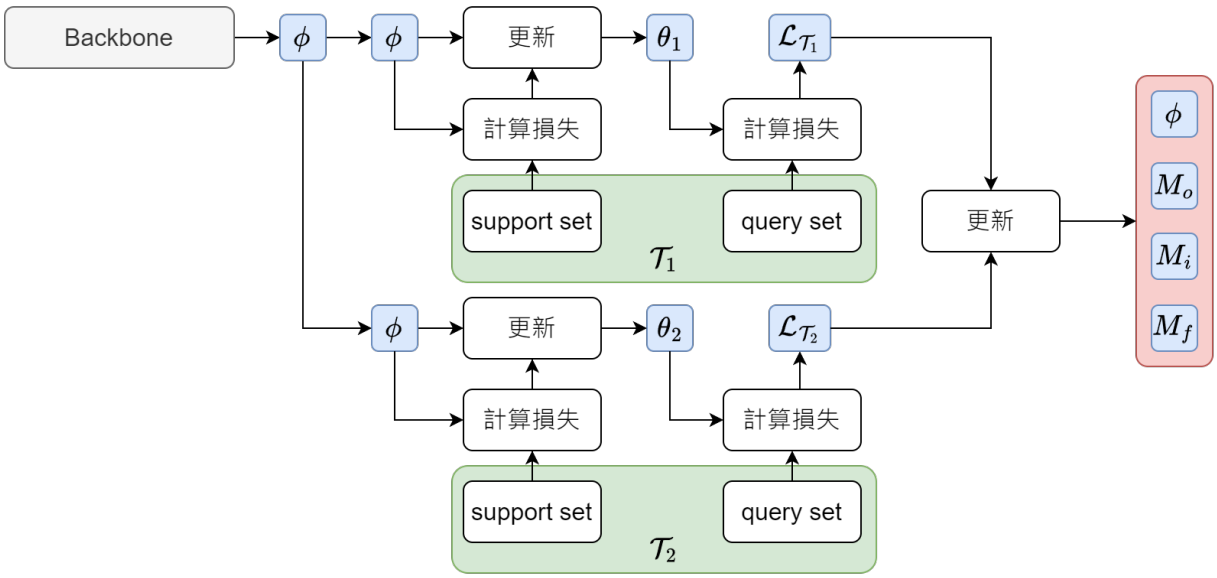

圖 5: Meta-Curvature 訓練的示意圖：此圖爲 Meta-Curvature 訓練的舉例說明。假設任務數量設定爲 2，模型的參數 ϕ 將被複製爲兩份。這兩份參數分別使用 $\mathcal{T}_1$ 與 $\mathcal{T}_2$ 進行參數更新。最後，通過在查詢集 (query set) 的損失總合更新參數 ϕ、M_o、M_i 與 M_f。

3 實驗設置

本節將描述本論文實驗中所採用的相關設置，包括模型使用的訓練集與測試集、聲音特徵的前處理方法、訓練過程中的學習率設置及各項超參數的設定。此外，還將說明用於評估模型表現的效能指標。

3.1 資料集

我們選擇使用 ESC-50 資料集 (Piczak, 2015) 作爲訓練與測試雛型模型的資料集。ESC-50 是一個環境聲音分類資料集，包含 2,000 個標註音檔，每個音檔長度爲 5 秒，採樣率爲 44.1kHz，檔案格式爲.wav。這些音檔來自 50 個不同的類別，每個類別各包含 40 筆資料。由於少樣本訓練的特殊性，我們將資料集依據類別進行分割：35 個類別用作訓練集，5 個類別作爲驗證集，其餘 10 個類別則用於測試。此分割方式參考了 Chou et al. (2019)。我們自行蒐集了一個道路警報資料集，該資料集包含四種類別的聲音，分別爲救護車聲、警車聲、消防車聲以及道路雜音。道路雜音類別擁有 300 筆音檔，而其他每種類別則各有 100 筆音檔。所有音檔均爲 5 秒長，且採樣率爲 16kHz。此資料集將用作訓練與測試道路警報模型的資料集。

3.2 聲音特徵前處理方法

首先，我們將所有音檔重新採樣至 16kHz，接著將音訊從波形訊號（waveform）轉換爲梅爾頻譜圖（mel-spectrogram），最後取對數作爲模型的輸入。在頻譜圖的設定中，使用 128 個梅爾濾波器 (Melfilter bank)，傅立葉轉換的窗長 (Window size) 設爲 1,024，跳躍長度 (Hop length) 爲 512。

3.3 參數設定

本論文的所有實驗數據均使用相同的訓練設定。每個模型訓練 200 個 epoch，並採用 Adam 優化器 (optimizer) (Kingma and Ba, 2014)，內循環學習率 α 設爲 0.4，外循環學習率 β 設爲 0.001，整個過程中學習率均保持固定，不使用自適應技術。在驗證階段，我們從驗證集中隨機抽取 200 個任務進行準確率測試；在測試階段，則從測試集中隨機抽取 5,000 個任務進行評估。任務由支持集與查詢集組成，模型先以支持集調適模型，然後在查詢集上進行測試以計算準確率。最終呈現之準確率爲所有任務的平均準確率。

3.4 評估準則

我們主要以準確率 (Accuracy) 作爲雛型模型的評估準則，準確率是正確預測的樣本數與總樣本數的比率。由於道路警報系統模型的測試集存在類別不平衡的現象，因此在道路警報系統模型的部份則使用 F1-Score 作爲評估準則。F1-Score 是精確率 (Precision) 和召回率 (Recall) 的調和平均數，公式 7 8 9 別分呈現精確率、召回率與 F1-Score 的計算方式：

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

模型架構	Accuracy(%)
CNN	71.2
CRNN	76.2
RCNN	80.1

表 1: RCNN 與基礎模型準確率比較：展示 CNN、CRNN 和 RCNN 模型在 ESC-50 測試集上進行 5,000 個任務抽樣後的平均準確率結果。

模型架構	Accuracy(%)
RCNN	80.1
RCNN(GELU)	80.8
SE-RCNN(GELU)	82.1

表 2: RCNN 改動準確率比較：呈現 RCNN 更換激勵函數與引入注意力機制後，於 ESC-50 測試集上進行 5,000 個任務抽樣後的平均準確率結果。

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

TP 代表正確預測爲正類的樣本數，FP 代表錯誤預測爲正類的樣本數，FN 代表錯誤預測爲負類的樣本數。

4 實驗結果

4.1 模型比較

表 1顯示我們所改進的 RCNN 模型相較於基礎模型展示了顯著的進步，這一進步主要體現在殘差模組的設計上。殘差模組通過堆疊多層卷積結構，使模型能夠有效地學習和捕捉更深層次的特徵關係。具體來說，這些堆疊的卷積層增強了模型的表達能力，使其能夠更好地捕捉數據中的複雜模式和細微變化。同時，RCNN 模型還對通道數進行了擴展，進一步增強了模型處理不同特徵的能力。此外，殘差模組引入了捷徑連接，不僅能減少梯度消失問題，還能提升深層神經網路的訓練穩定性和效率。捷徑連接通過直接將輸入傳遞到後面的層，形成一條捷徑路徑，使得梯度可以更順暢地反向傳播，從而有效避免深層網路常見的梯度消失問題。

少樣本情境	Accuracy(%)	記憶體佔用
1-shot	82.1	4.9GB
3-shot	94.4	10.3GB
5-shot	96.3	*9.4GB

表 3: shot 數與準確率關係：呈現在不同少樣本情境下訓練 SE-RCNN 模型後，於 ESC-50 測試集上進行 5,000 個任務抽樣的平均準確率結果與記憶體佔用情形。標記 * 表示因超出硬體極限而進行的參數調整後的記憶體佔用情形。

4.2 RCNN 改動比較

表 2 顯示更換激勵函數爲 GELU 與引入注意力機制對於 RCNN 準確率提昇的有效性。替換激勵函數後準確率提 0.7 %，這是因爲相比於 ReLU 在輸入小於零時會導致神經元關閉，可能會丟失一些有用的資訊，而 GELU 通過引入高斯分佈，使激勵函數能夠對輸入進行更加平滑的處理，從而保留更多的細節。引入 SE attention 後，模型的準確率又提升了 1.3 %，因爲 SE attention 的引入使模型能夠自適應地重新調整特徵的權重，從而增強對重要特徵的關注，抑制不重要的特徵。這種動態調整特徵權重的方法，有助於模型更加精確地捕捉關鍵資訊，從而提升整體性能。

4.3 K-shot 情境分析

此實驗主要目的是爲了解 K-shot 對於預測的影響趨勢，我們固定 5-way 進行訓練，分析不同 shot 數對準確率與硬體需求的影響。表 3 呈現在 5-way 不同 shot 數情境下使用 SE-RCNN 模型的訓練結果。隨著 shot 數量的增加，模型能夠獲取更豐富的資料和多樣的特徵，這使得模型更容易擷取各類別的共同特徵，從而提高模型的泛化能力。因此，表格中 5-shot 的訓練結果顯示出最佳的準確率。然而，隨著 shot 數量的增加，訓練對硬體的需求也相應上升。以本次訓練爲例，5-way 5-shot 訓練已達到我們訓練設備 NVIDIA 1080 Ti 顯示卡的記憶體極限，因此我們有對 5-way 5-shot 的訓練參數進行調整。後續道路警報系統實驗皆以 5-shot 情境下的 SE-RCNN 作爲雛型模型。

4.4 調適模型

少樣本訓練到模型應用的過程分爲兩個階段。第一階段是利用一定量的樣本來訓練一個雛型模型。例如，本文採用 ESC-50 資料集，通過元曲率方法訓練出一個雛型模型。第二階段則是對雛型模型進行調適，使其適應特定的目

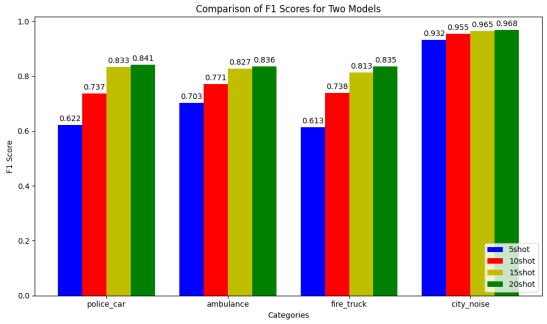

圖 6: F1-Score 比較圖：此表展示 SE-RCNN 模型在不同 shot 數下，對道路警報資料進行支持集抽樣以調適模型，並將剩餘樣本作爲測試集，重複 100 次後得到的各類別平均 F1-Score 結果。

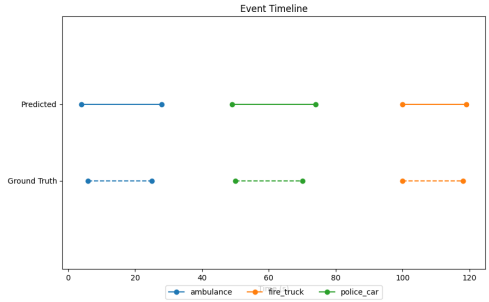

圖 7: 音檔預測示意圖：此圖不同顏色的線段代表不同事件，虛線爲事件時間的眞實標籤，實線爲事件時間之預測。

標任務。在本文中，我們將雛型模型調適至道路警報偵測任務。在第二階段，我們以 4-way 的形式對模型進行調適。道路警報模型是基於一個 4-way 的 SE-RCNN 雛型模型進行調適而成。我們使用道路警報資料集作爲調適階段的訓練集。模型的測試方式是通過從訓練集隨機抽樣生成支持集來調適模型，然後使用訓練集的剩餘樣本作爲測試集來評估模型的效能。這一過程重複進行 100 次，以確保結果的穩定性和可靠性，最終呈現各類別的平均 F1-score。圖 6 呈現在調適階段以不同 shot 數訓練的結果。由圖可以觀察到當 shot 越高效能越好，其中以 5-shot 到 10-shot 的提高效果最爲明顯。此外，道路警報模型的 5-shot 結果與 ESC-50 5-shot 結果之間存在一定的差距，我們認爲這主要與資料集的特性有關。爲了更貼近實際應用情況，我們的資料集並未對數據進行額外處理，這使得聲音更容易受到設備和外部環境的影響。相較於 ESC-50 資料集中的聲音，後者的音質較爲乾淨和清晰。

4.5 道路警報偵測系統

我們以章節 4.4 的模型作爲基礎實做道路警報偵測。在偵測系統中，對於每一筆輸入的音檔，系統內部設定了一個固定長度的偵測窗，該偵測窗的長度爲 5 秒。偵測窗以每 1 秒的步伐位移，經過整份音檔完成偵測。輸出的標籤會經過後處理，系統將連續相同類別的窗框區間定義爲事件發生區間，只有區間大於 7 秒的事件會被留下，過小的區間事件將被刪除以穩定輸出。圖 7 爲系統對於音檔預測的示意圖。從圖中可以看出，該系統已經能夠大致辨識出聲音事件，但由於採用片段級預測作爲輸出，若需進行精確定位，其精度仍然不足。因此，這將是我們接下來研究中需要努力改進的部分。

5 結論

本文以 MetaAudio 作爲基礎，嘗試引入殘差模組、GELU 激勵函數與 SE attention，以進一步提升基礎模型的表現。這些改進措施旨在加強模型對聲音特徵的捕捉和辨識能力。經過這些改進，使用元曲率少樣本學習方法，模型的分類準確率從 76.2% 提升至 82.1%。此外，我們以上述模型爲雛型，利用自行蒐集的道路警報資料集調適出一個道路警報偵測模型，並以此建構了一個道路警報偵測系統。然而，該系統仍存在偵測單位過大的問題，這將是我們後續努力的重點目標。

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. 2019. Learning to match transient sound events using attentional similarity for few-shot sound recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 26–30. IEEE.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. 2022. Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. *arXiv preprint arXiv:2206.05876*.

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.

Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi. 2022. Metaaudio: A few-shot audio classification benchmark. In *International Conference on Artificial Neural Networks*, pages 219–230. Springer.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Eunbyung Park and Junier B Oliva. 2019. Meta-curvature. *Advances in neural information processing systems*, 32.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Gege Song, Zhulin Tao, Xianglin Huang, Gang Cao, Wei Liu, and Lifang Yang. 2020. Hybrid attention-based prototypical network for unfamiliar restaurant food image few-shot recognition. *Ieee Access*, 8:14893–14900.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.