運用不同音訊長度於遷移式學習以提升電鋸聲音識別能力之研究
A Study on Using Different Audio Lengths in Transfer Learning for Improving Chainsaw Sound Recognition

張家瑋 Jia-Wei Chang
jiaweichang.gary@gmail.com

胡忠雲 Zhong-Yun Hu
e871223eeee@gmail.com

國立臺中科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taichung University of Science and Technology

摘要

在山林中，由於聲音的多元複雜及環境中諸多的雜訊，電鋸聲音的識別是富有挑戰性的任務。本研究認為以不同的聲音長度對於模型的訓練結果可能有所差異，故以簡易的 LeNet 模型結合了平均池化層設計出能夠接受任意長度音訊的識別模型。本研究主要分析不同聲音長度對於模型訓練之影響以及短至長與長至短音訊的遷移學習結果。本實驗皆以 ESC-10 資料集來訓練模型並以自行蒐集的電鋸聲資料集驗證模型的準確度。實驗結果表明(1)以 1 秒、3 秒、5 秒資料集分別訓練的三個模型，在 1 秒、3 秒與 5 秒的電鋸聲驗證集中，各達到 74%~78%、74%~77%與 79%~83%的準確度。(2)以 1 秒→3 秒→5 秒的 ESC-10 資料遷移學習的模型於 1 秒、3 秒與 5 秒電鋸聲驗證集中分別達到 85.28%、88.67%與 91.8%準確度，均較原訓練方法有所明顯提升。(3)在遷移式學習中，相較於長至短秒數的遷移訓練，以短至長秒數的遷移訓練得到了較佳的結果；尤其在 5 秒的電鋸聲驗證集中相差了 14%的準確度。

Abstract

Chainsaw sound recognition is a challenging task because of the complexity of sound and the excessive noises in mountain environments. This study aims to discuss the influence of different sound lengths on the accuracy of model training. Therefore, this study used LeNet, a simple model with few parameters, and adopted the design of average pooling to enable the proposed models to receive audio of any length. In performance comparison, we mainly compared the influence of different audio lengths and further tested the transfer learning from short-to-long and long-to-short audio. In experiments, we used the ESC-10 dataset for training models and validated their performance via the self-collected chainsaw-audio dataset. The experimental results show that (a) the models trained with different audio lengths (1s, 3s, and 5s) have accuracy from 74%~78%, 74%~77%, and 79%~83% on the self-collected dataset. (b) The generalization of the previous models is significantly improved by transfer learning, the models achieved 85.28%, 88.67%, and 91.8% of accuracy. (c) In transfer learning, the model learned from short-to-long audios can achieve better results than that learned from long-to-short audios, especially being differed 14% of accuracy on 5s chainsaw-audios.

關鍵字：聲音辨識、環境聲音分類、電鋸聲音識別、遷移學習

Keywords: Voice Recognition, Environmental Sound Classification, Chainsaw Sound Recognition, Transfer Learning

1 緒論

近年來對於環境保護的意識逐漸在社會大眾中受到重視，其中針對森林的相關議題除了天災的野火、土石流，就是針對人為因素的防範，在人為的事件中對於森林破壞度最大的就是盜伐，透過監視器或是人工巡邏對於保護一個森林來說成本以及保護力度都不足

夠，但如果透過聲音監控的方式，可以將布置防護網的成本降低，也可以更加即時的反應盜伐事件的發生。這個任務是屬於環境聲音分類(Environmental Sound Classification, ESC)的範疇，在進行環境聲音分類的任務之前需要將聲音經過預處理，預處理的方法有很多，也能取得許多不同的特徵以供模型使用。過零率(Zhang and Kuo, 2001)、小波特徵(Valero and Alías, 2012)、梅爾倒頻譜係數(MFCC)(Uzkent et al., 2012)。目前機器學習與深度學習已經被廣泛的應用於環境聲音分類任務上。支持向量機(Support Vector Machine，SVM) (Chu et al., 2009; Piczak, 2015b)、隨機森林分類器(Random Forest Classifier，RF) (Piczak, 2015b)、高斯混合模型(Gausasian Mixture Model，GMM) (Piczak, 2015b; Dhanalakshmi et al., 2011)都是經典的機械學習方法。但是機器學習一開始的訓練很耗費時間和成本。如果沒有充足的資料，難以訓練出可用的模型。近年來，深度學習技術已被很好的應用於從聲音信號中提取高辨識度特徵以執行環境聲音分類。提取有用特徵以及對於細微聲音仍然保持良好的泛化能力使深度學習成為了環境聲音分類的首選方式。環境聲音分類與語音辨識任務的不同之處在於，環境聲音分類所要識別的聲音通常都是零散的，同一類的聲音所轉換出的頻譜圖可能表現出相當大的落差，聲音模式可以是連續的、不規則的、瞬間的、而且大部分會包含許多吵雜或是無聲的帧，我們輸入的聲音長度也有可能不同。並且如果需要將模型應用於森林之中的小型監控設備，對於模型的大小、複雜度以及聲音的長度都有較大的限制，因此本篇論文想要研究在一個簡單的模型中，所以本篇嘗試透過改變訓練時的音訊秒數以進行遷移式學習(Zhuang et al., 2020; Liao et al., 2021; Hung and Chang., 2021)來研究模型針對訓練集外的電鋸聲音判斷的敏感度。本研究其餘章節的組織如下：第二節說明本研究會使用之環境聲音模型原始架構相關工作，第三節介紹本研究所使用的資料集以及解釋本研究實行之方法論，第四節為本研究模型訓練結果比較以及模型對於資料集音訊外的聲音判斷能力結果，第五章對實驗結果進行相關討論，第六章總結本研究結果。

2 相關研究

本章節介紹使用基於深度學習的模型進行環經聲音分類的相關工作。

2.1 將頻譜圖應用於 CNN

由於將音訊轉換成頻譜圖後能得到二維的特徵，所以頻譜圖一直是聲音深度學習模型喜歡使用的預處理方式。在 CNN 問世後(Piczak, 2015a)首次提出將頻譜圖特徵作為輸入並執行 ESC 的 2D-CNN，根據研究結果表示與 SVM、RF、GMM 等機器學習模型相比 PiczakCNN 顯著的提高了辨識的準確度，受到 PiczakCNN 的啟發，愈來愈多的人將頻譜圖輸入不同的 CNN 模型，也有人結合了預訓練網路都得到了極佳的效果(如 GoogleNet(Szegedy et al., 2015)和 AlexNet(Krizhevsky et al., 2012))。

2.2 LeNet-5

本篇研究所使用之模型參考自 LeNet-5(LeCun et al., 2015)並進行一些修改以符合訓練任務需求，在 90 年代，由於 SVM 等算法的發展，深度學習的發展受到了很大的阻礙。但 LeCun 等人(LeCun et al., 2015)堅持不懈，依然在該領域苦苦研究。1998 年，LeCun 提出了 LeNet-5 網絡用來解決手寫識別的問題。LeNet-5 被譽為是卷積神經網絡的「Hello Word」，足以見到這篇論文的重要性。該模型共有 7 層，共有 3 個卷基層、2 個平均池化層以及 2 個全連接層。

2.3 二元自適應均值池化層

要進行本篇想研究的方案之前，需要先想辦法使模型能輸入不同維度大小的聲音資訊，於是在原有模型卷積層後展平層之前加入了 AdaptiveAvgPool2d 此方法的使用概念類似於全域性池化層(Global Average Pooling, GAP)(Lin et al., 2013)，使模型能輸入不同維度的聲音資料，正常的平均池化需要自己計算窗口以及步伐，但 AdaptiveAvgPool2d 能夠只輸入想要輸出的資料維度大小，它會自動的計算窗口以及步伐使得輸出格式符合模型要求。

2.4 遷移學習

在某些領域中標籤的標記昂貴，導致訓練資料的不足，容易使訓練出來的模型發生過擬合的狀況，也就是對於訓練資料外的資料泛化能力不足，導致模型沒有實務價值，遷移學習中有兩個常用的方法，特徵萃取和微調，特徵萃取是指先以預先訓練好的模型作為資料特徵提取的部分，為目標任務提取有用的特徵，微調技術是將原有任務訓練好的模型以及參數應用於目標訓練任務，使目標訓練任務能有較佳的初始梯度位置進行訓練，能達到較快收斂以及增加準確度的功效，遷移式學習在過去的研究取得不錯的成果，因此本篇研究提出的模型訓練方式便基於微調技術，研究是否能夠在不改變模型複雜度的情況透過改變輸入音訊長度進行遷移訓練以此提升準確度。

3 方法論

3.1 資料集

- ESC-10 資料集(Piczak et al., 2015b):

ESC-10 資料集為 ESC-50 資料集的子集，內包含 400 個室內外環境錄音的標記集合，適用於環境聲音分類的基準測試方法，該數據集的音訊都是由 5 秒長的紀錄組成，採樣率為 44100Hz，被平均分類為 10 個類別，其中一類為電鋸聲，每個類別都有 40 條音訊，此數據集內的標籤已預先安排了 5-fold 以進行交叉驗證，確保同一原始源文件的片段包含在同一個 fold 中。

- 電鋸聲音集：

本研究自行蒐集的包含電鋸聲音的聲音片段，沒有包含任何來自 ESC-10、ESC-50 的音訊資料，數據集的音訊都是由 5 秒長的紀錄組成，採樣率為 44100Hz，特別一提的是這些片段不全是乾淨的電鋸聲，以模擬在現實中需要判斷時會有雜音的情況。

3.2 透過不同秒數進行訓練模型之架構

本章共有四個小節，第一小節說明資料預處理，第二小節介紹模型實做細節以及實驗的參數設定，第三小節說明實驗設計，第四小節說明實驗環境與超參數設定。

3.2.1 資料預處理

從給定的聲學訊號中提取了頻譜圖特徵。採樣率為 44100Hz，帧移設置為 512，窗口長度為 2048，濾波器個數為 128，最高頻率為 22050，最低頻率為 20。在此研究中使用了 Python 中的 Libroas 庫(McFee et al., 2015)來提取頻譜訊號，由於本研究想要透過不同秒數來訓練模型以及進行遷移式訓練，想要提取短中長三種不同長度以做出區別所以選擇了 1 秒、3 秒、5 秒。轉換出來的特徵大小分別為(128,87,1)、(128,259,1)、(128,431,1)。由於聲音都是 5 秒片段，所以提取 1 秒及 3 秒聲音時資料分別會增加 5 倍及 3 倍，圖 1 表示聲音片段提取的方式。本實驗設計了兩種模型訓練方式分別為：(1)正常的 ESC-10 標籤，標籤由 0 至 9 共 10 個標籤進行分類，以及(2)將電鋸聲以外的聲音標籤都設為 0，電鋸聲標籤設為 1，進行二元分類。

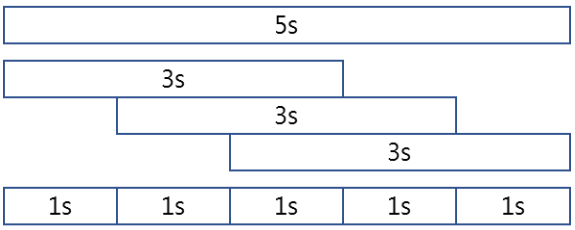

圖 1: 不同聲音長度之提取示意圖

3.2.2 模型實做細節以及實驗的參數設定

圖 2 展示了本篇研究所使用之模型架構。模型相關參數如下。

- A1:輸入大小為(1, 128, W)，W 為 1 秒、三秒以及 5 秒轉換成頻譜圖後的寬度。
- A2:為一個 2D-CNN 卷積層，輸入通道數為 1，輸出通道數為 16，kernel_size 為 5，stride 為 1，padding 為 0。
- A3:為一個最大池化層，kernel_size 為 2，stride 為 2，padding 為 0。
- A4:為一個 2D-CNN 卷積層，輸入通道數為 16，輸出通道數為 32，kernel_size 為 5，stride 為 1，padding 為 0。
- A5:為一個最大池化層，kernel_size 為 2，stride 為 2，padding 為 0。

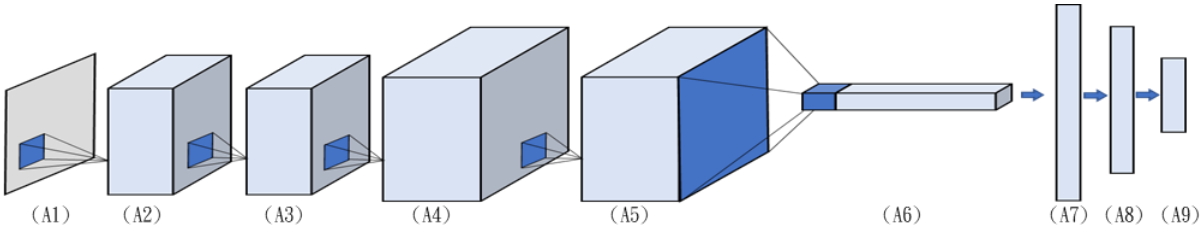

圖 2: LeNet 結合 Global Average Pooling 之模型架構圖

- A6:為一個二元自適應平均池化層，依照通道方向進行自適應平均池化，輸出為一為陣列長度 32。
- A7:為全連接層，節點數為 120。
- A8:為全連接層，節點數為 84。
- A9:為全連接層，節點數為 10 或者 2，依照實驗項目而定。

3.2.3 實驗設計

本研究以不同長度 (1 秒、3 秒、5 秒) 的 ESC-10 資料集來訓練模型，並將訓練好的三個模型分別使用本研究自行蒐集電鋸聲資料集之 1 秒、3 秒與 5 秒的音檔來驗證電鋸聲之識別能力。本研究著重於(1)觀察不同長度聲音的訓練對於準確度的影響；(2)依照秒數由短到長(1 秒→3 秒→5 秒)與由長到短(5 秒→3 秒→1 秒)的方式進行遷移學習訓練，並與先前的模型進行比較。其中，模型訓練統一使用 ESC-10 資料集以 5-Fold 交叉驗證來進行；而圖 3 至圖 6 的模型效能比較，則統一使用本研究自行蒐集之電鋸資料集，該資料集中不包含 ESC-10 之電鋸音檔。

3.2.4 實驗環境與超參數設定

所有模型都是在具有 8GB RAM 和 NVIDIA GeForce RTX3060 6G GPU 上進行開發及運行，所提出之實驗方法使用在 Windows10 作業系統上運行 Python 的開源 Pytorch 1.12 庫開發。批次大小為 64，使用 Adam 優化器 (Kingma and Ba, 2014)用於優化，學習率為 0.0002 每訓練 10 次將學習率降為原來的一半，損失函數方法為交叉熵(CrossEntropy Loss) (Zhang and Sabuncu, 2018)，共訓練 30 次，遷移式訓練的部分就是將模型以上述參數用不同長度音訊資料重複訓練。

4 實驗結果

本章節將實驗結果分為兩個階段，第一階段式模型正常訓練的結果，第二階段為模型進行遷移式訓練的結果。圖 3、圖 4 與圖 5 中的長條皆代表 5-fold 驗證的平均準確度，長條上的誤差線高點為 5-fold 裡最高準確度，低點為 5-fold 裡最低準確度。

4.1.1 模型依二元分類訓練後的準確度

圖 3 為模型使用二元分類為最終結果進行訓練後分別餵入不同秒數的判斷準確度。模型共有 3 個 分別以 1 秒、3 秒、5 秒進行訓練，並都餵入 1 秒、3 秒、5 秒進行預測。3 個模型對於 1 秒測試音訊分別有 53.52%、60.04%、53.68%的準確度，都高於 3 秒測試音訊的 49.47%、55.4%、44.87%，以及五秒測試音訊的 44%、50%、39.8%，結果來說二元分類模型對於電鋸聲音預測準確度最高的是以 1 秒進行預測。

4.1.2 模型依 ESC-10 分類訓練後的準確度

圖 4 為模型使用 ESC-10 分類為最終結果進行訓練後分別餵入不同秒數的判斷準確度。模型共有 3 個 分別以 1 秒、三秒、5 秒進行訓練，並都餵入 1 秒、3 秒、5 秒進行預測。3 個模型對於 1 秒測試音訊分別有 74.16%、74.16%、79.32%的準確度，3 秒測試音訊分別有 78.2%、76.13%、83.53%的準確度，5 秒測試音訊分別有 78.8%、77.6%、83.2%的準確度，結果來說不管幾秒訓練的模型，模型對於不同秒數的測試資料預測的準確度都差不多，但可以看出使用 5 秒進行訓練的模型準確度較高，在測試資料方面餵入較長秒數的預測準確度普遍較高，而且所有的預測準確度都比都比使用二元預測的方式高很多。

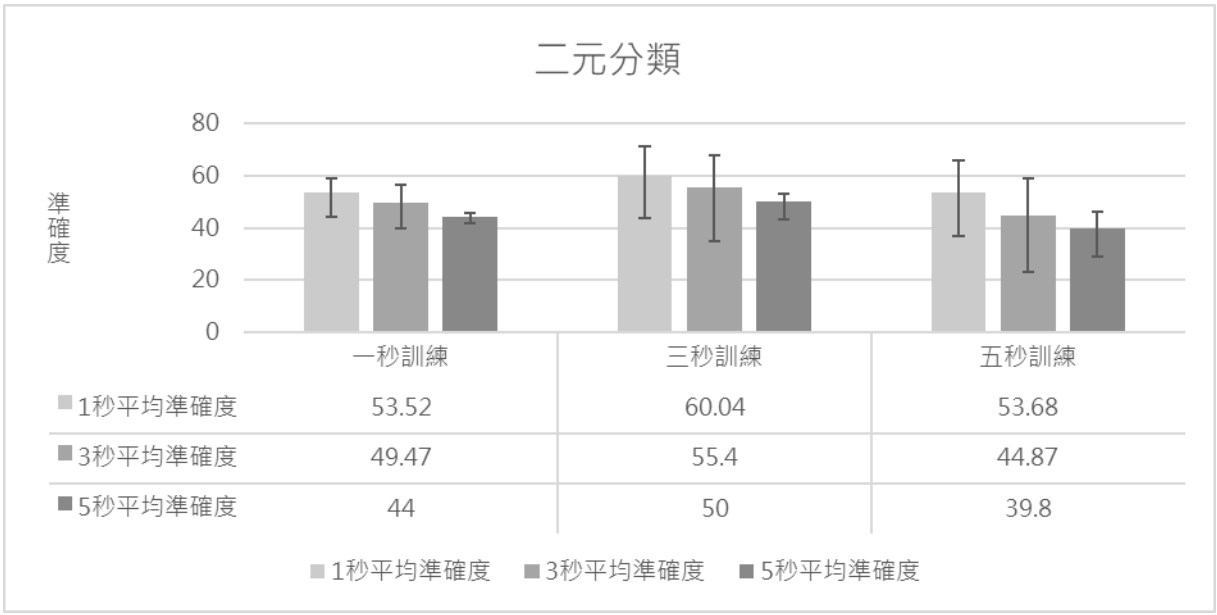

圖 3: 使用二元分類訓練後的準確度直方圖，模型分別以 1 秒、3 秒以及 5 秒進行單獨訓練，並且使用 1 秒、3 秒以及 5 秒的測試音訊進行準確度測試。

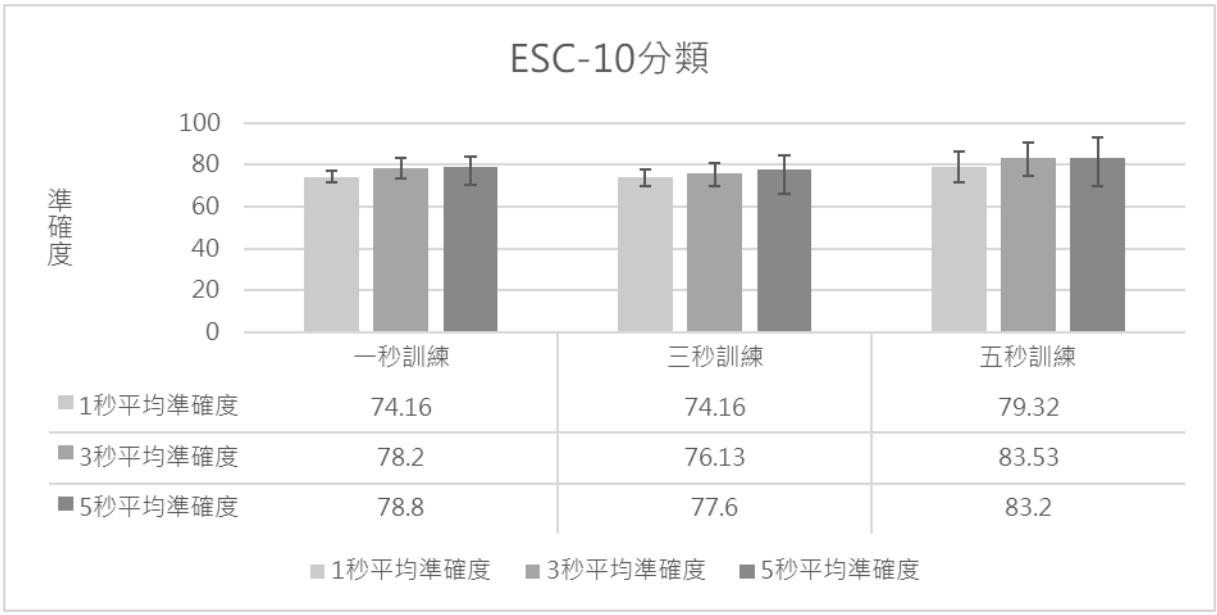

圖 4: 使用 ESC-10 分類訓練後的準確度直方圖，模型分別以 1 秒、3 秒以及 5 秒進行單獨訓練，並且使用 1 秒、3 秒以及 5 秒的測試音訊進行準確度測試。

4.2 模型使用遷移式訓練後的準確度

圖 5 為模型使用遷移式訓練後的準確度，模型分別以(1 秒→3 秒→5 秒)進行訓練，以及(5 秒→3 秒→1 秒)秒進行訓練，並都餵入 1 秒，3 秒，5 秒進行預測。2 個模型對於 1 秒測試音訊分別有 85.28%、74.52%的準確度，3 秒測試音訊分別有 88.67%、77.06%的準確度，五秒測試音訊分別有 91.8%、77.8%的準確度。與圖 4 表現最好的模型相比，以 5 秒進行訓練的模型的測試數據進行比較，可以發現將秒數由小到大(1 秒→3 秒→5 秒)進行訓練的模型，在 5 秒音訊的準確度提升了 8.6%，3 秒音訊提升了 5.14%，1 秒音訊提升了 5.96%，由大到小(5 秒→3 秒→1 秒)進行訓練的模型表現出較差結果的，而且在各秒數的準確度比起圖 4 中的任何模型都沒有進步。

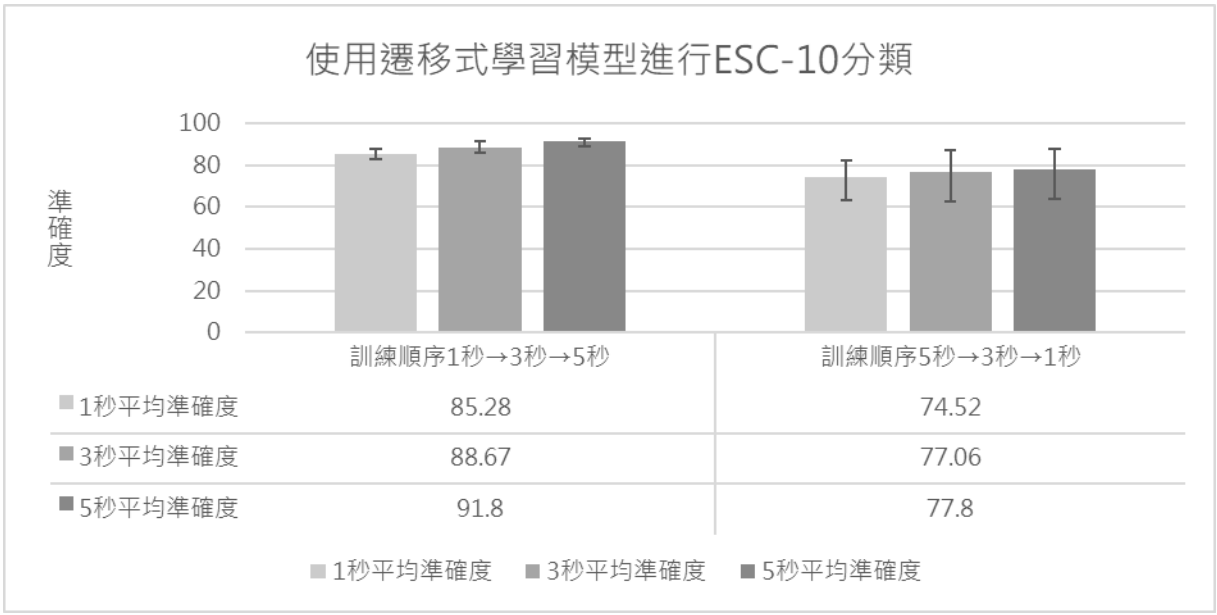

圖 5: 使用遷移式學習後模型進行 ESC-10 分類的準確度直方圖，模型分別以(1 秒→3 秒→5 秒)以及(5 秒→3 秒→1 秒)的順序進行遷移式學習，並且使用 1 秒、3 秒以及 5 秒的測試音訊進行準確度測試。

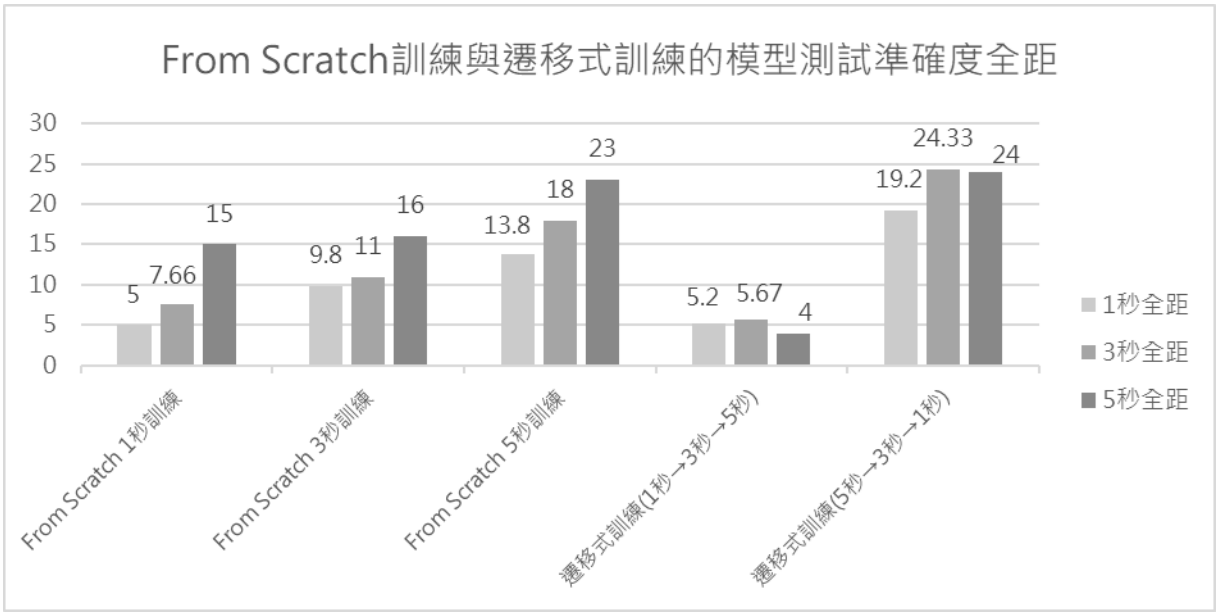

圖 6: 分別使用 From Scratch 以及遷移式學習訓練模型進行 ESC-10 分類 K-Fold 的預測準確度全距直方圖，模型分別以 1 秒、3 秒以及 5 秒進行單獨訓練以及(1 秒→3 秒→5 秒)、(5 秒→3 秒→1 秒)的順序進行遷移式學習，並且使用 1 秒、3 秒以及 5 秒的測試音訊進行預測值全距計算。對於 K-Fold 的預測準確度全距，全距計算方式為把 K-Fold 中預測最高準確度的值減去預測最低準確度的值。

4.3 模型使用遷移式訓練後對於 K-Fold 的預測全距

圖 6 為模型使用遷移式訓練後對於 K-Fold 的預測值全距，計算方式為將 K-Fold 預測最高準確度的值減去預測最低準確度的值。前三個模型為沒有經過遷移式學習只使用單一秒數(1 秒、3 秒、5 秒)進行訓練的模型，1 秒預測準確度值全距分別為 5%、7.66%、15%，3 秒預測準確度全距分別為 9.8%、11%、16%，5 秒預測準確度全距分別為 13.8%、18%、23%，後面為兩個模型使用遷移式訓練後的準確度，模型分別以(1 秒→3 秒→5 秒)進行訓練，以及(5 秒→3 秒→1 秒)秒進行，1 秒預測準確度全距分別為 5.2%、19.2%，3 秒預測準確度全距分別為 5.67%、24.32%，5 秒預測準確度全距分別為 4%、2.4%，訓練結果來說可以看到以(1 秒→3 秒→5 秒)進行遷移式訓練後

能有效的降低預測值全距，並且從圖 4 以及圖 5 的準確度中可以看到在降低全距的同時還可以大量的提升準確度，對於 5 秒的測試資料提升了 8.6%的準確度並降低了高達 19%的全距，對於 3 秒的測試資料提升了 5.14%的準確度並降低了 12.33%的全距，對於 1 秒的測試資料提升了 5.96%的準確度並降低了 8.6%的全距，研究表明遷移式訓練出來的模型增加了對於音訊特徵的提取性能增加了模型的泛化性還能提升模型判斷的準確度。

5 討論

從實驗結果可以看到以二元分類的方式進行訓練的模型成效不佳，推測是因為訓練時的資料分布太過偏斜因為餵進去的資料不是電鋸聲以及電鋸聲的音訊比例是 9 比 1 導致模型泛化能力降低，以正常 ESC-10 標籤進行訓練的模型就算用與訓練時使用的秒數不同的音訊進行判斷也有相當的準確度，有意思的是可以看到給予模型較長秒數進行預測的準確度普遍較高不管模型是用幾秒訓練的，但是這兩個訓練方式可以看到在 5-Fold 的準確度差異較大最高準確度與最低準確度差異較大，表現較好的以 ESC-10 分類的模型在輸入較長秒數進行預測時的最高最低準確度相差最多。使用遷移式訓練的模型中以遞增秒數方式進行訓練可以看到其在 5 秒的判斷準確度較高，3 秒、1 秒也有所提升，以遞減秒數方式進行訓練可以看到所有的判斷準確度都沒有進步而且誤差變大了，1 秒預測準確度全距分別為 5.2%、19.2%，3 秒預測準確度全距分別為 5.67%、24.32%，5 秒預測準確度全距分別為 4%、2.4%，所以經過研究表示，如果要將模型使用不同長度音訊進行遷移式訓練，從短音訊訓練到長音訊能得到較佳的結果，如果由長音訊訓練至短音訊並不會提升判斷的準確度，並且提高了模型的誤差，從短音訊訓練到長音訊其最高準確度與最低準確度差異明顯變小，因此正確的遷移式學習的確對於電鋸聲音的識別有所裨益。

6 結論

本研究提出一個問題使模型能接受不同於訓練音訊長度的音訊進行預測以及探討遷移式學習對於此種模型的幫助，不同訓練秒數對於模型的泛化能力的比較，實驗比較了兩種不同的標註方式以及在表現較好的標註方式上進一步的使用遷移式學習進行訓練，結果證明了模型能有效的進行泛化對於不同於訓練音訊長度的音訊也有著不差的準確度，在遷移式學習方面可以看到這種訓練方式能有效的提升模型的泛化能力以及準確度。

References

Chu, S., Narayanan, S., & Kuo, C. C. J. (2009). Environmental sound recognition with time–frequency audio features. IEEE Transactions on Audio, Speech, and Language Processing, 17(6), 1142-1158.

Dhanalakshmi, P., Palanivel, S., & Ramalingam, V. (2011). Classification of audio signals using AANN and GMM. Applied soft computing, 11(1), 716-723.

Hung, J. C., & Chang, J. W. (2021). Multi-level transfer learning for improving the performance of deep neural networks: theory and practice from the tasks of facial emotion recognition and named entity recognition. Applied Soft Computing, 109, 107491.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Liao, J. Y., Lin, Y. H., Lin, K. C., & Chang, J. W. (2021, December). 以遷移學習改善深度神經網路模型於中文歌詞情緒辨識 (Using Transfer Learning to Improve Deep Neural Networks for Lyrics Emotion Recognition in Chinese). In International Journal of Computational Linguistics & Chinese Language Processing, Volume 26, Number 2, December 2021.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8, pp. 18-25).

Piczak, K. J. (2015a, September). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP) (pp. 1-6). IEEE.

Piczak, K. J. (2015b, October). ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1015-1018).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVMs with a new set of features. International Journal of Innovative Computing, Information and Control, 8(5), 3511-3524.

Valero, X., & Alías, F. (2012, August). Gammatone wavelet features for sound classification in surveillance applications. In 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO) (pp. 1658-1662). IEEE.

Zhang, T., & Kuo, C. C. J. (2001). Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on speech and audio processing, 9(4), 441-457.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems, 31.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.