The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

環境音分類使用大規模預訓練模型以及半監督式訓練之初步研究
A Preliminary Study on Environmental Sound Classification Leveraging Large-Scale Pretrained Model and Semi-Supervised Learning

You-Sheng Tsao, Tien-Hong Lo, Jiun-Ting Li, Shi-Yan Weng, Berlin Chen
Department of Computer Science and Information Engineering, National Taiwan Normal University
{60947058s, teinhonglo, 60947036s, 60947007s, berlin}@ntnu.edu.tw

摘要

隨著智慧裝置的應用日漸普及，環境音分類技術的研究也越加受到重視。本論文探究環境音分類使用大規模聲音預訓練模型以及半監督式模型訓練模式。為此，我們首先使用大規模聲音預訓練模型來發展環境音分類方法；並且在假設標記資料匱乏的情境下，基於遷移學習(Transfer Learning)的概念下，使用近期被提出的 FixMatch 訓練演算法以及 SpecAugment 資料擴增技術來達到半監督訓練的目的。在環境音分類標竿資料集 UrbanSound8K 的實驗顯示，我們所提出的方法能較現有的基礎方法有 2.4%準確率提升。

Abstract

With the widespread commercialization of smart devices, research on environmental sound classification has gained more and more attention in recent years. In this paper, we set out to make effective use of large-scale audio pretrained model and semi-supervised model training paradigm for environmental sound classification. To this end, an environmental sound classification method is first put forward, whose component model is built on top a large-scale audio pretrained model. Further, to simulate a low-resource sound classification setting where only limited supervised examples are made available, we instantiate the notion of transfer learning with a recently proposed training algorithm (namely, FixMatch) and a data augmentation method (namely, SpecAugment) to achieve the goal of semi-supervised model training. Experiments conducted on benchmark dataset UrbanSound8K reveal that our classification method can lead to an accuracy improvement of 2.4% in relation to a current baseline method.

關鍵字：環境音分類、遷移學習、半監督學習
Keywords: Environmental Sound Classification, Transfer learning, Semi-supervised learning

1 緒論

隨著人工智慧與硬體技術的普及，在智慧裝置上應用語音辨識來下達指令已然成為司空見慣的事情，但是除了語音以外，若能夠辨識音訊種類也能夠為智慧裝置帶來更多的可能性，像是使用智慧音箱來對居家環境進行感知，以避免意外發生，又或者在智慧型手機上透過辨識周圍的環境音來套用相應的收音設定等。

本文從音訊標記 (Audio Tagging) 中的環境音分類 (Environmental Sound Classification) 任務著手，透過對輸入的聲音片段進行分類任務來辨別該聲音的種類。我們注意到，在發展環境音辨識時，尤其是針對特定環境的音訊標記任務，容易面臨到資料稀缺的問題，儘管相較於語音辨識，對音訊標記的資料集標註或許不需耗費龐大的人力資源，但對於監督式學習來說，如果想要讓模型能夠有更好的表現，勢必得提供大量的標記資料給模型進行訓練，若能使用少許的標註資料就能夠讓模型有相當的表現，又或者能夠利用未標註資料讓模型有辦法適應資料的分布，便可以在有限的資源下提升模型的能力。

為了突破訓練在目標領域(Target Domain)時標記資料稀缺的情況，我們提出一個結合兩

種做法的架構：首先使用預訓練模型來解決資料不足以讓模型學習特徵的問題，有了這樣的作法便能夠快速且強力的將訓練結果提升到一定的水準；接著，我們進一步以半監督式的遷移學習配合預訓練模型進行訓練，這樣的方式除了能夠在標註資料的音訊分類上保持該有表現外，也能夠善用未標註的資料，在目標領域的資料上有更好的擬合。細節上，我們使用音訊大規模預訓練模型專案 PANN (Pretrained Audio Neural Networks) (Kong et al., 2020) 作為基底，進行監督式的遷移學習，而在半監督學習的作法上，我們使用架構簡單但效果卓越的 FixMatch (Sohn et al., 2020) 對未標註的測試資料擬合，並為了實現 FixMatch 的技術需求，我們在音訊標記的任務上引入近期在語音辨識熱門的的資料增補技術 SpecAugment (Park et al., 2019)，對這個系統則使用環境音資料集 UrbanSound8K (Salamon et al., 2014) 進行評估，並分別對 FixMatch 的參數與方法上進行效果的比較，據我們所知，儘管在此類任務上有眾多個別使用預訓練或半監督學習的研究，但將預訓練模型與 FixMatch 結合仍尚未被討論過。

接下來會依序介紹相關研究，描述使用的模型與半監督訓練方法，在實驗設定章節會簡介資料集的結構與處理，並詳述訓練的參數，在實驗結果的部份，首先會對提出的架構與其他方法進行比較，再分析各項參數的調整，最後一節則是結論。

2 相關研究

這個章節主要會介紹音訊標記的相關做法、音訊標記之預訓練模型概述，以及本論文所聚焦的半監督訓練方法考察。

2.1 音訊標記

早期的做法中，音訊標記的流程與自動語音辨識的做法類似，同樣將音訊透過人為定義的函數提取特徵，再將這些特徵輸入如高斯混合模型 (Gaussian Mixture Model, GMM) 或是隱含馬可夫模型 (Hidden Markov Model, HMM) (Vuegen et al., 2013; Mesaros et al., 2010) 以得到機率的分佈後，使用這些模型做為辨識器來使用。近年來深度學習的發展如日中天，使用卷積網路架構 (Convolutional Neural Network, CNN) 來進行特徵辨識的做法逐漸變成主流，在電腦視覺的領域中，CNN 成功證明了它的有效性 (Krizhevsky et al., 2017)，而在音訊辨識方面，無論是對特定資料集的表現或者是知名的音訊辨識比賽 DCASE[1] 中，表現優異的做法也大多使用 CNN 作為基礎模型，再加上額外的特徵或是使用新穎的訓練手法 (Su et al., 2019; Sharma et al., 2020) 來增加模型的表現。

在 Google 發表了 Audioset 資料集後，音訊標記的發展便有了大幅的進步，該資料集包含 5000 多個小時從 Youtube 影片分離的音訊，並分成 527 種分類，而如同電腦視覺領域的 ImageNet (Deng et al., 2009) 一樣，在如此規模的資料幫助下，訓練出來的模型一般都有很不錯的泛化能力，方便進行後續的延伸應用，這也就引起大家對於預訓練模型的興趣。除了本篇論文引用的 PANN 以外，進一步的做法如 PSLA (Gong et al., 2021) 則是使用更多額外的訓練手法來增加對模型對 Audioset 的適應，而除了使用音訊資料集預訓練外，近來也有透過加入與分類相應的不同形式資料，如文字、圖像等讓模型能夠對於各個分類學到更強健的關聯性 (Jaegle et al., 2021; Guzhov et al., 2021) 或是一並使用圖像資料集預訓練的研究，如 Palanisamy et al. (2020) 就提出，將頻譜當成圖像在 ImageNet 預訓練模型上進行遷移學習，可以在更短的訓練次數內達到相當的辨識率，這樣的作法也在前述提到的 PSLA 中被應用。但綜合考量這些做法的易取用性以及結果上的表現後，我們挑選了易於修改的 PANN 做為實驗的框架。

2.2 半監督學習

在語音辨識的領域中，為了善用沒有譯文的資料，半監督學習也是一個熱門的研究領域，畢竟語音的資料集相較於普遍的分類任務，會需要花費更多資源進行標註，在這樣的背景下，將半監督學習套用在音訊標記的做法便迅速獲得我們的注意。

為數不少的半監督訓練方法是使用偽標籤 (Pseudo-Label) 來處理未標註資料集，具體做法為預先使用訓練過的模型來辨識這些未標

[1] http://dcase.community/challenge2020/index

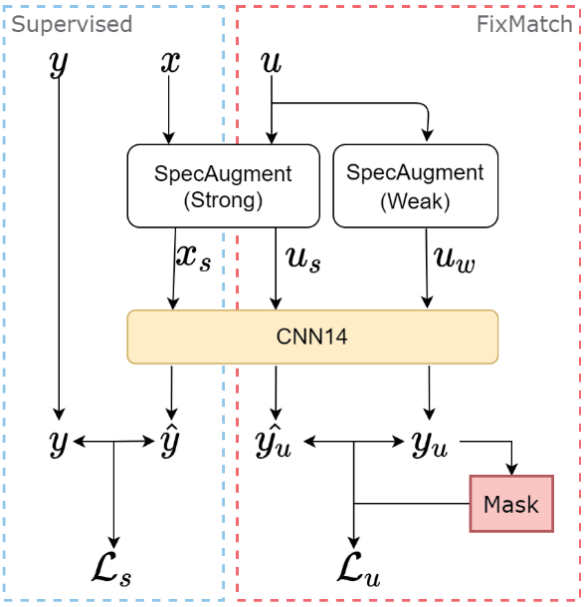

圖 1. 預訓練模型結合半監督方式訓練流程，藍色區塊為監督式學習，紅色為半監督學習，在一次的迭代中結合兩者的損失來進行訓練。

註資料，再根據如銳化函數 (Sharpening)、信心度等設定的條件來過濾及調整偽標籤的分布，最後便能將這些加上偽標籤的資料當成標註資料放進標註資料集進行訓練。

在近期常見的做法中，分別有對模型進行迭代，每次的新迭代都以新的模型加上前次計算的偽標籤進行訓練的 Teacher-Student 模式 (Xie et al., 2020)，以及整個訓練流程均使用同一個模型，並使用一致性正則化 (Consistency Regularization) (Sajjadi et al., 2016) 為核心概念的 MixMatch (Berthelot et al., 2019) 系列研究，前者需要多次迭代模型使得實作上稍微複雜，因此在這篇論文中我們使用了 MixMatch 的延伸研究，將流程去蕪存菁的 FixMatch 來進行實作，詳細作法會於下個章節詳述。

3 研究方法

為了弭平標註資料稀缺的問題，我們使用預訓練模型 PANN 讓模型有一定的能力對音訊特徵辨認，除了能夠短時間讓模型收斂，在收斂時也能夠達到比從頭開始訓練更好的效果，而為了應對未來模型在推論時會遇到的資料分布，我們進一步的使用了 FixMatch 對目標域進行偽標籤一致性的訓練，以更好的對未標註資料集做擬合。我們將提出的架構展示如圖 1，並對各部件做詳細的介紹。

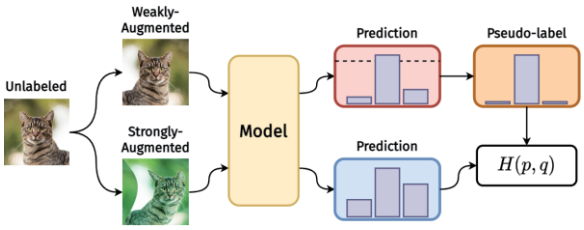

圖 2. FixMatch 訓練流程圖解

3.1 PANN

PANN 的主要目標是提供一系列不同架構的預訓練模型，訓練資料基於 Google 所發表的 AudioSet 資料集，期望能像 ImageNet 一般，以如此規模的資料集作為特徵抽取或是遷移學習的基礎模型。

在該論文中他們針對了不同的超參數與訓練方式做了多種實驗，發現使用 14 層的 CNN 模型能在嘗試過的架構中取得最佳的結果，並提出了一個對時間維度擷取的 1D-CNN 作為額外的特徵，稱做 Wavegram，將此特徵與 CNN-14 結合，成為他們的論文中成效最好的模型，但由於在初步實驗中對於 Wavegram 套用資料增補的成效不彰，最後我們只使用了 CNN-14 的結構。

PANN 在訓練的策略上，根據 AudioSet 中每個類別所包含的樣本數不一致問題做均勻取樣，實際做法為先從所有類別中抽樣，再從挑中的類別中隨機挑選樣本，以這樣的方式來避免訓練過程中因為特定類別的樣本太多而造成模型過於偏袒。

PANN 也使用了 Mixup (Zhang et al., 2018; Cances et al., 2021) 與 SpecAugment 等資料增補技術來增強模型的表現，在後續的遷移學習實驗中我們也沿用了這些資料增補技術。以這些方式訓練 AudioSet 能夠在 Wavegram-Logmel-CNN 與 CNN-14 的模型上，以平均正確率均值 (mean Average Precision, mAP)，即對所有分類的平均正確率作為評量標準，分別得到 0.439 與 0.431 的成績。

除此之外，考慮到訓練這樣的模型是為了便於在下游任務上使用，該論文中近一步的比較將 PANN 做為特徵抽取器，或是以微調 (Fine-tune) 方式進行遷移學習的方法優劣，前者會固定輸入側的訓練層參數，後者則是在輸出層換上新的全連接層進行訓練，在最後的實驗結果中則顯示，以微調方式進行訓練能夠在多數的下游任務中取得較佳的結果。

3.2 FixMatch

FixMatch 的訓練流程如圖 2 所示，主要的概念為一致性正則化 (Consistency Regularization) 的使用，翻成白話即是，即使對一個訓練樣本做了資料增補 (Data Augmentation) 之後，因為來源的樣本一致，結果的分布應該會是在同樣的領域內，所以模型需對這兩個樣本給出相近的標籤，在使用這樣的訓練限制下，透過訓練將兩個樣本的損失最小化，就可以運用這些未標記資料來增加模型的泛化性。

在取得偽標籤的做法中，對於同個未標註資料樣本 u 將會分別套用弱增補 $\mathrm{u}_{\mathrm{weak}}$ 以及強增補 $\mathrm{u}_{\mathrm{strong}}$，弱樣本經過模型得到預測後會使用閾值 τ 將可能性低於 τ 的預測遮住得到 mask，並將結果作為該樣本的偽標籤：

$$\mathrm{y_u} = P_m(\mathrm{u_{weak}}) \quad (1)$$

$$\mathrm{mask} = 1(\max(y_\mathrm{u}) \geq \tau) \quad (2)$$

有了偽標籤便能與同個樣本的強增補計算損失函數，將預測的損失能最小化：

$$\mathcal{L}_\mathrm{u} = \mathrm{mask} \cdot \mathrm{CE}\left(P_m(\mathrm{u_{strong}}), \mathrm{argmax}(y_\mathrm{u})\right) \quad (3)$$

之後使用交叉熵 (Cross-Entropy) 作為損失函數，將監督學習的損失與未標註資料損失相加，成為最終的損失函數：

$$\mathcal{L}_s = CE(P_m(x), y) \quad (4)$$

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_\mathrm{u} \quad (5)$$

在此處的 λ_u 為在兩者的損失做取捨的權重，我們參考的兩篇論文均將參數固定為 0.5。

也因為未標註資料會被過濾，故每次訓練的樣本數會是標註樣本的 μ 倍，這裡我們也參考原始論文的設定，使用 μ =7。

4 實驗設定

4.1 資料集

我們使用 UrbanSound8K 來進行效果的比較，這個資料集提供都市中多種類的環境音，包含了 10 種類別共 8,732 個標記音檔，每個音檔長度均少於 4 秒，並由官方提供 10 等分的交叉驗證 (10-fold cross validation)，每個等分大略包含了 870 個音檔，半監督訓練時的資料佔比則採訓練：驗證：未標註 = 1:1:8 的方式來驗證半監督訓練的有效性。

SpecAugment	Frequency		Time	
	max drop	drop nums	max drop	drop nums
Weak	8	1	32	1
Strong	35	2	64	2

表 1. SpecAugment 參數設定

輸入音檔的處理上，參考 PANN 提供之遷移學習專案的參數，將音檔重新取樣至 16kHz 以符合模型設定，並用 64 mel-bins 轉換至 log-mel spectrograms，並且為了實驗的簡單化，我們沿用專案的設定，所有音檔在讀取後填充至 5 秒長度。。其餘固定的訓練參數為學習率 = 5e-4，批數量 (batch_size) = 32。

4.2 資料增補

資料增補的部分，我們使用不同參數的 SpecAugment 作為半監督訓練的強增補與弱增補，監督式訓練則套用強增補參數，參考了 Weninger et al. (Weninger et al., 2020) 的設定以及 PANN 預設的參數，列於表 1。

在 Mixup 的部分我們比較了在遷移學習上使用與否對結果的影響。而在初步實驗中，我們也試著使用 Label-smoothing (Müller et al., 2020) 來增加模型的在偽標籤訓練過程的穩健性，或嘗試對標記資料加入 Pitch Shift 以及 Background Noise 進行實驗，但效果並沒有進一步的提升，故在最終實驗中沒有使用這三種方法。

4.3 訓練方式

PANN 的架構基於原論文的遷移學習專案[2]，並使用微調方式訓練整個模型及分類器層，預訓練模型則採用 CNN-14-16k 作為基準，原因除了對於 Audioset 的表現較好外，其架構相較 Wavegram-Logmel-CNN 更為簡潔，考慮到多了一種特徵增加訓練時間，進步卻有限的取捨中，我們便捨棄了 Wavegram 這個特徵；另外，在套用 FixMatch 的實驗裏，儘管我們嘗試將 SpecAugment 套用至 Wavegram，但訓練的效果上不但不會進步，甚至會傷害模型的表現，這樣的結果導致依賴資料增補技術的 FixMatch 實作變的困難，我們推測若沒有對 Wavegram 一併進行資料增補，FixMatch 在

[2] https://github.com/qiuqiangkong/audioset_tagging_cnn

	Accuracy	**$AUC_{weighted}$**
CNN14	39.55	84.00
PANN	74.55	**95.54**
+ FixMatch	**75.23**	93.28
-- w/o mixup	73.91	92.60

表 2. 僅使用單個等分資料與相同迭代次數的情形下，比較加入預訓練模型、FixMatch 以及去除 Mixup 時的表現，這裡的 FixMatch 閾值 $\tau = 0.95$

FixMatch	**Accuracy**	**$AUC_{weighted}$**
$\tau = 0.95$	75.23	93.28
$\tau = 0.85$	75.44	91.79
$\tau = 0.75$	**79.66**	**94.94**
+ Gradient Accum.	78.41	93.77

表 3. 使用不同閾值訓練之結果，並針對表現最好的閾值進行梯度累加實驗

訓練的過程中很有可能會過於依賴 Wavegram 這個相對容易的特徵，且預訓練模型也沒有訓練到 Wavegram 進行資料增補的部分，這些因素都增加了模型收斂的困難度。

我們將進行 FixMatch 訓練時的批數量減半至 16，為了應對 FixMatch 與 Mixup 合併使用的批數量需求，合併使用時未標註樣本的批數量會變成 $2 * \mu *$ batch_size，因此我們額外對梯度累加 (Gradient Accumulation) 技術進行實驗，以便於和監督式學習於同樣的參數上比較，並參照原始論文的做法，使用餘弦退火學習率排程器 (Cosine Annealing Learing Rate Schedular)，將其設定在每 35 輪 (epoch) 調整學習率，在原始論文及我們多次的實驗中發現，這樣的設定能夠有效的避免模型在訓練過程中被未標記資料過度影響而導致成效的衰退，或是訓練無法收斂。

5 實驗結果

在此章節中，我們先以一個等分的資料模擬資料稀缺的情境，呈現使用預訓練模型與 FixMatch 的差異，接著再比較 FixMatch 訓練時，使用 Mixup、調整閾值、梯度累加技術，及優化器 (Optimizer) 的差異，最後再以最佳參數進行 10-fold 交叉驗證的訓練，與其他模型進行比較。除了進行交叉驗證的訓練以外，其他均僅使用第一個等分訓練，第二個等分驗證，其餘等分做為未標記或不使用。

我們使用準確率 (Accuracy) 與 Area Under the ROC Curve (AUC ROC) 分數進行效能的評估，其中 AUC ROC 所測量的是模型能夠正確判斷出正負樣本的能力，該評量計算接收者操作特徵 (Receiver Operating Characteristic, ROC) 的曲線下面積 (Area Under Curve, AUC)，ROC 曲線若越往上凸則代表模型整體表現較佳，透過計算曲線下面積便可得到一個分數來衡量模型的表現。

5.1 使用 FixMatch 進行訓練

我們在一開始的實驗中先參考 Cances et al. (2021) 一文，套用他們對 US8K 訓練的 $\tau = 0.95$，在表 2 的部分呈現相同迭代次數下，使用隨機權重從頭訓練，以及逐步套用本論文提出的架構訓練的效果比較，可以看到在相同迭代次數時，PANN 可以很輕易的大幅度超越沒有經過預訓練的模型，這也是使用預訓練模型的優勢所在，並且再進一步使用 FixMatch 訓練的時候，因為有了未標註資料的幫助，所以能夠再推進 1% 的相對準確度。

接著我們實驗了使用 Mixup 與否的差異，這個技術的限制在於，訓練時會先取得批數量兩倍的樣本，再將樣本兩兩融合，這樣的資源需求在一般訓練上可能差異不大，但若配合上 FixMatch 必須先取得批數量 7 倍的未標註樣本進行過濾的限制後，一次迭代所需的樣本數就會來到原本的 16 倍之多，但在我們實驗的數據中顯示，不使用 Mixup 會讓模型表現的比僅用監督學習還差的情況，在這裡我們推測因為 PANN 是經過 Mixup 訓練過的模型，若在半監督遷移學習時不使用的話，這一階段的學習難度會就大幅下降，並且在難度低的情況下多次對目標域擬合，就會造成整體預測結果的衰退。

接下來會比較不同閾值的結果，以及前文提到之大量資料需求下，使用梯度累加的差異。儘管相關論文 (Sohn et al., 2020; Cances et al., 2021) 對於閾值的設定均有至少高於 0.75 的共識，但在預訓練模型的影響下，是否需要額外調整閾值仍值得探討。直覺上，我們會認為模型需要對預測結果非常肯定，這樣的預測才會對模型有幫助，我們在前一個實驗中選擇了 0.95 來實驗，但在圖像分類的原論文中，卻是以 0.75 得到最佳表現，於是我們依序的從 0.95 至 0.75 實驗了三個權重，得到結果如表 3。可以發現隨著閾值的下降表現也逐漸的進步，我們的解釋是，歸功於預訓練

FixMatch $\tau = 0.75$	Accuracy	$AUC_{weighted}$
SGD	71.29	94.86
ADAM	**79.66**	**94.94**
Ranger	75.17	94.54

表 4. 在 τ = 0.75 使用不同優化器訓練之結果

的優勢，使得在訓練上我們可以給比較低的閾值，讓半監督的結果不會被遷移學習的初期被某些少數超過閾值，但仍有可能分類錯誤的樣本給誤導，我們在分析資料集的過程中發現到，儘管使用了所有資料集進行監督學習，還是有模稜兩可的樣本容易混淆。

在得到最佳閾值為 0.75 後，我們使用梯度累加方式，試圖模擬批數量 32 的訓練結果，結果卻得到 1%的分數下降，推測是儘管在同樣的迭代次數下，多次更新的結果還是對半監督訓練比較有優勢的。

最後則比較了不同優化器在此次架構中的影響，我們比較了三種優化器，分別為 FixMatch 原始論文中表現最好的 SGD，PANN 預設的 AdamW，以及在機器學習比賽中號稱可以推進成績的 Ranger (Wright, 2019)，結果如表 4。出乎意料的，儘管訓練中我們使用了與原論文中一致的排程器，但在同樣的訓練次數中 SGD 卻是落後一大截，反而是 AdamW 能夠穩定的成長並得到最佳的表現，而 Ranger 雖然表現得不錯，但訓練過程中的驗證集卻會得到不穩定的辨識率，導致難以預測整體的結果。

5.2 與其他方法的比較

在監督式學習的部分，我們以兩種資料量來進行比較，表 5 中標有 100% 為使用所有可用標記資料來進行訓練，也就是資料充足的情況，而 10% 則是使用單個等分進行訓練，並且列出幾個使用所有標記資料進行訓練的方法做為此資料集目前準確率上限的參考。而在半監督學習的部分，我們主要與 Cances et al. (2021) 的實驗結果進行比較，該結果使用不含預訓練權重的 ResNet 模型進行 FixMatch 訓練，此表格內的分數均採用交叉驗證所得到的平均分數來呈現。

從監督訓練的結果發現，在預訓練模型的幫助下，使用所有資料訓練 PANN 最佳可以達到 95.5% 的準確度，相較於 ResNet 多出 13.45%，就算僅用單個等分，在同個結構上

	UrbanSound8K **Supervised**		Accuracy (%)	$AUC_{weighted}$ (%)
Others	ResNet (Cances et al., 2021)	– 10%	76.25	--
		– 100%	82.04	--
	AudioCLIP (Guzhov et al., 2021)		90.07	--
	TSCNN-DS (Su et al., 2019)		97.20	--
	ADCNN-5 (Sharma et al., 2020)		97.52	--
Ours	CNN14	– 10%	69.83	94.40
	PANN	– 10%	81.68	97.97
		– 100%	95.49	99.83
Semi-Supervised				
	ResNet + FixMatch (Cances et al., 2021)		81.73	--
	CNN14 + FixMatch		79.42	97.60
	PANN + FixMatch		**84.07**	**98.35**

表 5. 實驗結果，表格呈現交叉驗證的分數，比較相關研究及使用預訓練與半監督學習的效果，標註 CNN14 為乾淨的模型，PANN 則為預訓練模型

也比乾淨模型高出 11.85% 的準確度，使用極少比例的資料就足以與使用所有資料訓練的 ResNet 匹敵，而接下來為了能充分利用未標註資料，我們便進一步比較使用半監督訓練的效果。

我們將前個小節中得到的最佳參數對 FixMatch 進行訓練並呈現於同個表格，使用 τ = 0.75、AdamW 優化器且不使用梯度累加的情況下，可以提高準確度 2.39%。儘管在這樣的實驗結果上進步的幅度不大，但仍顯示了加入未標記資料可以讓模型學得更多，並贏過其他兩者，對於這樣小幅度的成長，我們推測可能是因為 PANN 已經充分地對音訊特徵擬合，使得 FixMatch 的長處，也就是適應目標領域特徵的能力，無法明顯的表現，不過以資源的消耗來說，使用預訓練模型的優勢就是能快速的立於良好的起跑點，對比的 ResNet 模型必須以每批次 256 個樣本訓練 300 輪 (epoch) 才能得到 81.73% 的成績，PANN 只需以 16 個樣本訓練 70 輪就能得到相近、甚至更佳的準確度，而這樣的訓練方式讓不管是標記資料或未標記資料都能對模型有所助益。

6 結論

此篇論文中，我們呈現了利用語音辨識技術的 SpecAugment 套用 FixMatch 半監督訓練於大規模預訓練模型的結果，並與未使用預訓練模型的作法相比，得到了微幅的成長，為我們所知在音訊標記領域中，首次以 FixMatch 搭配預訓練模型的做法，儘管這大部分歸功於預訓練模型對音訊特徵的適應，但相較於未使用預訓練的情況，可以大幅縮短訓練的時間，同時達到準確度的增加。在未來的半監督學習研究上，我們計畫能夠在訓練能時應對未標記的例外樣本進行排除，如 FixMatch 的延伸研究 OpenMatch (Saito et al., 2021)，使得模型在訓練時能夠更好的利用未標記資料。

參考文獻

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv:1905.02249 [cs, stat]*, October. arXiv: 1905.02249.

Léo Cances, Etienne Labbé, and Thomas Pellegrini. 2021. Improving Deep-learning-based Semi-supervised Audio Tagging with Mixup. *arXiv:2102.08183 [cs, eess]*, February. arXiv: 2102.08183.

Jia Deng, Wei Dong, R. Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *arXiv:2102.01243 [cs, eess]*, May. arXiv: 2102.01243.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. AudioCLIP: Extending CLIP to Image, Text and Audio. *arXiv:2106.13043 [cs, eess]*, June. arXiv: 2106.13043.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General Perception with Iterative Attention. *arXiv:2103.03206 [cs, eess]*, March. arXiv: 2103.03206.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *arXiv:1912.10211 [cs, eess]*, August. arXiv: 1912.10211.

Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *2010 18th European Signal Processing Conference*, pages 1267–1271. August.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. When Does Label Smoothing Help? *arXiv:1906.02629 [cs, stat]*, June. arXiv: 1906.02629.

Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. 2020. Rethinking CNN Models for Audio Classification. *arXiv:2007.11154 [cs, eess]*, November. arXiv: 2007.11154.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*:2613–2617, September. arXiv: 1904.08779.

Kuniaki Saito, Donghyun Kim, and Kate Saenko. 2021. OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers. *arXiv:2105.14148 [cs]*, May. arXiv: 2105.14148.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. *arXiv:1606.04586 [cs]*, June. arXiv: 1606.04586.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, Orlando Florida USA, November. ACM.

Jivitesh Sharma, Ole-Christoffer Granmo, and Morten Goodwin. 2020. Environment Sound Classification Using Multiple Feature Channels and Attention Based Deep Convolutional Neural Network. In *Interspeech 2020*, pages 1186–1190. ISCA, October.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv:2001.07685 [cs, stat]*, November. arXiv: 2001.07685.

Yu Su, Ke Zhang, Jingyu Wang, and Kurosh Madani. 2019. Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. *Sensors*, 19(7):1733, January.

Lode Vuegen, Bert Van Den Broeck, Peter Karsmakers, Jort Florent Gemmeke, and Hugo Van hamme. 2013. An MFCC GMM approach for event detection and classification. October.

Felix Weninger, Franco Mana, Roberto Gemello, Jesús Andrés-Ferrer, and Puming Zhan. 2020. Semi-Supervised Learning with Data Augmentation for End-to-End ASR. *arXiv:2007.13876 [cs, eess]*, July. arXiv: 2007.13876.

Less Wright. 2019. New Deep Learning Optimizer, Ranger: Synergistic combination of RAdam + LookAhead for the best of both. September.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with Noisy Student improves ImageNet classification. *arXiv:1911.04252 [cs, stat]*, June. arXiv: 1911.04252.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]*, April. arXiv: 1710.09412.