

NANYANG TECHNOLOGICAL UNIVERSITY

用于声学事件分类的持续学习

Xiao Yang

计算机科学与工程学院

2022

arXiv:2512.17932v1 [eess.AS] 10 Dec 2025

南洋理工大学

MSAI硕士项目 MSAI/21/044

端到端声学事件分类

提交人：Xiao Yang，指导教授：
Prof.Chng Eng Siong

计算机科学与工程学院

2022

原创性声明

本人特此证明，本报告所载的工作是原创研究（和/或研究综述）的成果，不含剽窃材料，且未曾提交给其他任何大学或机构以申请更高学位。

[学生在此处签名]

入学证件上的学生姓名 日期： [DD/MM/YYYY]

导师声明

我已审阅本报告的内容和呈现形式，并声明报告不存在抄袭，且语法清晰度足以接受审查。据我所知，除作者归属声明中所注明的部分外，研究（和/或研究综述）及文字撰写均为候选人本人完成。我确认相关研究按照南洋理工大学的伦理政策和诚信标准开展，且研究数据均如实呈现，不存在偏见。

[导师在此处签名]



教职员档案上的导师姓名 日期： [
DD/MM/YYYY]

作者归属声明

请从以下选项中选择一项，并根据情况划去另一项：

(A) 本报告**不**包含任何发表于同行评审期刊的论文或在会议上获接收、且列我为作者的论文中的材料。(B) 本报告**包含**以下发表于同行评审期刊的论文和/或在会议上获接收、且列我为作者的论文中的材料。

(B)

如果您选择上述选项（B），请提供与本报告相关的出版物（期刊/会议）列表。请明确说明您在该项工作中的贡献。

1. Yang Xiao*、Xubo Liu*、James King、Arshdeep Singh、Eng Siong Chng、Mark D. Plumbley。设备端环境声音持续学习分类。DCASE研讨会。（2022年）

（a）在这篇论文中，我的贡献主要是开展实验并分析结果。我还通过实现这些方法，为研究作出了贡献。

(b) Xubo 和我提出了核心研究构想，并与 James 共同准备了论文初稿。Arshdeep 为所采用的 BC-ResNet 模型作出了贡献。Mark D. Plumbley 教授、Wenwu Wang 教授和 Eng Siong Chng 教授参与了论文初稿的修订和综述。

2. Yang Xiao、Nana Hou、Eng Siong Chng。《Rainbow Keywords：用于在线口语关键词检测的高效增量学习》。INTERSPEECH会议。（2022年）

(a) 在本文中，我提出了核心研究构想，实现了所提出的方法，并准备了论文初稿。我还通过开展实验和分析结果，为这项研究作出了贡献。

(b) 我的合著者主要参与了论文初稿的修订和综述。

[请在此处签署]

学生姓名（须与入学证件一致） 日
期： [日/月/年]

目录

摘要 iii

致谢 iv

缩略词 v

图目录 viii

表目录 ix

1 引言 1
1.1 背景 .11.2 研究动机.21.3 论文的主要贡献 .21.3.1 用于关键词检测的持续学习 .21.3.2 用于环境声音分类的持续学习 .21.4 论文结构 .3

2 文献综述 4
2.1 关键词检测 .4
2.1.1 定义与背景 .42.1.2 关键词检出方法 .42.1.3 关键词检测数据集 .8

2.2 环境声音分类 . 11
2.2.1 定义与背景 .112.2.2 环境声音分类方法 .112.2.3 声音分类数据集 .16

2.3 持续学习 . 18
2.3.1 定义与背景 .182.3.2 持续学习方法 .19

3 持续学习的口语关键词检测 2 0
3.1 相关工 orks 20
3.2 M方法 20
3.2.1 多样性感知采样器 .213.2.2 数据增强 .223.2.3 知识蒸馏损失 .22

3.3 实验与结果 . 23
3.3.1 数据集 .233.3.2 实验设置 .233.3.3 结果 .25

3.4 Summary 27

4 环境声音分类的持续学习 284.1 相关工作 .284.2 方法 .284.2.1 基于回放的持续学习 .
294.2.2 提出的 MUA 方法（不确定性++） .314.3 实验与结果 .314.3.1 环境声音分类模型 .
314.3.2 数据集.314.3.3 实验设置 .324.3.4 评估指标 .324.3.5 参考基线 .324.3.6 结果 .334.4 总
结 .34**5 结论与建议 35**

5.1 结论 .355.2 未来工作 .355.2.1 不同信噪比条件下的持续学习 .355.2.2 多语言持续学习 .
36

附录 A **45**

摘要

在计算资源（例如模型大小、运行内存）受限的情况下，如何在不发生灾难性遗忘的前提下持续学习新类别，是端侧声学事件分类面临的一项挑战。为缓解这一问题，我们针对口语关键词检测和环境声音分类提出了两种新颖的多样性感知增量学习方法。我们通过衡量逐样本分类不确定性来选择用于训练的历史数据。对于口语关键词检测应用，所提出的RK方法引入了一种多样性感知采样器，通过计算分类不确定性，从历史关键词和新输入关键词中选取多样化的集合。因此，RK方法能够在增量学习新任务的同时避免遗忘先验知识。此外，RK方法还提出了数据增强和知识蒸馏损失函数，以便在边缘设备上进行高效的内存管理。对于环境声音分类应用，我们通过观察数据的分类概率在分类器嵌入上添加并行扰动后如何变化来衡量不确定性。与对原始数据添加扰动相比，这种方法可以显著降低计算成本。

实验结果表明，所提出的RK方法在*Google Speech Command* 数据集上的平均准确率比最佳基线高出4.2个百分点，同时所需内存更少。在*DCASE 2019* 任务 *1*和*ESC-50* 数据集上的实验结果表明，我们提出的方法在分类准确率和计算效率方面均优于基线持续学习方法，说明该方法能够针对设备端环境声音分类，高效地逐步学习新类别，同时避免灾难性遗忘问题。

关键词： 声学事件分类、环境声音分类、关键词检测、持续学习。

致谢

无数人支持我完成这篇学位论文。Chng Eng Siong教授对我的分析和论述方式提出了极其宝贵的意见，有时甚至会在深夜或清晨回复邮件。我非常珍惜每周与团队一起参加的会议。作为SpeechLab@NTU的新成员，他总是为我的出版物提供正确的方向。感谢我的导师Chng教授，感谢您的耐心、指导与支持。我非常感激您愿意接收我成为您的学生，并在这一年里始终对我抱有信心。

写作期间，还有许多人给了我宝贵的建议，包括 Ng Dian Wen 等人。Dian Wen 是一位出色的合作者，也是对我耐心十足的兄长。今年，他在项目上的指导和想法让我受益匪浅。此外，Chen Chen、Yuchen 和实验室里的其他朋友也给予了我帮助，并与我进行了许多高质量的讨论。我会非常怀念这段时光。祝愿他们的研究之路越走越好。

我尤其要感谢 Hou Nana博士。您的丰富学识和细致的编辑让我受益良多。若没有您的指导，我绝不可能学会如何开展研究，也无法发表我的第一篇会议论文。

感谢我的父母，始终陪伴在我身边；即使我自己不这么觉得，你们也总是告诉我我很棒。爸爸，谢谢你给予我的一切爱，也总是提醒我最终的目标。

有人说：“巨人照着镜子，却什么也看不见。”现在，我将迈向人生的下一站。最后，我想感谢日夜努力的自己。

缩略词

CNN 卷积神经网络**CL** 持续学习**RK** 提出的 Rainbow Keyword 方法**RCL** 基于回放的持续学习**MUA** 记忆更新算法 **KWS** 关键词检测**ASR** 自动语音识别**FFNN** 全连接前馈神经网络**ReLU** 修正线性单元**MFCC** 梅尔频率倒谱系数**DS-CNN**深度可分离卷积神经网络**RNN** 循环神经网络**BiLSTM**双向长短期记忆网络**ViT**视觉 Transformer**SSL** 自监督表征学习**MISP**基于多模型信息的语音处理**DCASE** 声学场景与事件检测和分类 **CRNN** 卷积循环神经网络**ESC** 环境声音分类**GMM** 高斯混合模型

AST 音频频谱图 Transformer**SOTA** 最先进水平**SSAST** 自监督 AST**MSPM** 掩码频谱图图块建模**类别增量学习** 类别增量学习**任务增量学习** 任务增量学习**蒙特卡罗** 蒙特卡罗**知识蒸馏**知识蒸馏**GSC** Google Speech Command 数据集**准确率** 平均准确率**后向迁移** 向后迁移

图目录

2.1 关键词检测系统概述.*4 2.2* 完整架构，以及放大的残差块 *[33] .6 2.3 [52]*中的关键词 *Transformer* 架构。音频经过预处理，转换为梅尔尺度频谱图，并沿时域划分为互不重叠的图块。再加上学习得到的类别令牌，这些构成了多层 *Transformer*编码器的输入令牌。与 *ViT [50],* 一样，每个令牌都会加上学习得到的位置嵌入。类别令牌的输出经过线性分类头处理，用于进行最终的类别预测.*7 2.4* 在下游应用中使用自监督表征学习的框架，摘自 *[57].8 2.5 [58]* 中的词语列表，这些词语收录于 *Google Speech Commands* 数据集 *v1*（前六行）和 *v2*（所有行）中。词语按标准化的 *10* 个关键词（前两行）和非关键词（后五行）分类.*9 2.6 MISP2021-AVWWS* 语料库概述。 *[P*：表示存在唤醒词，*N*：表示不存在唤醒词*] .9 2.7*： *[24]*中的 *BCResBlock*。*BC-ResNet* 块包含一个频率深度卷积和一个子频谱归一化层。随后对特征沿频率维度取平均，再进行时间深度可分离卷积。时间特征在残差连接处广播至二维特征。在过渡块中，前端额外加入一个 *1x1* 卷积，以改变通道数，且不使用恒等捷径连接。. . . . . *12 2.8*： *[25]*中的 *BC-ResNet-Mod*。每一行表示由一个或多个相同模块组成的序列，重复 *n* 次；输入形状为频率×时间×通道，时间步总数为 *T*。*13 2.9 [82]*中的音频频谱图 *Transformer*（*AST*）架构。二维音频频谱图被分割为一系列有重叠的 *16×16*图块，随后线性投影为一系列一维图块嵌入。每个图块嵌入都加上可学习的位置嵌入。序列前端还会添加一个分类令牌。输出嵌入作为 *Transformer* 的输入，分类令牌的输出则通过线性层用于分类.*14 2.10 [83]*中的自监督 *AST*。二维音频频谱图被分割为一系列互不重叠的*16×16* 图块，随后线性投影为一系列一维图块嵌入 *E*。每个图块嵌入都加上可学习的位置嵌入 *P*，然后输入 *Transformer*编码器。*Transformer* 的输出 *O* 用作频谱图图块表征。在自监督预训练期间，模型会随机掩码部分频谱图图块，并要求模型：*1*）从所有被掩码的图块中找出每个被掩码位置对应的正确图块；*2*）重建被掩码的图块。这两个前置任务旨在促使 *AST* 模型学习音频数据的时间结构和频率结构。微调期间，对所有图块表征 {*O*} 进行均值池化，并使用线性分类头进行分类.*15 2.11 [84]* 中的类别.*16 2.12 [23]* 中的类别.*17*

2.13 在持续学习中，互斥任务按顺序学习。任务增量学习在评估期间可以访问任务 *ID*，而更具挑战性的类别增量学习则不能。本学位论文研究的是类别增量学习*.18 3.1* 所提出的彩虹关键词方法的框图。具体而言，D_S^t 表示任务 τ_t的输入音频流数据。D_E^t 和 D_E^{t-1}分别表示任务τ_t和 τ_{t-1}的样本。我们按唯一关键词将 $D_E^{t-1} \cup D_S^t$分组为子集 D_c, $c = 1..N^t$，其中 N^t 表示 $D_E^{t-1} \cup D_S^t$集合中唯一关键词的总数。x、$\hat{}x$ 和 K 分别表示 D_c中的每个样本、x 的五种扰动以及五种扰动策略。“计算 *u(x)*”表示根据公式*3.3*计算 *u(x).21 3.2 TC-ResNet-8* 的架构。它分别使用 *ResNet-8* 作为骨干卷积神经网络。*FC* 表示全连接层。注意，“*s*”、“*c*”和“*k*”分别表示步幅、通道大小和宽度乘数*.23 3.3* 所提出的 *RK* 方法与其他有竞争力的基线方法在增加任务数量时的对比研究准确率（%）。图 *(a)* 至 *(d)* 分别表示任务数量为 *(= 20, 10, 5, 4)* 时的实验*.25 3.4* 所提出的 *RK* 方法与其他有竞争力的基线方法在不同内存大小下的对比研究准确率（%）*.26 4.1* 原生不确定性方法与我们所提出方法的框图。具体而言，朴素方法通过波形向 x 添加扰动，并生成多个波形作为 $\hat{}x$。我们的方法输入嵌入 e，并生成扰动后的嵌入 $\hat{}e$，这意味着我们只保存嵌入。模型骨干网络的输出只计算一次。“计算 *u(x)*”表示根据公式（*4.3*）计算 *u(x)*。K 表示通过扰动方法生成的扰动数量*.30 5.1* 不确定性采样器的 *Python* 代码*.45*

表格目录

3.1 所提出的知识蒸馏损失的对比研究中的平均准确率（*ACC*）和反向迁移（*BWT*）。*L* 表示 *RK* 的内存大小。 .253.2 所提出的数据增强方法的对比研究中的平均准确率（*ACC*）和反向迁移（*BWT*）。*"NoAugment"*表示实验中未应用数据增强。*L* 表示 *RK* 的内存大小。 .263.3 近期最先进训练策略的对比研究中的准确率和效率指标。为公平比较，我们采用 *TC-ResNet-8* 模型作为所有训练策略的测试平台。内存大小 *L* 设为 *[500, 1500, 3000]*在后续 *RK* 实验中。 274.1 所提出的记忆更新算法的对比研究中的准确率（*ACC*）和反向迁移（*BWT*）。 .324.2 所提出的扰动方法的对比研究中的准确率（*ACC*）和反向迁移（*BWT*）。*K* 表示通过扰动方法生成的扰动数量。 .334.3 所提出的不确定性方法的对比研究中的平均时间（秒）++ 。*K* 表示通过扰动方法生成的扰动数量。 34

第1章

引言

1.1 背景

音频分类是指为音频片段分配标签的一系列任务 [1]。音频分类有许多应用，例如声学场景分类 [2], 声音事件检测 [3] 和关键词检测 [4]。音频分类是信号处理和机器学习领域的重要研究课题。音频分类在许多现实应用中发挥着关键作用，包括声学监测 [5], 医疗保健 [6] 和多媒体索引 [7]。

卷积神经网络（CNN）等神经网络方法已用于音频分类，并取得了最先进的性能 [8]。在许多现实场景中，音频分类模型需要部署在移动设备等资源受限平台上 [9]。因此，当前基于深度学习的音频分类系统通常会在紧凑型模型中使用有限类别进行训练，以降低计算开销并减小模型体积 [10]。

因此，当运行时遇到目标域中未见过的类别时，使用源域数据训练的模型性能可能会显著下降。当模型开发者希望扩展待分类的音频类别时，一种方法是使用新类别的数据对模型进行微调。然而，这种方法可能会在微调过程中丢弃先前学到的知识，这也称为灾难性遗忘问题 [11]。另一种可能的解决方案是使用历史数据与新数据的混合数据重新训练分类模型。然而，在现实的设备端场景中，这种方法会消耗大量资源和时间。由于基于重新训练的解决方案计算开销较大，因此设计高效且有效的方法，使训练好的设备端音频分类模型能够适应新的声音类别至关重要。

持续学习（CL） [12] 旨在随着时间的推移持续学习新知识，同时保留并重新利用先前学到的知识。近年来，在图像分类 [13], 、机器人学 [14] 和自然语言处理 [15]等深度学习任务中，持续学习方法已展现出良好前景，其结果优于微调方法。此外，一些研究人员还探索了持续学习在音频处理领域的应用，例如 [10] 和 [16]。然而，关于设备端应用中的持续学习（例如设备端音频分类），现有文献关注较少，这正是本学位论文的研究重点。设备端场景通常受到存储和内存空间 [17], 的限制，这可能给依赖外部存储器来恢复历史数据的持续学习带来挑战。因此，可在设备上运行的音频分类模型的容量可能受到限制，从而在持续学习新的声音类别时容易遗忘旧知识。

1.2 研究动机

鉴于计算资源（例如模型大小、运行内存）受到限制，在设备端音频分类中持续学习新类别而不发生灾难性遗忘是一项具有挑战性的问题。本学位论文的目标是寻找解决这一问题的方法。我们将研究用于分类的最先进持续学习方法，然后针对不同的音频分类应用提出最合适的方法。

1.3 学位论文的主要贡献

本学位论文提出了两种持续学习方法，用于解决两种不同设备端音频分类任务中的遗忘问题。

1.3.1 关键词检测的持续学习

我们提出了一种名为 Rainbow Keywords（RK）的新型多样性感知持续学习方法，以解决上述问题；该方法无需任务 ID 信息，且所需参数更少。具体而言，所提出的 RK 方法引入了一种多样性感知采样器，通过计算分类不确定性，从历史关键词和新输入关键词中选取少量但具有多样性的样本。因此，即使仅使用有限的历史样本，模型在学习新关键词时也不会遗忘先验知识。此外，我们采用混合标签数据增强，进一步提高所选样本的多样性，从而提升性能。同时，我们提出了一种知识蒸馏损失函数，以确保先验知识能够通过有限的所选样本得以保留。我们在*Google Speech Command* 数据集上按照先前工作的设置 [13, 18]开展实验。实验结果表明，所提出的 RK 方法在平均准确率方面比最佳基线绝对提升了 4.2%，且所需内存更少。相关脚本可在 GitHub 上获取 [1]。

1.3.2 环境声音分类的持续学习

我们研究了用于设备端环境声音分类的基于回放的持续学习（RCL）方法。首先，我们研究现有记忆更新算法（MUA）方法的性能，例如蓄水池 [19],原型 [20] 和 不确定性 [21]（如第4.2.1节所述）在设备端环境声音分类的 RCL 中的表现。

我们通过实验证明，不确定性 [21] 方法在我们的场景中表现最佳。此外，我们提出了不确定性++，这是一种基于不确定性 方法的简单高效 MUA 方法。与不确定性 方法不同，我们提出的不确定性++ 将扰动引入分类器的嵌入层。因此，在测量数据不确定性时，计算成本（例如运行内存和时间）可以显著降低。我们在 DCASE 2019 Task1 [22] 和 ESC-50 [23]数据集上，使用端侧模型 BC-ResNet-Mod（~86k 个参数） [24, 25]评估了我们方法的性能。实验结果表明，不确定性++ 在分类准确率方面优于现有 MUA 方法，显示出其在实际端侧音频应用中的潜力。我们提出的方法与模型无关，且易于应用。我们的代码已发布在 GitHub[2]。

[1] https://github.com/swagshaw/Rainbow-Keywords[2] https://github.com/swagshaw/ASC-CL

1.4 论文结构

本学位论文分为五章，各章概述如下：

- 第2章全面综述了音频分类领域两个典型应用中的相关工作，涵盖从传统方法到当前深度学习方法的演进。本章还介绍了训练和测试中使用的各种热门基准。随后，第2章还概述了持续学习方法。

- 第3章概述了所提出的用于关键词检测的 Rainbow Keywords 方法，并详细介绍了基于 *Google Speech Commands* 数据集开展的实验。

- 第4章概述了所提出的用于环境声音分类的不确定性++ 方法。随后，我们在 DCASE 2019 任务1 [22] 和 ESC-50 [23] 数据集上，使用设备端模型 BC-ResNet-Mod（~86k 个参数） [24, 25]进行了性能评估。

- 第5章总结了本学位论文，并概述了未来的工作。

第2章

文献综述

音频分类是音频处理的重要领域之一，已广泛应用于语音助手等多种现实场景。人类能够听到的声音频率范围为20 Hz至20 kHz，这取决于施加在鼓膜上的压力 [26]。基于其来源，我们将音频分类任务简要划分为两类：语音和自然声音。在本学位论文中，我们针对这两类中的典型任务提出了持续学习方法，分别是关键词检测和环境声音分类。

因此，为全面综述相关研究，2.1 节简要介绍了基于深度学习的现代关键词检测方法。接下来的2.2节综述了环境声音分类的各种方法。随后，2.3 节探讨了持续学习方法及其演变。

2.1 关键词检测

2.1.1 定义与背景

口语关键词检测（KWS） [4] 旨在识别音频输入中的特定关键词。它是许多现实应用中的主要模块，例如苹果 Siri 和 Google Home，这些应用广泛用于边缘设备。语音助手的一个显著特点是，要使用它们，必须先通过口头唤醒词或关键词将其激活。这样便无需运行计算开销高得多的自动语音识别（ASR）。具体而言，关键词检测（KWS）可定义为：在包含语音的音频流中识别关键词的任务。此外，除用于激活语音助手外，KWS 还广泛应用于语音数据挖掘、音频索引和电话路由等领域。 [27]。多年来，人们一直在探索各种用于 KWS 的技术。

2.1.2 关键词检出方法

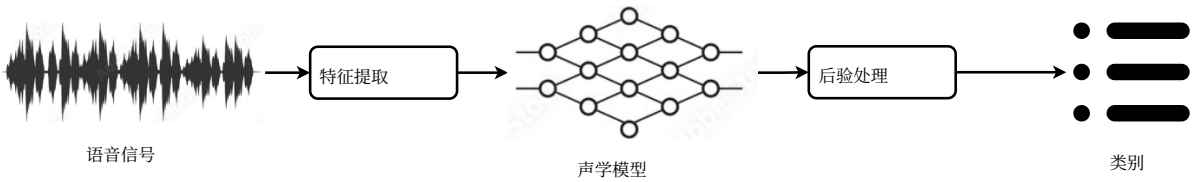


图2.1：关键词检测系统概览

图2.1　展示了现代深度语音关键词检测系统的一般流程。该系统　m

由三个主要模块组成：它从音频特征中尝试识别不同的关键词和填充（非关键词）类别；3）后处理器处理时间序列后验概率，以确定输入信号中可能存在的关键词。

本节介绍声学模型，它是深度口语关键词检测系统的核心。自然的做法是在降低计算复杂度的同时，设计出更准确的模型。

全连接神经网络：

2014年，深度口语关键词检测开始采用当时最流行的神经网络架构类型所构建的声学模型。全连接前馈神经网络（FFNN） [28]。由三个全连接隐藏层组成的简单堆叠结构，每层有128个神经元，并采用修正线性单元（ReLU）激活函数，随后接一个Softmax输出层，其参数量和（当时的）状态数更少，而且性能显著优于在干净和嘈杂声学条件下的关键词/填充隐马尔可夫模型系统。然而，由于目标转向设计更准确、更稳健且计算开销更低的声学模型，全连接FFNN很快便退居次要地位。

与全连接FFNN密切相关且计算开销更低的替代方案包括单值分解滤波器（SVDF） [1, 29, 30]和脉冲神经网络 [31]。一种密切相关且计算开销较低的替代方案是全连接FFNN，它采用单值分解滤波器（SVDF）和脉冲神经网络。SVDF于 [30], 提出，通过低秩近似来近似全连接层，使首个深度关键词检测系统的FFNN声学模型 [28] 在不影响性能的情况下将规模缩减了75%。脉冲神经网络（SNN）以事件驱动方式处理信息。如果关键词检测中的信息是稀疏的，这种方式便能显著降低计算负载 [31]。

卷积神经网络：

[32] 是2015年从全连接前馈神经网络转向卷积神经网络的信号。人们欣赏卷积神经网络利用局部音频时频相关性进行深度关键词检测声学建模的能力；它所需的参数更少，性能也优于全连接前馈神经网络。这也是卷积神经网络的吸引力之一。通过调整滤波器、步幅、卷积核和池化尺寸等各种超参数，可以轻松限制模型，以满足计算限制。此外，无须牺牲太多性能即可实现这一点。

Tang 和 Lin [33] 是《探索深度关键词检测中的深度残差学习》的原作者。他们还引入了扩张卷积，以扩大网络的感受野并捕捉更长的时频模式，同时不增加参数数量 [34–36]。因此，Tang 和 Lin 在参数更少的情况下，关键词检测性能显著优于标准卷积神经网络 [32] ，并于2018年实现了新的最先进水平。

这一成功为后续研究提供了很大的动力。人们正在考虑采用深度残差学习。Choi 等人 [37] 将梅尔频率倒谱系数（MFCC）作为深度残差学习框架（TC-ResNet）的输入通道。这种方法有助于克服网络深度有限时难以同时捕获高频和低频特征的挑战；二维扩张卷积也能在很大程度上实现这一点，因为它扩大了网络的感受野。我们提出，与参数数量相同的二维卷积相比，时间卷积显著降低了计算负载。因此，TC-ResNet 的关键词检测性能与 Tang 和 Lin 的 [33]相当，同时显著降低了其在移动设备 [37]上的延迟和每秒浮点运算次数。与关键词检测、标准卷积神经网络、卷积循环神经网络（CRNN）以及带注意力机制的循环神经网络相比，TC-ResNet 的延迟和模型大小均处于最低之列，并且性能更优（另见下一节）；它也优于基于竞争性声学模型的方法（另见下一节）。因此，我们采用 TC-ResNet-8 [37]作为测试平台，评估所提出的彩虹关键词方法。

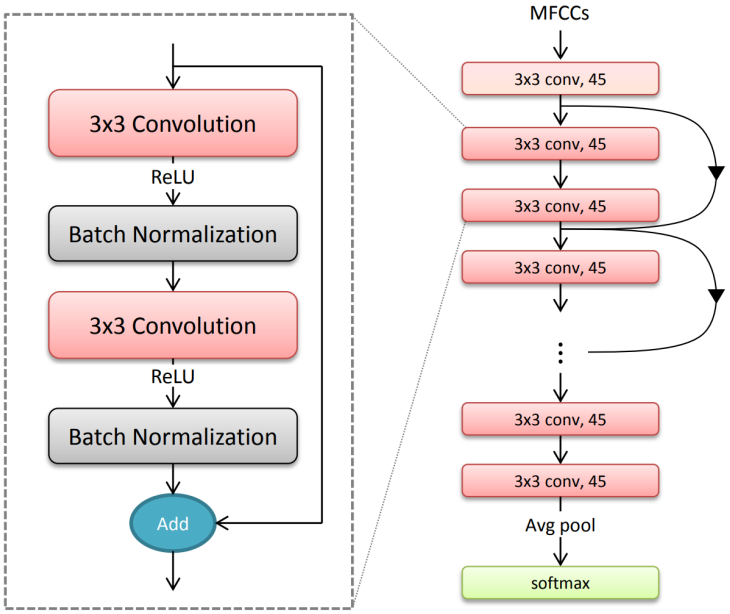


图 2.2：完整架构，以及从 *[33]* 中放大的残差块

另一种颇具吸引力的方法 [38] 能够减少计算量。标准卷积神经网络的大小取决于深度可分离卷积。深度可分离卷积将标准卷积分解为深度卷积和逐点（1×1）卷积，并将深度卷积的输出组合起来，生成新的特征图 [39]。深度可分离卷积神经网络（DS-CNN）是为嵌入式系统实现高性能声学模型的绝佳选择。

综述 [4], 总结指出，现代基于卷积神经网络的声学模型理想情况下应包含以下三个方面：

- 一种利用长时频依赖关系的机制，例如使用时域卷积 [37] 或扩张卷积。
- 深度可分离卷积 [39] 可在不牺牲性能的情况下，大幅降低模型的内存占用和计算量。
- 残差连接能够快速、有效地训练更深的模型，从而提升关键词检测性能。

循环神经网络：

语音是一种具有很强时间依赖性的时间序列。因此，使用循环神经网络（RNN）进行声学建模是很自然的选择。如果延迟不是严格限制，可以使用双向长短期记忆网络（BiLSTM） [40,41] 来捕捉因果关系和反因果关系，从而提升关键词检测的性能。或者看看 KWS 的双向 GRU，地址为 [42]。在关键词检测等场景中，由于存在时间依赖性，建模耗时不长时，GRU 比 LSTM 更好，因为它占用的内存更少。以相近或相当的性能更快地训练。更好 [43]。

卷积神经网络在建模长期依赖关系方面可能存在困难。为了解决这一问题，我们可以将其与循环神经网络结合，构建所谓的卷积循环神经网络。因此，卷积循环神经网络兼具两者的优势。首先，卷积层对语音的局部频谱和时间依赖关系进行建模，然后循环层通过对后续内容进行建模，来刻画语音信号中的长期时间依赖关系。一些研究探讨了

使用单向或双向长短期记忆网络或门控循环单元，在深度语音关键词检出中采用卷积循环神经网络进行声学建模 [40–43]。

Transformer:

Transformer架构近期在多个领域取得了最先进的成果，包括蛋白质序列 [44], 文本 [45,46], 符号音乐 [47], 视频 [48,49] 和图像理解 [50,51]。这体现了一种更广泛的趋势：单一神经网络架构能够跨越多种数据领域和任务进行泛化。注意力机制也已应用于关键词检测 [12, 13], , 但仅作为其他架构（如卷积神经网络或循环神经网络）的扩展。受简单的视觉 Transformer（ViT）模型在计算机视觉领域的出色表现 [50]以及提升其数据效率的技术启发， [52] （如图2.3）提出了将这一架构应用于关键词检测的改进方案，并发现无需额外数据，该方案在 Google Speech Commands 数据集上的表现可媲美或超越现有模型。

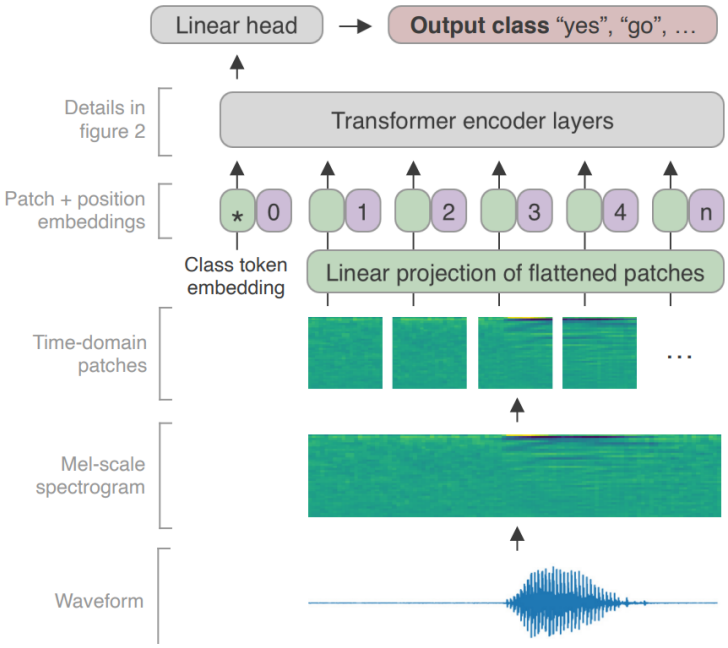


图 2.3： *[52]*提出的关键词 *Transformer*架构。音频经过预处理后转换为梅尔尺度频谱图，并沿时域划分为互不重叠的图块。这些图块与一个学习得到的类别令牌共同构成多层 *Transformer*编码器的输入令牌。与视觉 *Transformer* 一样， *[50]*, 每个令牌都会加上一个学习得到的位置嵌入。类别令牌的输出经过线性分类头处理，用于生成最终的类别预测。

对于用户自定义的关键词，由于无法要求用户提供大量样本，因此没有可用的大型数据集。所以，这可以视为一个少样本学习问题。自监督表征学习（SSL）方法能够分别利用少量和大量的有标签及无标签数据进行微调和预学习，从而发挥作用。SSL 方法有望构建一个通用模型，使其能够惠及各种任务和领域。这类方法已在自然语言处理和计算机视觉领域取得成功，在许多下游场景中提升了性能，同时减少了所需的标签数量。最近， [53] 比较了几种广泛使用的 SSL 模型，以确定哪种预训练模型最适合少样本学习中的关键词检测。他们的结果表明，HuBERT [54] 的结果最佳，并且对少样本样本的变化具有鲁棒性。 [55]将 Wav2Vec 2.0 [56] , 这一 SSL 模型融入其关键词检测模型中。而 [55]达到了最先进的

那一年的艺术表演。

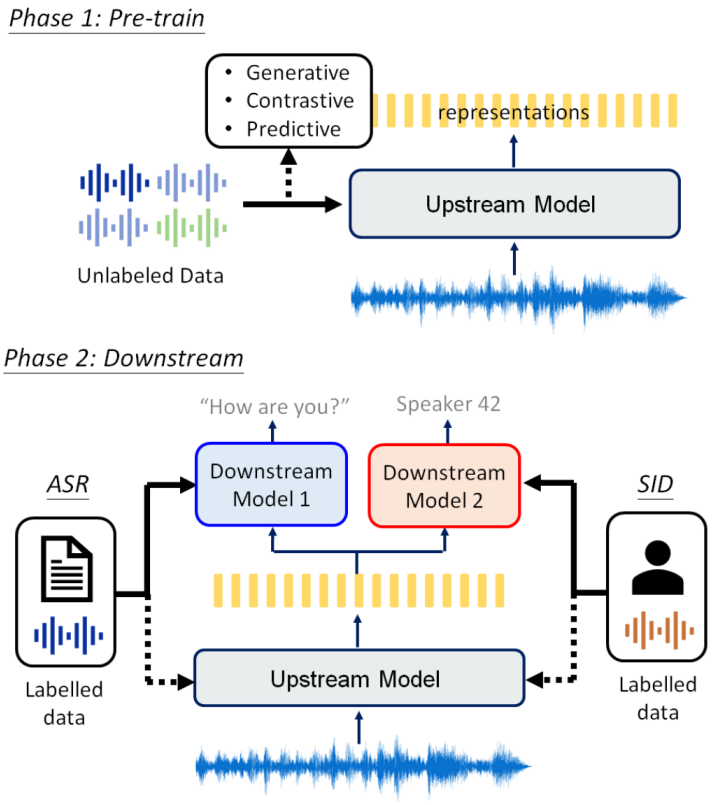


图 2.4：在下游应用中使用自监督表征学习的框架，摘自 *[57]*。

尽管已经取得了诸多进展，我们仍可以预期，自监督学习在关键词检测中的应用在未来会持续成为热门话题。虽然已有一些研究探讨如何利用自监督学习增强关键词检测，但我们仍发现一些尚待解决的问题，将在下一节讨论。如何利用无标签数据和自监督学习技术，更高效地增强关键词检测，是一项重要且具有挑战性的研究任务，也将成为未来工作的一个潜在方向。

2.1.3 关键词检测数据集

Google Speech Commands

公开发布的 Google Speech Commands 数据集 [58]已成为深度关键词检测开发与评估中最著名的开放基准。这个众包数据库以 16 kHz 的采样率，通过手机和笔记本电脑麦克风采集，因此在一定程度上带有噪声。该数据集的第一个版本由 1881 位说话人录制，包含 64727 段时长为一秒或更短的语音片段，每段包含 30 个不同单词中的一个。

第一版与第二版的主要区别在于，后者于2018年公开发布，并增加了5个单词（总计35个单词）、更多语音片段（105829个）和更多说话人（2618位）。图 2.5 列出了 Google Speech Commands 数据集 v1（前六行）和 v2（所有行）包含的单词。为了便于关键词检测技术的复现与比较，该基准还对训练集、开发集和测试集进行了标准化，并统一了实验框架中的其他关键环节，包括涉及

Version 1 (*v1*)	Version 2 (*v2*)						
		yes	no	up	down	left	KW
		right	on	off	stop	go	
		zero	one	two	three	four	Non-KW
		five	six	seven	eight	nine	
		bed	bird	cat	dog	happy	
		house	Marvin	Sheila	tree	wow	
		backward	forward	follow	learn	visual	

图 2.5：以下单词列表来自 *[58]*，包含在 *Google Speech Commands* 数据集 *v1*（前六行）和 *v2*（所有行）中。单词按标准化的 *10* 个关键词（前两行）和非关键词（最后五行）分类。

背景噪声。

基于多模态信息的语音处理（MISP）挑战赛

该数据集 [59] 考虑了以下场景：几个人在客厅边看电视边聊天，并与智能音箱/电视互动。在示意图中，六位说话人正在聊天，同时使用多台设备并行录制音频和视频。在真实客厅中，可能有一些变量会影响对话和/或采集到的音频与视频，例如，电视可能会打开或关闭，对话可能发生在白天或夜间等。此外，通过观察真实客厅中发生的实际对话，我们发现说话人可能分成若干组，分别讨论不同话题。这是一种常见的自然对话现象。与所有说话人讨论同一话题的情况相比，分组讨论会导致音频中的重叠率更高。录制期间，我们对上述变量进行控制，以尽可能覆盖更多真实场景。

Dataset	Training		Dev		Eval	Total
	P	N	P	N		
Duration (h)	5.67	112.86	0.62	2.77	2.87	124.79
Session	89	89	10	10	19	118
Room	25	25	5	5	8	38
Participant	258	258	35	35	54	347
Male	81	81	11	11	31	123
Female	177	177	24	24	23	224

图2.6： *MISP2021-AVWWS* 语料库概览。 *[P*：表示存在唤醒词，*N*：表示不存在唤醒词*]*

使用了三种类型的录音设备。录音设备的类型取决于其与说话者的距离。远距离录音设备是一组由 6 个样本同步的全向麦克风组成的线性麦克风阵列，放置在距离说话者 3-5 米处。相邻麦克风之间的距离为 35 毫米。远距离线性麦克风阵列的录音被录制到笔记本电脑上。在距离说话者 1-1.5 米处，我们放置了一组由 2 个样本同步的全向麦克风组成的线性麦克风阵列。

相邻麦克风之间的距离为 92毫米。为便于转录，每位说话者都在下巴中央佩戴了一个高保真麦克风。中间线性麦克风阵列和每个近距离高保真麦克风的音频均通过调音台录制。

任务 1 使用的数据库包含 124.79小时的视听数据。图 2.6 展示了视听数据在训练集、开发集和评估集中的划分，并列出了会话数量、房间类型以及男性/女性说话人的数量等详细信息。唤醒词为“Xiao T Xiao T”。该数据集包含 118 个会话。每个对话会话中的说话人数量从 1 到 6 不等。数据集中说话人总数为 347 人。所有说话人均为以普通话为母语的中国人，且没有明显口音。录音期间推荐了各种对话主题。由于最终排名仅取决于远场录音的结果，评估集仅包含来自远场设备的录音，而中场和近场录音则包含在训练集和开发集中。此外，还提供了一些真实噪声数据。

2.2 环境声音分类

2.2.1 定义与背景

环境声音分类（ESC）旨在将音频录音归入预定义的环境声音类别 [60]。检测任务的目标声音因应用场景而异；但对于通用声音事件检测系统，目标声音是鸟鸣、汽车驶过和脚步声等环境声音。在文献中，为了将环境声音分析领域与更成熟的语音或音乐分析任务区分开来，这些声音有时被称为 非语音 和 非音乐声音 [61], 。声音事件检测任务的目的也不同于典型的语音或音乐分析任务，因为人们对语音、音乐和环境声音的感知方式也不同：聆听音乐侧重于声音的审美特质，语音感知侧重于语言或副语言信息，而日常聆听则旨在识别声源 [62]。

近年来，设备端环境声音分类 [17, 63, 64] 日益受到研究关注，正如声学场景与事件检测和分类（DCASE）2022年挑战赛任务1“低复杂度声学场景分类”所体现的那样 [65]。这种计算复杂度较低的声音分类系统可部署在移动平台和嵌入式平台上，用于许多实际音频应用，例如声学监控 [5], 、生物声学监测 [66] 和多媒体索引 [7]。

可能的声音类别数量是无限的，因为任何物体或生物都可能作为自然发生的事件产生声音。

2.2.2 环境声音分类方法

近年来的研究表明，深度神经网络在图像分类和语音识别等多个领域带来了巨大进步，如今也已成为环境声音分析与分类的主流方法 [67, 68]。其主要缺点是需要大量数据进行训练。由于该领域仍然缺乏大型强标注数据集，因此大型数据集的需求给声音事件检测带来了难题。采用弱标签和迁移学习的先进训练策略为解决数据方面的不足提供了可行方案，但系统的整体架构通常不会发生显著变化。

卷积神经网络：

传统的卷积神经网络架构会使用多个连续的卷积和池化操作模块，分别沿时间维度和特征维度进行特征学习与下采样。作为替代方案，Ren 等人使用了基于扩张卷积核的空洞卷积神经网络 [69]。这类卷积核无需进行中间池化操作，也能获得相当的感受野大小。Koutine 等人表明，如果通过限制感受野大小对其进行正则化，ESC 系统的性能可以得到提升 [70]。

在大多数基于卷积神经网络的架构中，只有最后一个卷积层的激活值会连接到最终分类层。作为一种替代方案，Yang 等人采用了多尺度特征方法，并进一步处理了来自所有中间特征图的激活值 [71]。此外，作者使用了 Xception 网络架构，其中卷积操作被拆分为深度（空间）卷积和逐点（通道）卷积，以减少可训练参数的数量。另一种相关方法是将二维卷积分解为两个一维卷积核，以分别建模声音的瞬态特征和长期特征 [72]。不同对称性的影响

非对称卷积核形状也由Wang等人进行了系统评估。 [73]。

为改善特征学习，人们提出了对常见卷积神经网络架构的若干扩展。Basbug 和 Sert 借鉴了计算机视觉中的空间金字塔池化策略，在该策略中，特征图经过池化后，在不同的空间分辨率下进行组合 [74]。Marchi等人将基于频谱图特征的一阶和二阶时间导数作为额外的输入通道，以便检测幅度快速增大的瞬态短时事件 [75]。

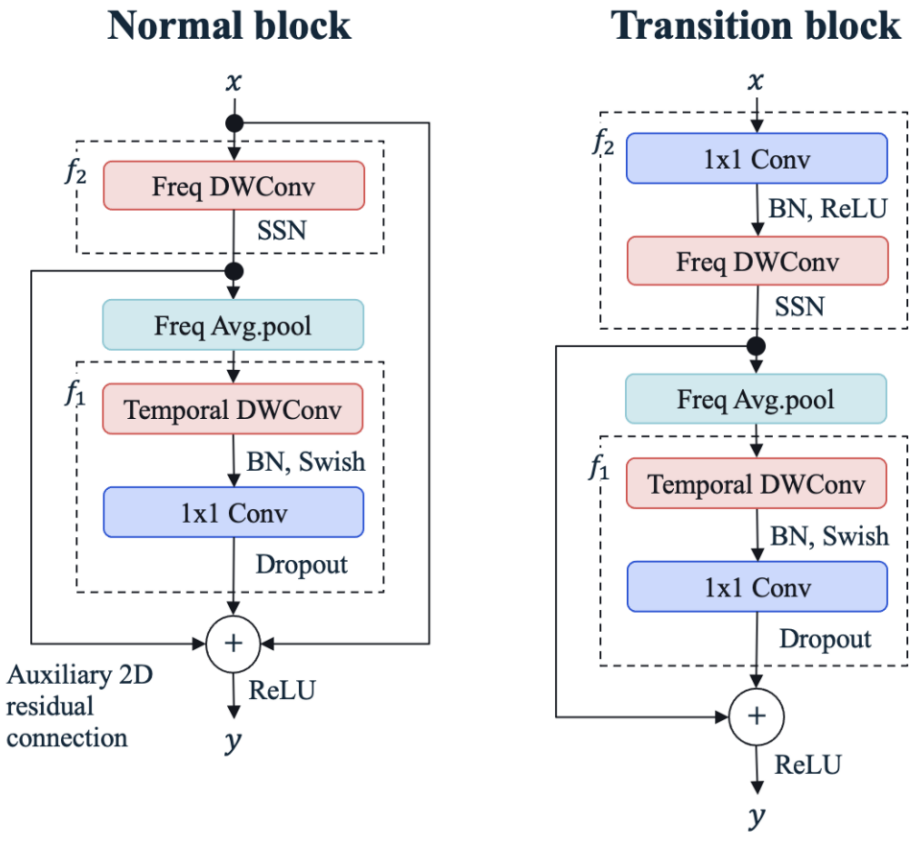


图 2.7： ：来自 *[24]* 的 *BCResBlock*。*BC-ResNet* 块包含一个带有子频谱归一化的频率深度卷积。随后，对特征沿频率维度求平均，再进行时间深度可分离卷积。在残差连接处，时间特征被广播到二维特征中。在过渡块中，前端还额外添加了一个 *1x1* 卷积，以改变通道数，且不使用恒等捷径连接。

近年来，已有许多研究提出了更高效、性能更优的环境声音分类方法。Kim等人提出了相关方法 [25]，旨在利用未见设备的泛化能力，同时保持轻量级模型的性能。为了设计参数数量较少的低复杂度网络，他们采用了广播残差网络（BC-ResNet） [24], 作为基线架构，将一维和二维卷积神经网络特征结合使用，以提高效率。虽然 [24] 针对的是人声，但我们的目标是对音频场景进行分类。因此，我们对网络进行了两项修改，即限制感受野，并使用最大池化代替空洞卷积，以适应输入域的差异。所提出的架构如图2.8所示，是一种名为改进型 BC-ResNet（ BC-ResNet-Mod）的全卷积网络。该模型的前端采用 5x5 卷积，步幅为 2x2，以进行下采样，后接 BC-ResBlocks [24]。模型共包含 9 个 BC-ResBlocks，如2.7所示，并设有两个最大池化层，感受野大小为 109x109。他们还在全局平均池化之前加入最后一个 1x1 卷积，使模型能够分别对每个感受野进行分类，并通过求平均将分类结果集成起来。

Input	Operator	n	Channels
$256 \times T \times 1$	conv2d 5x5, stride 2	-	2c
$128 \times T/2 \times 2c$	stage1: BC-ResBlock	2	c
$128 \times T/2 \times c$	max-pool 2x2	-	-
$64 \times T/4 \times c$	stage2: BC-ResBlock	2	1.5c
$64 \times T/4 \times 1.5c$	max-pool 2x2	-	-
$32 \times T/8 \times 1.5c$	stage3: BC-ResBlock	2	2c
$32 \times T/8 \times 2c$	stage4: BC-ResBlock	3	2.5c
$32 \times T/8 \times 2.5c$	conv2d 1x1	-	num class
$32 \times T/8 \times$ num class	avgpool	-	-
$1 \times 1 \times$ num class	-	-	-

图 2.8：：来自 的 *BC-ResNet-Mod*。每一行都是一个由一个或多个相同模块组成的序列，重复 n 次；输入形状为频率 × 时间 × 通道，总时间步数为 T。

BC-ResNets 在时间维度上使用空洞卷积，以在保持网络时间分辨率的同时获得更大的感受野。他们发现，在音频场景领域，无需完全保留时间分辨率，因此在网络中部插入最大池化层，而不使用空洞卷积。

BC-ResNet-Mod 实现了两个目标：1）在参数数量方面实现高效设计；2）适应设备不均衡的数据集。因此，在我们的实验中，我们使用 BC-ResNet-Mod-4，它会在提取频谱和时间特征之前，将输入通道维度增加到 80。

前馈神经网络：

前馈神经网络（FNN）用于多种环境声音分类算法。Bisot 等人使用 FNN 架构，将非负矩阵分解和音频信号的常数 Q 变换所提取的特征进行拼接 [76]。Takahashi 等人将 FNN 与多个高斯混合模型（GMM）分类器结合起来，以建模各个声学场景 [77]。

卷积循环神经网络：

用于声音事件检测的一种通用网络架构是卷积循环神经网络（CRNN），它包含卷积层和循环层，各自承担特定的作用 [78]。卷积层充当特征提取器，旨在通过连续的卷积和非线性变换，从网络输入端提供的时频表示中学习判别性特征。循环层则负责学习输入特征序列中的时间依赖关系。

Transformer：

为了更好地捕捉长程全局上下文，近期的一种趋势是在卷积神经网络之上添加自注意力机制。这类卷积神经网络-注意力混合模型在许多音频分类任务中都取得了最先进水平（SOTA）的结果，例如音频事件分类 [79, 80], 口语关键词检测 [42],和情感识别 [81]。然而，鉴于纯基于注意力的模型在视觉领域的成功 [50,51], ，我们有理由探讨卷积神经网络对于音频分类是否仍然不可或缺。

Gong等人提出了音频频谱图 Transformer（AST） [82], 这是一种无卷积、纯粹基于注意力的模型，可直接应用于音频频谱图，并且即使在最低层也能捕获长程全局上下文。AST（图 2.9）在各种

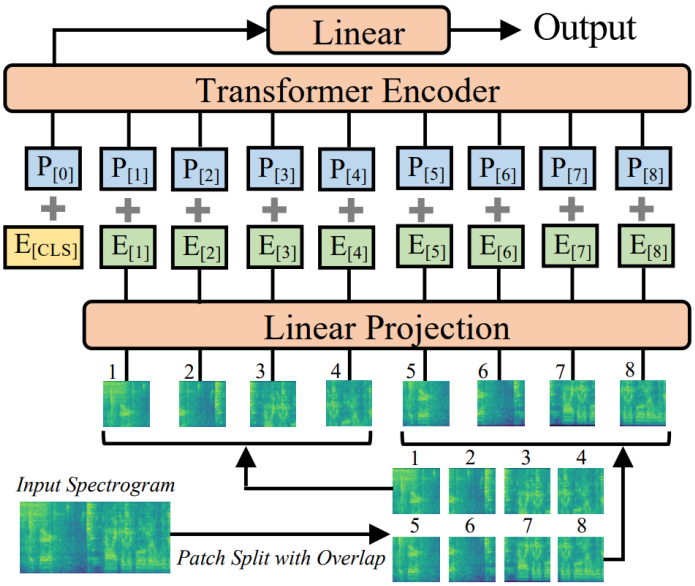


图2.9： 音频频谱图 *Transformer*（*AST*）架构，源自 *[82]*。二维音频频谱图被拆分为一系列 *16×16* 有重叠的图块，然后线性投影为一系列一维图块嵌入。每个图块嵌入都加上一个可学习的位置嵌入。序列前置一个额外的分类令牌。输出嵌入作为 *Transformer* 的输入，分类令牌的输出则通过线性层用于分类

音频分类任务和数据集上均优于最先进的系统。

尽管标注音频和语音数据成本高昂，他们仍探索了自监督 AST（SSAST），利用未标注数据缓解对数据量的需求问题。在 [83], Gong等人提出了一种新颖的联合判别式与生成式掩码频谱图图块建模（MSPM）自监督学习（SSL）框架，能够在标注数据有限的情况下显著提升 AST 的性能。据我们所知，MSPM 是音频和语音领域首个基于图块的自监督学习框架，而 SSAST 是首个纯粹基于自注意力的自监督音频分类模型。他们还表明，同时使用语音和音频数据集进行预训练，能够显著提升模型的泛化能力，并且其性能优于使用单一领域数据集进行预训练。因此，SSAST 模型在语音和音频下游任务上均表现出色。

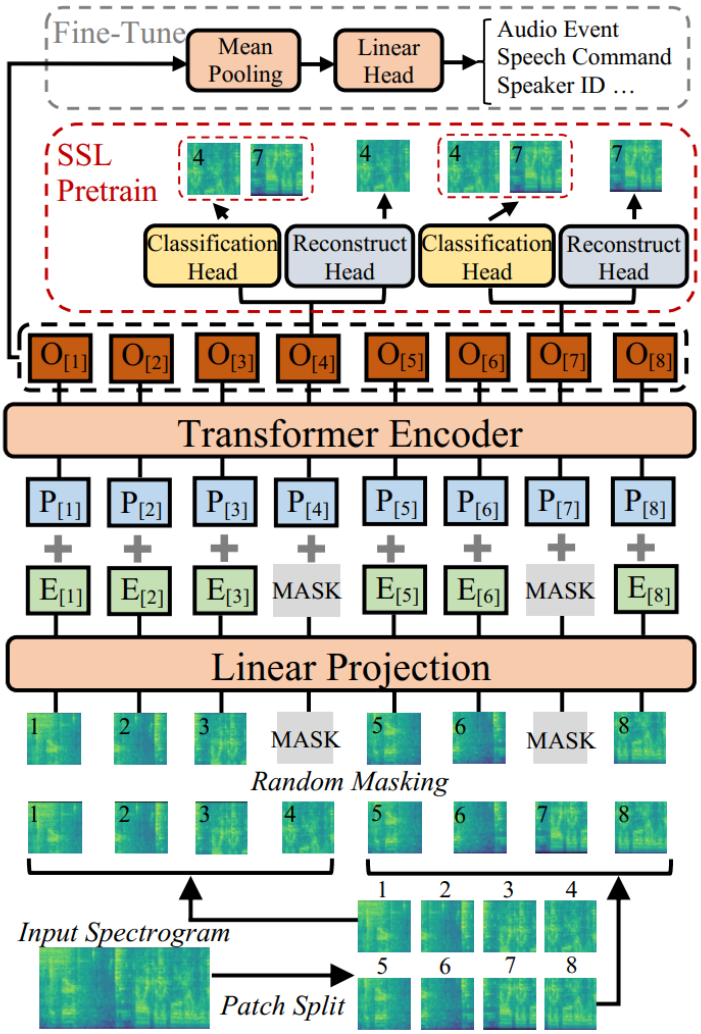


图2.10：来自 *[83]* 的自监督 *AST*。二维音频频谱图被拆分为一系列互不重叠的 *16×16* 图块，然后通过线性投影转换为一系列一维图块嵌入 *E*。每个图块嵌入都与可学习的位置嵌入 *P* 相加，然后输入 *Transformer*编码器。*Transformer* 的输出 *O* 被用作频谱图图块表征。在自监督预训练期间，他们会随机屏蔽一部分频谱图图块，并要求模型：*1*）从所有被屏蔽的图块中，找出每个被屏蔽位置对应的正确图块；*2*）重建被屏蔽的图块。这两个前置任务旨在促使 *AST* 模型学习音频数据的时间结构和频率结构。在微调期间，他们对所有图块表征进行均值池化{*O*}，并使用线性分类头进行分类。

2.2.3 声音分类数据集

音频集

音频集 [84] 是一个大规模数据集，包含经过人工标注的音频事件，旨在弥合图像与音频研究在数据可用性方面的差距。我们依据文献和人工整理结果，构建了一个包含 632 个音频类别、精心设计的层级本体，并收集人工标注员对 YouTube 视频中 10 秒片段是否存在特定音频类别的判断。我们根据元数据、上下文（例如链接）和内容分析进行搜索，选取片段供标注。最终的数据集包含 1,789,621 个片段（4,971 小时），涵盖 485 个音频事件类别，每个类别至少有 100 个实例。

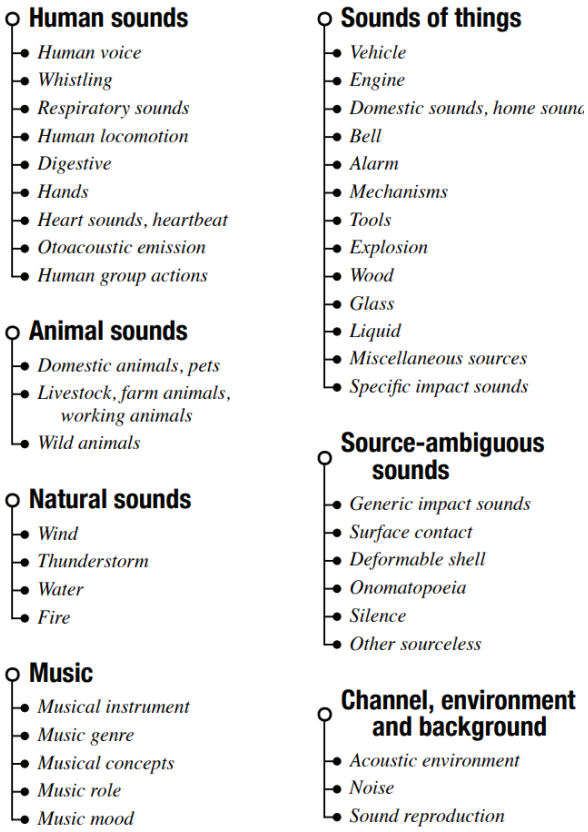


图2.11： *[84]*中的类别

TAU 2019 城市声学场景

TAU 2019 城市声学场景数据集 [22], 由以下声学场景的录音组成：机场、室内购物中心、地铁站、步行街、公共广场、中等交通流量街道、乘坐有轨电车、乘坐公交车、乘坐地下地铁以及城市公园。

该任务所使用的数据集是 TUT 2018 城市声学场景数据集的扩展版，录制于欧洲多个城市。TUT 2018 城市声学场景数据集包含来自巴塞罗那、赫尔辛基、伦敦、巴黎、斯德哥尔摩和维也纳的录音；TAU 2019 城市声学场景数据集则在此基础上增加了里斯本、阿姆斯特丹、里昂、马德里、米兰和布拉格。录音由四台设备同时录制。

ESC-50

ESC－50 [23] 数据集由 2,000 条带标签的环境录音组成，平均分为 50 个类别（每个类别包含 40 个音频片段）。为方便起见，这些类别被归入 5 个宽泛定义的主要类别中（每个类别包含 10 个类别）：

- 动物声音，
- 自然声景和水声，
- 人类非语音声音，
- 室内／家庭声音，
- 室外／城市噪声。

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

图2.12： *[23]*中的类别

该数据集涵盖了各种声源——其中一些非常常见（如笑声、猫叫声、狗吠声），一些相当独特（如玻璃破碎声、刷牙声），还有一些声源之间的差异更加细微（如直升机噪声和飞机噪声）。

2.3 持续学习

2.3.1 定义与背景

持续学习旨在开发能够利用新数据持续学习、解决新任务，同时保留从先前任务中学到的知识的人工智能系统 [20, 85]。在大多数持续学习（CL）场景中，学习者会在一系列示例训练会话中依次接触各项任务。在此期间，训练仅使用单个任务的数据。每次训练后，学习者都应能够在未见过的数据上完成所有先前接触过的任务。这种学习模型的生物学启发十分明显，因为它反映了人类获取并整合新知识的方式。当面对新的学习任务时，学习者会利用已有知识，并将新学到的知识整合到先前任务中。

这与常见的监督学习范式形成鲜明对比。在这些范式中，所有任务的带标签数据都会在一次深度网络训练过程中共同提供。持续学习者在评估所有先前学习过的任务时，一次只能访问其中一个任务的数据。持续学习的主要挑战是学习当前任务的数据，同时避免遗忘先前学过的任务。朴素的微调方法在领域迁移问题上非常有效，但由于缺少先前任务的数据，最终得到的分类器无法对该任务的数据进行分类。在先前学过的任务上，性能大幅下降的现象称为灾难性遗忘 [11]。持续学习旨在防止灾难性遗忘，同时避免僵化 问题，因为僵化会妨碍模型适应新任务。

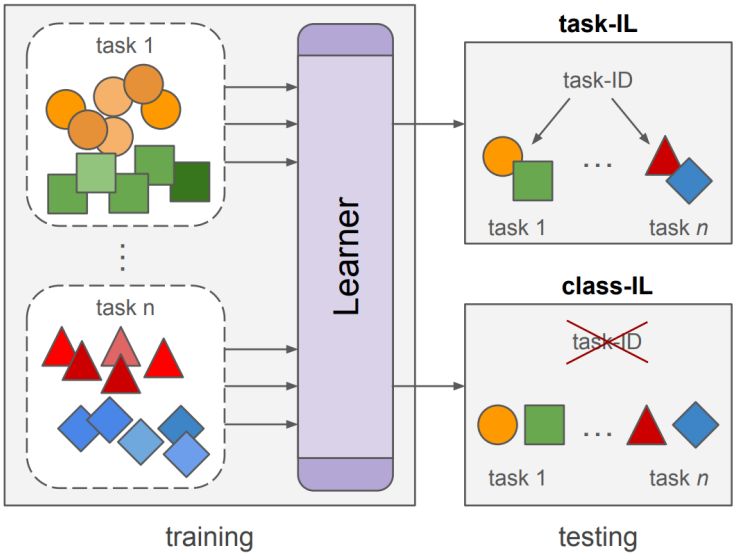


图2.1： 在持续学习中，互斥任务按顺序学习。任务增量学习在评估时可以获取任务 *ID*，而更具挑战性的类别增量学习则无法获取。类别增量学习是本学位论文的研究主题。

持续学习将训练划分为一系列任务。在任何一次训练过程中，学习者只能访问当前任务的数据。某些方法还可以选择利用少量从先前任务中保存的数据。大多数早期方法都考虑了这种情况，称为任务增量学习（任务 IL）：算法在推理时可以访问任务 ID。这显然有一个优势：方法无需区分来自不同任务的类别。近年来，研究者开始着手解决难度更高的类别增量学习（类别-IL）场景。在这类场景中，学习者在推理时无法访问任务 ID，因此必须能够区分所有类别（见图2.13）。推理时不提供任务 ID 的场景通常

包括逐渐提高类别粒度的场景（例如，在图像中检测越来越多的目标类别）。过去几年提出了多种类别-IL 技术，本文也重点关注持续学习中的类别-IL 技术。

过去几年中，持续学习的普及程度显著提升，这主要是由工业和社会应用的需求所推动的。逐步吸收知识可以解决一些问题：

- **内存限制：** 某些系统在可存储的数据量方面存在物理限制，无法存储其能够显示的全部数据，因此不能依赖联合训练策略。这类系统只能存储用于执行任务的有限样本集，并且必须以增量方式进行学习。
- **数据安全／隐私限制：** 对于需要从不应永久存储的数据中学习的系统，持续学习可以提供一种解决方案。政府法律可能会限制客户将数据存储在中央位置（例如，用于移动设备上的应用程序）。

- **可持续人工智能：** 训练深度学习算法可能成本高昂。每次新的数据更新都重新训练此类系统，会产生巨大的碳足迹，而且这一问题在未来几年很可能会日益严重。持续学习提供了计算方面高效的算法，仅需在系统更新时处理新数据。

2.3.2 持续学习方法

本节介绍几种用于应对上述持续学习挑战的方法，并将其分为三大类：基于正则化的方法、基于回放的方法和偏差校正方法。

正则化方法

有些方法会在分类损失中加入正则化项，以减轻灾难性遗忘。有些方法对权重进行正则化，并估计网络中每个参数的重要性度量 [86], ，另一些方法则着重于记住特征表示的重要性 [87]。

回放方法

回放方法会保留少量样本（样本回放），或生成合成样本。通过回放先前任务中存储或生成的数据，回放方法旨在防止遗忘先前任务。大多数回放方法会结合使用样本来应对任务间混淆，并采用其他方法处理导致灾难性遗忘的原因。将样本回放用于持续学习，最早由《增量分类器与表示学习》（iCaRL）提出 [20]。此后，这一技术已应用于大多数持续学习方法。在接下来的两个章节中，我们将重点讨论应用样本时需要作出的选择。

偏差校正方法

偏差校正方法旨在解决任务近因偏差问题，即增量学习的网络倾向于偏向最近学习任务中的类别。造成这一问题的主要原因是，在训练结束时，网络见过许多最后一个任务中的类别样本，却没有见过（或在使用回放时只见过极少量）早期任务中的样本。Wu 等人提出了一种简单有效的方法来防止任务近因偏差 [88], ，并将其方法称为偏差校正（BiC）。他们在网络中添加了一个专门用于校正任务偏差的额外层。训练分为两个阶段。第一阶段使用交叉熵损失训练新任务。随后，在第二阶段，他们从训练数据中拆分出一小部分，作为验证集。

第3章

口语关键词检测的持续学习

本章提出一种名为 Rainbow Keywords（RK）的新型多样性感知持续学习方法，以解决上述问题；该方法无需任务 ID 信息，且所需参数更少。具体而言，所提出的 RK 方法通过计算分类不确定性，引入多样性感知采样器，从历史关键词和新输入关键词中选取少量但多样的样本。因此，即使只能利用有限的历史样本，模型在学习新关键词时也不会遗忘先验知识。此外，我们采用混合标签数据增强，进一步提升所选样本的多样性，从而提高性能。同时，我们提出一种知识蒸馏损失函数，以确保先验知识能够通过有限的所选样本得以保留。我们按照先前工作的设置，在*Google Speech Command* 数据集上开展实验 [13, 18]。实验结果表明，与最佳基线相比，所提出的 RK 方法在平均准确率方面绝对提升了 4.2%，且所需内存更少。相关脚本可在 GitHub 上获取 [1]。

3.1 相关工作

以往的工作 [89, 90] 利用少样本微调，使用来自目标域的训练数据调整关键词检测模型，以适应新的场景。然而，调整后模型在源域数据上的性能可能较差，这也称为灾难性遗忘</style问题 [11]。近期工作 [10]提出了一种用于小型关键词检出的渐进式持续学习 [91]策略，以缓解使用目标域数据调整基于源域数据训练的模型时出现的灾难性遗忘问题。这种方法有两方面的局限。首先，该方法需要将任务 ID 作为辅助信息来学习不同任务的知识，而在实践中这一信息并非总是可用。其次，随着任务数量增加，模型占用的存储空间也会增大。对于轻量级边缘设备而言，这样的存储空间开销难以承受。

3.2 方法

对于在线关键词检出系统，我们假设模型应能在一系列任务中识别所有关键词，且不会发生灾难性遗忘。对于每个任务 τ_t，我们都有输入对 (x_t, y_t)，其中 x_t 表示音频话语，y_t 表示关键词标签。我们的目标是最小化截至当前任务 τ_t 的所有关键词N^t 的交叉熵损失 [85]，其定义如下：

$$L_{CE}(x,y) = \sum_{i=1}^{N^t} y_i log \frac{exp(o_i)}{\sum_{j=1}^{N^t} exp(o_j)} \tag{3.1}$$

[1]https://github.com/swagshaw/Rainbow-Keywords

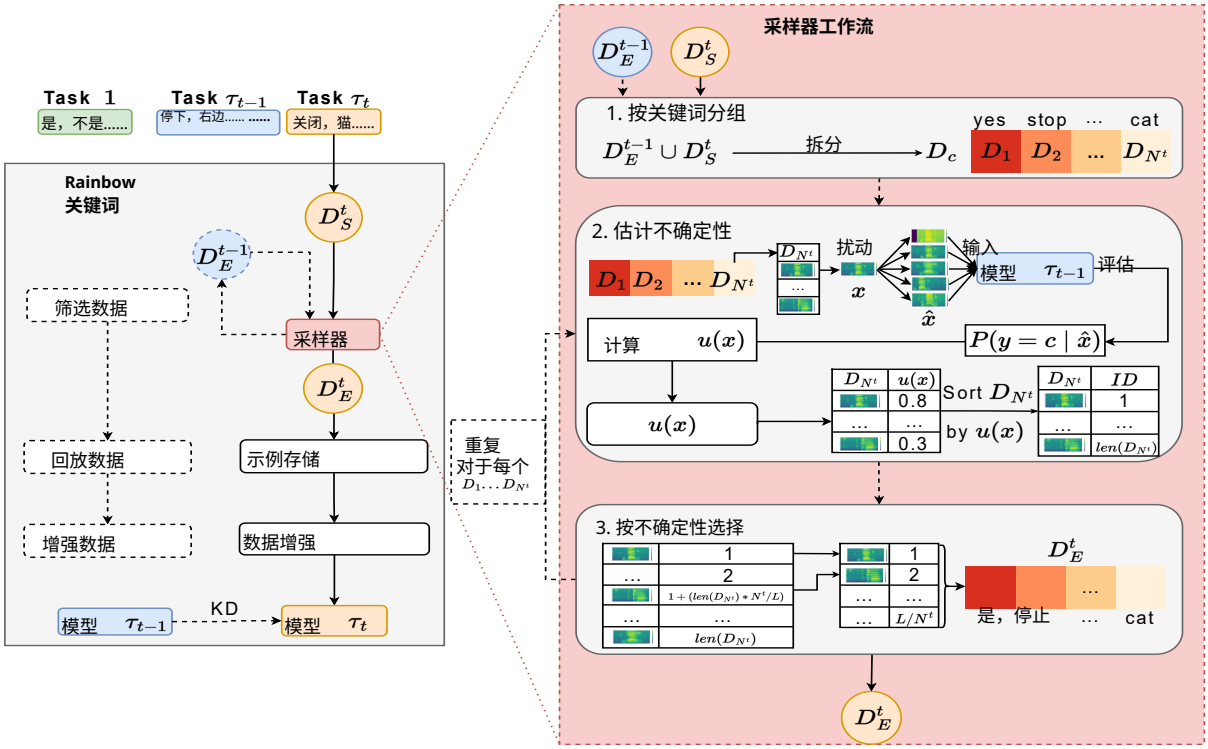


图 *3.1*：所提出的彩虹关键词方法的框图。具体而言，D_S^t 表示任务 τ_t 的输入音频流数据。D_E^t 和 D_E^{t-1}分别表示任务 τ_t 和 τ_{t-1} 的样本。我们根据唯一关键词将 $D_E^{t-1} \cup D_S^t$ 分组成子集 D_c,$c = 1..N^t$，其中 N^t 表示 $D_E^{t-1} \cup D_S^t$ 集中唯一关键词的总数。x、$\hat{x}$ 和 K 分别表示 D_c中的每个样本、x 的五种扰动以及五种扰动策略。“计算 $u(x)$”表示根据公式 *3.3* 计算 $u(x)$。

其中 o 表示模型在任务 τ_t中的输出逻辑值。

3.2.1 多样性感知采样器

多样性感知采样器旨在选择 多样化样本 以高效管理内存。此类多样化样本由各样本在特征空间中的相对位置定义，该位置通过分类模型推理样本时的不确定性进行估计 [21]。所需的三个步骤如图3.1右侧红框所示：

按关键词拆分： 第一步汇集历史样本 D_E^{t-1}和输入数据 D_S^t，并按唯一关键词将它们分组为子集 D_c,$c = 1..N^t$，其中 N^t 表示$D_E^{t-1} \cup D_S^t$ 集中唯一关键词的总数。

估计不确定性： 第二步使用蒙特卡洛（MC）方法估计 D_c 中每个样本 [92], 的不确定性，其定义见公式 3.2。

$$P(y = c \mid x) = \int p(y = c \mid \hat{x}) p(\hat{x} \mid x) d\hat{x} \tag{3.2}$$

其中x、$\hat{x}$、y 分别表示关键词D_{N^t}的每个音频话语、x的五种扰动，以及x的标签。因此，音频话语x的不确定性表示为$u(x)$：

$$u(x) \approx 1 - \frac{1}{K} \sum_{k=1}^{K} P(y = c \mid \hat{x}_k) \tag{3.3}$$

算法 1：多样性感知记忆更新。输入：L 表示内存大小，N^t 表示任务 τ_t 之前已见类别的数量，D_S^t 表示任务 τ_t 的输入数据，D_E^{t-1}表示任务 τ_{t-1} 之后的历史样本。输出：学习任务 τ_t 后的 D_E^t 个样本。1 初始化 $D_E^t = \{\}$；2 k_c = floor(L/N^t)；3 对类别 c = 1,2, ..., N^t 执行 4 $D_c = \{(x, y) | y = c, (x, y) \in D_S^t \cup D_E^{t-1}\}$；5 按照由公式 3.3 计算的 u(x) 对 D_c 排序。6 对 j = 1,2, ..., k_c 执行 7 i = j∗ $|D_c|/k_c$；8 $D_E^t += D_c$[i]；9 结束；10 结束

其中K 表示五种扰动策略，包括削波失真 [93], 时间掩蔽 [93],移位 [94], 音高偏移 [94] 和频率掩蔽 [93]。$u(x)$越大，表明模型对预测这些扰动的置信度越低。

按不确定性选择： 第三步从D_c 中按不确定性$u(x)$降序排列，并以$len(D_c) * N^t/L$为步长，选取L个样本。因此，最多样化的样本会被纳入D_E^t。只有这些样本可用于训练。

算法 1 总结了我们提出的多样性感知记忆更新算法。我们将内存大小设为 L，以确定设备中最多可保留的样本数。我们在已见类别（N^t）之间分配相同数量的记忆槽（k_c）。将样本分配至记忆槽后，我们计算任务 $\tau_t D_S^t$）中的流式样本和历史记忆样本（D_E^{t-1}）的不确定性，然后按不确定性对所有这些样本（D_c）排序。从排序后的列表中，我们按间隔 $|D_c|/k_c$ 选取样本，以确保多样性。

3.2.2 数据增强

由于D_E^t 中的样本受内存限制而数量较少，我们采用数据增强来进一步增加D_E^t的多样性。具体而言，我们随机混合两个音频话语，在不增加额外存储的情况下增加训练数据量。

3.2.3 知识蒸馏损失

近期研究 [20, 88, 95] 表明，知识蒸馏（KD）有助于在教师—学生模型之间传递知识。受这一理论启发，我们将任务 τ_{t-1} 的模型视为教师模型，将任务 τ_t 的模型视为学生模型。我们提出一种知识蒸馏损失，使学生模型能够保留教师模型的先验知识，从而避免灾难性遗忘，其形式如下：

$$\begin{aligned} \sigma(o_i^t(x); N^{t-1}) &= \frac{exp(o_i^t/T)}{\sum_{j=1}^{N^{t-1}} exp(o_j^t/T)} \\ L_{KD}(o^t(x), o^{t-1}(x)) &= \sum_{i=1}^{N^{t-1}} \sigma(o_i^t(x)) log \sigma(o_i^{t-1}(x)) \end{aligned} \tag{3.4}$$

其中，$o^{t-1}(x)$ 和 $o^t(x)$ 分别表示教师模型和学生模型的输出逻辑值。N^t表示截至任务 τ_t的所有关键词。 σ 表示由温度T参数化的知识蒸馏 Softmax函数。温度T 是知识蒸馏的经验超参数，设为 2.0。因此，我们旨在最小化截至当前任务 τ_t 的所有关键词N^t 的总损失，其形式如下：

$$L_{total}(x,y) = \lambda L_{CE}(x,y) + \quad (1-\lambda) L_{KD}(o^t(x), o^{t-1}(x)) \tag{3.5}$$

其中L_{CE} 表示公式3.1中定义的交叉熵损失，L_{KD} 表示上文定义的知识蒸馏损失。 λ 表示定义为 $\sqrt{1 - N^{t-1}/N^t}$ 的实验超参数。

3.3 实验与结果

3.3.1 数据集

我们在 *GoogleSpeech Command* 数据集 v1（GSC） [58], 上进行实验，该数据集包含 64,727 个时长为一秒的音频片段，涵盖 30 个英语关键词类别。我们使用 80% 的数据进行训练，20% 的数据进行测试。在我们的实验中，GSC 中的所有音频片段均以 16kHz 进行采样。

3.3.2 实验设置

网络配置

我们采用 TC-ResNet-8 [37] 作为测试平台，以评估所提出的彩虹关键词方法。它包含一个一维卷积层，后接三个残差块，每个残差块由一维卷积、批归一化和 ReLU 激活函数组成。每层有 {16,24,32,48} 个通道（如图 3.2所示）。

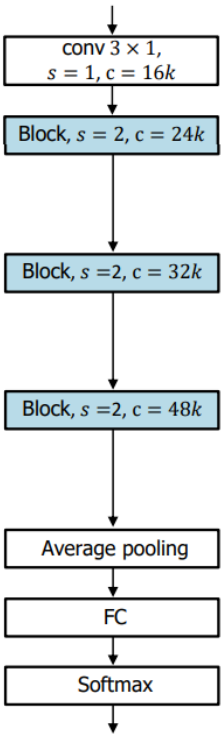


图 3.2：*TC-ResNet-8* 的架构。它分别使用 *ResNet-8* 作为骨干卷积神经网络。*FC* 表示全连接层。请注意，*'s'*、*'c'* 和 *'k'* 分别表示步幅、通道大小和宽度乘数。

我们首先使用包含 15 个唯一关键词的 GSC 数据集中的 32,000 个音频片段对 TC-ResNet-8 模型进行预训练。为了评估所提出的 RK 方法的学习能力，我们将其余数据拆分为 5 个任务。每个任务包含 3 个新的唯一关键词，这些关键词在先前任务中未曾出现。为模拟边缘设备的运行条件，我们将样本的最大数量设为 L，因为边缘设备的内存有限 [13, 18]。

在 训练阶段，我们使用梅尔频率倒谱系数（MFCC = 40）作为输入。

参考基线

我们构建了八种基线方法进行比较。

- **微调训练**：针对每个新任务调整 TC-ResNet-8 模型，不采用类别增量学习（CIL）策略。我们将其视为下界基线。
- **NR [96]：** 这是一种类别增量学习方法，会从先前任务中随机选择训练样本，用于后续训练。
- **iCaRL [20]：** 是一种类增量学习方法，会选择靠近其所属类别均值的样本。随后，iCaRL 会利用这些样本进行后续训练。
- **EWC [86]：** 是一种类增量学习方法，通过引入二次惩罚项，对模型中对过去任务至关重要的参数进行正则化。参数的重要性通过费 Fisher 信息矩阵近似计算。
- **RWalk [97]：** 是一种改进 EWC 的类增量学习方法。费 Fisher 信息矩阵近似 [86] 和在线路径积分 [98]相结合，用于计算每个参数的重要性。RWalk 还会选择历史样本，并利用它们进行后续训练。
- **BiC [88]：** 是一种近期提出、受到更多关注的类增量学习方法。BiC 引入了一个额外的层，用于校正网络的任务偏差。BiC 还采用与 iCaRL 相同的采样方法来选择历史样本。
- **PCL-KWS [10]：** 是一种用于口语关键词检测的持续学习方法。具体而言，PCL-KWS 包含多个任务专用子网络，用于记忆先前关键词的知识。随后，该方法引入关键词感知网络缩放机制，以减少网络参数。PCL-KWS 需要 task-ID 来选择相应的子网络。
- **联合训练：** 使用整个数据集训练 TC-ResNet-8 模型，不受任何限制。我们将其视为上限基线。

指标

我们根据准确率和效率指标报告性能。准确率指标包括平均准确率 (ACC)和向后迁移 (BWT) [99]。具体而言，“平均准确率”表示整个训练结束后所有已学习任务的平均准确率。“BWT”评估学习新任务后所有先前任务的准确率变化，反映遗忘程度。效率指标包括参数和内存。“参数”衡量策略中模型的参数总量。“内存”表示每个任务中全部训练数据所需的内存。

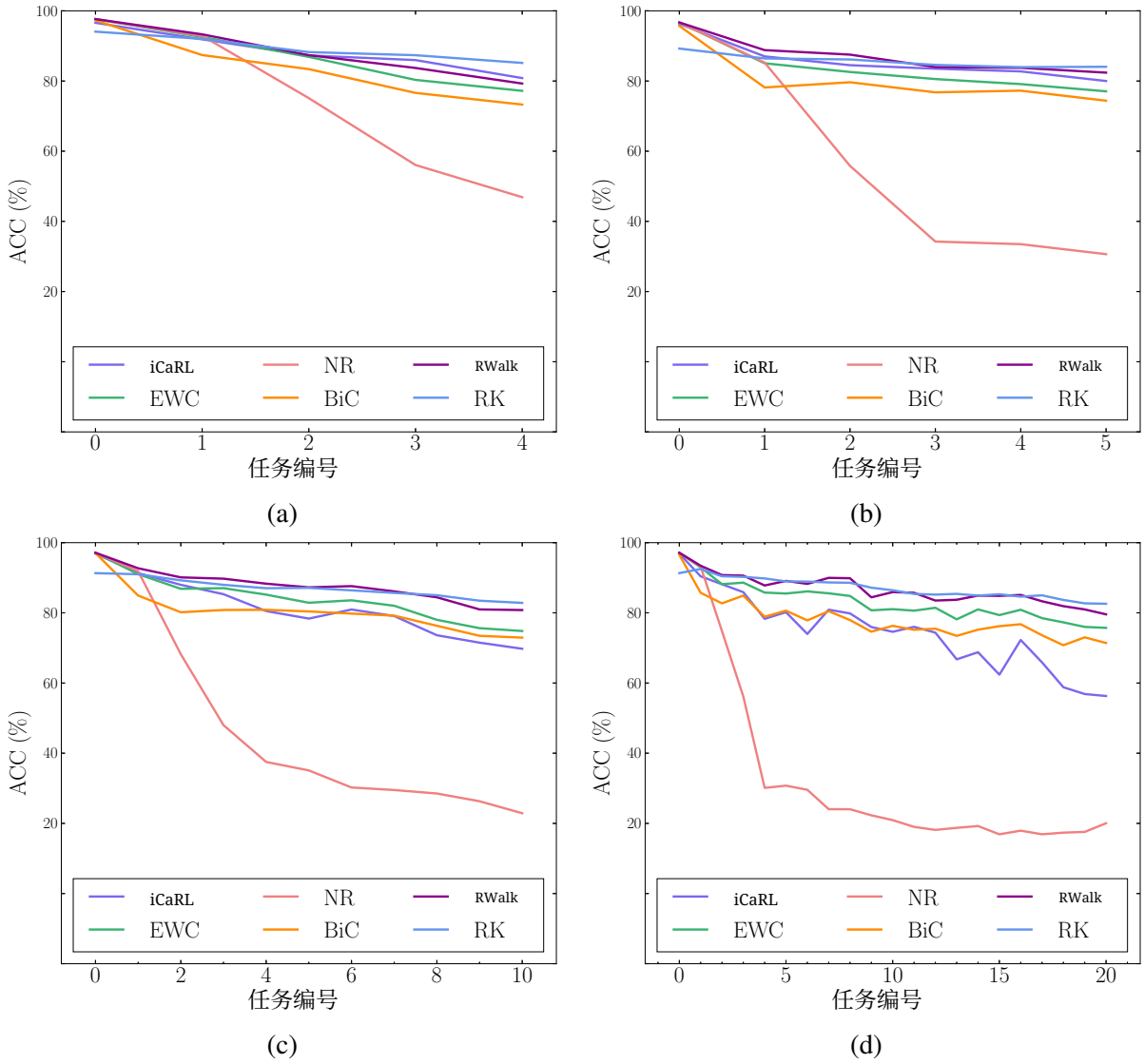


图 3.3： 在对所提出的 *RK* 方法与其他有竞争力的基线方法进行的任务数量递增对比研究中，准确率（*%*）。图 *(a)* 至 *(d)* 分别表示任务数量为（*=20*、*10*、*5*、*4*）时的实验。

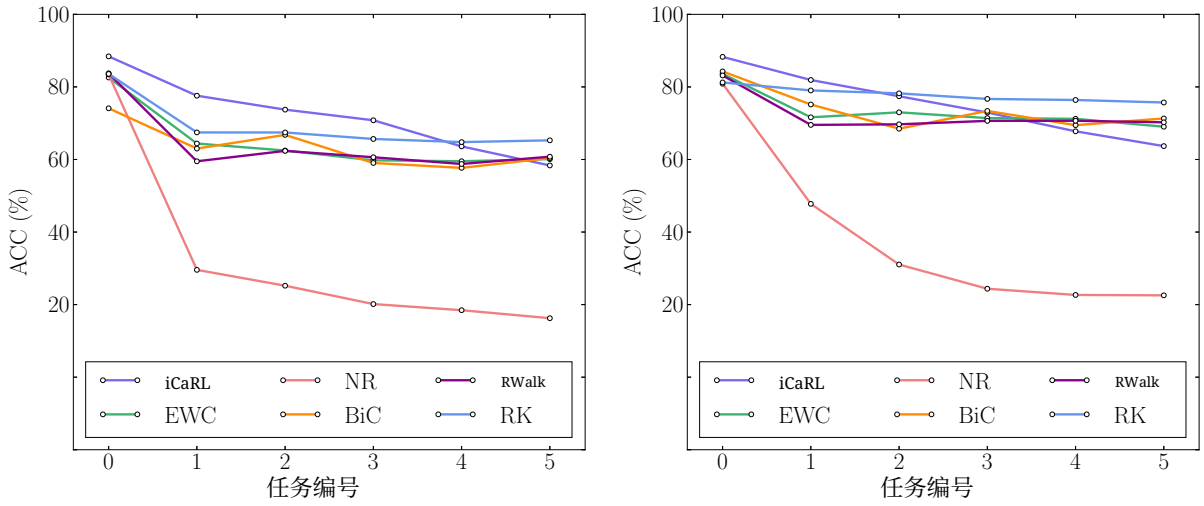
表 3.1： 对所提出的知识蒸馏损失进行对比研究所得的平均准确率（*ACC*）和后向迁移（*BWT*）。*L* 表示 *RK* 的内存大小。

方法（L=500）	**知识蒸馏损失**	**准确率（↑）**	**后向迁移（↑）**
Rainbow 关键词	NO	0.779	-0.033
Rainbow 关键词	YES	**0.779**	**-0.015**

3.3.3 结果

知识蒸馏损失的影响

我们首先分析并总结采用知识蒸馏损失时的性能。如表 3.1所示，我们发现，所提出的知识蒸馏损失函数在BWT方面实现了54.5%的相对提升。



(a) 内存大小 L = 300　　(b) 内存大小 L = 500

图 3.4: 所提出的 *RK* 方法与其他有竞争力的基线方法在不同内存大小下的准确率（%）对比研究。

表3.2: 所提出的数据增强方法的平均准确率（*ACC*）和向后迁移（*BWT*）对比研究。“*NoAugment*”表示实验中未应用数据增强。*L* 表示 *RK* 的内存大小。

方法（L=500）	**准确率（↑）**	**向后迁移（↑）**
NoAugment	0.779	-0.033
SpecAugment [93]	0.783	**-0.006**
Mixup [100]	**0.828**	-0.036

不同任务数量对 RK 方法的影响

我们进一步分析并总结了所提出的RK方法在任务数量增加时的性能，如图 3.3所示。横轴表示任务数量（= 20、10、5、4），纵轴表示准确率的评估指标。例如，当任务数量设为20时，我们使用包含10个唯一关键词的21,000个音频片段对TC-ResNet-8进行预训练，并将其余数据拆分为20个任务，每个任务包含1个唯一的未见关键词。相应的准确率是每个任务完成训练后测试集的准确率。此处所提出的RK方法的内存大小 *L* 为1500。我们观察到，随着任务数量增加，所提出的RK方法取得了最佳准确率表现。尽管随着任务数量增加，保留先验知识的难度也随之增加，所提出的RK方法仍能达到85.0%以上的准确率，显著优于其他基线方法。

不同内存大小对 RK 方法的影响

随后，我们报告不同内存大小（L=300 或 500）对 RK 方法的影响，如图3.4所示。我们将所有方法的内存大小限制为与 RK 方法相同。我们发现，尽管所有方法的性能都随着内存大小的增加（即，可用训练数据增多）而显著提升，所提出的RK方法仍优于其他方法。此外，即使内存大小有限（L=300），所提出的RK方法也能达到超过 65.0% 的准确率，远高于其他方法。

本研究由新加坡国家研究基金会通过其 AI Singapore 计划资助（AISG 奖项编号：AISG-100E-2018-006）。

表 3.3： 近期最先进训练策略对比研究中的准确率和效率指标。为确保公平比较，我们采用 *TC-ResNet-8* 模型作为所有训练策略的测试平台。后续 *RK* 实验中的内存大小 L 设为 *[500, 1500, 3000]*。

方法	准确率(↑)	后向迁移(↑)	参数	内存
微调	0.262	-0.372	64.48K	162.4M
EWC	0.835	-0.064	129.96K	162.4M
PCL-KWS	0.836	-0.041	406.9K	162.4M
NR	0.560	-0.163	64.48K	178.6M
iCaRL	0.846	-0.057	75.29K	178.6M
BiC	0.793	-0.085	64.48K	178.6M
RWalk	0.871	-0.045	129.96K	178.6M
RK-500	0.828	-0.036	129.96K	**16.2M**
RK-1500	0.887	-0.012	129.96K	48.6M
RK-3000	**0.913**	**-0.012**	129.96K	97.2M
联合	0.940	-	64.48K	1624.4M

数据增强的影响

我们还总结了两种数据增强方法对所提出的RK方法的影响，如表3.2所示。我们发现，所有数据增强方法都能提高准确率。采用“Mixup”数据增强时性能最佳。因此，后续我们采用“Mixup”数据增强。

与其他竞争方法的基准比较

表3.3 总结了所提出的 RK 与其他竞争方法在平均准确率（ACC）、反向迁移（BWT）、参数和内存方面的比较。任务编号设为 5。我们观察到，所提出的 RK 在内存大小为 L 3000 时取得了最佳性能。与最佳基线方法 RWalk 相比，所提出的 RK-3000 在使用更少内存的情况下，平均准确率绝对提升了 4.2%，其性能更接近上限。此外，即使只有 16.2M 训练数据，我们的方法 RK-500 也能达到与其他基线方法相当的性能，适用于边缘设备。

3.4 总结

本文提出了一种新颖的多样性感知类增量学习方法，称为 Rainbow Keywords（RK）方法，旨在以更少的内存避免灾难性遗忘。实验结果表明，与最佳基线相比，所提出的RK方法在平均准确率方面实现了4.2%的绝对提升。消融研究也表明，所提出的数据增强和知识蒸馏损失在边缘设备上相当有效。未来，我们计划将该方法应用并调整至多语言关键词搜索系统中 [101–103]。

本文的计算 工作部分借助国家超级计算中心的资源完成 re,
新加坡（h ttps://www.nscc.sg）。

第4章

用于环境声音分类的持续学习

本章研究了用于设备端环境声音分类的基于回放的持续学习（RCL）方法。我们首先考察现有记忆更新算法（MUA）方法（例如 蓄水池 [19], 原型 [20] 和 不确定性 [21]（如第4.2.1节所述）在设备端环境声音分类的RCL中的性能。

我们通过实验证明，不确定性 [21] 方法在我们的场景中表现最佳。此外，我们提出了不确定性++，这是一种基于不确定性 方法的简单而高效的 MUA 方法。与不确定性 方法不同，我们提出的不确定性++ 将扰动引入分类器的嵌入层。因此，在测量数据不确定性时，计算成本（例如运行内存和时间）可显著降低。我们使用设备端模型 BC-ResNet-Mod（~86k 个参数）[24, 25]，在 DCASE 2019 Task1 [22] 和 ESC-50 [23]数据集上评估了我们方法的性能。实验结果表明，不确定性++ 在分类准确率方面优于现有的 MUA 方法，展现出其在真实世界端侧音频应用中的潜力。我们提出的方法与模型无关，且易于应用。我们的代码已发布在 GitHub[1]上。

4.1 相关工作

最近，基于回放的持续学习方法在关键词检测 [10, 16] 和声音事件检测 [104]等音频任务中展现出了良好的结果，优于基于正则化的方法。然而，文献对设备端应用中的持续学习关注较少，例如设备端环境声音分类；本文将重点研究这一问题。设备端场景通常受到存储和内存空间 [17],的限制，这可能给依赖外部存储器来恢复历史数据的基于回放的持续学习带来挑战。因此，可在设备上运行的声音分类模型可能受到容量限制，从而在持续学习新的声音类别时容易遗忘旧知识。

4.2 方法

本节首先介绍基于回放的持续学习和四种记忆更新算法，随后介绍所提出的不确定性++ 算法。

[1] https://github.com/swagshaw/ASC-CL

4.2.1 基于回放的持续学习

按照环境声音分类的持续学习设定 [12, 104, 105]，我们假设模型 M 应在一系列任务中识别所有类别$T = \{\tau_0, \ldots, \tau_t\}$，且不会发生灾难性遗忘。对于每个任务 $\tau \in T$，我们有输入对(x, y)和类别C，其中x表示音频波形，y表示类别$c \in C$。我们的目标是最小化当前任务中所有类别C的交叉熵损失 τ，其形式如下：

$$L_{CE}(\tau) = \sum_{c \in C} y_c log \frac{exp(M(x)_c)}{\sum_{c \in C} exp(M(x)_c)}, \tag{4.1}$$

其中，$M(x)$表示模型M对输入x的输出。

从先前任务中学习到的参数在学习新类别后可能会被覆盖，这种现象也称为灾难性遗忘。为缓解这一问题，我们引入基于回放的方法。基于回放的方法利用一块称为“回放缓冲区”的内存区域，临时存储历史训练样本，以维持性能。

在真实的设备端场景中，使用全部历史数据和新数据的混合数据重新训练声音分类模型会消耗大量资源和时间。为缓解这一问题，基于回放的方法只访问部分历史数据，以节省存储空间。在这种情况下，如何通过记忆更新算法选择样本并将其存入回放缓冲区，是关键所在。

具体而言，在任务 τ_t 的训练过程中，回放缓冲区会存储从先前已学习的任务 $\{\tau_0, \tau_1, \ldots, \tau_{t-1}\}$ 中选出的训练样本，并构建任务 τ_t 的训练数据缓冲区 $\hat{D}_t$，其形式如下：

$$\hat{D}_t = g(\hat{D}_{t-1}) \cup D_t, \tag{4.2}$$

其中，g 是记忆更新算法，[16], $\hat{D}_{t-1}$ 是任务 τ_{t-1} 的训练数据缓冲区，而D_t 是新任务的输入数据。

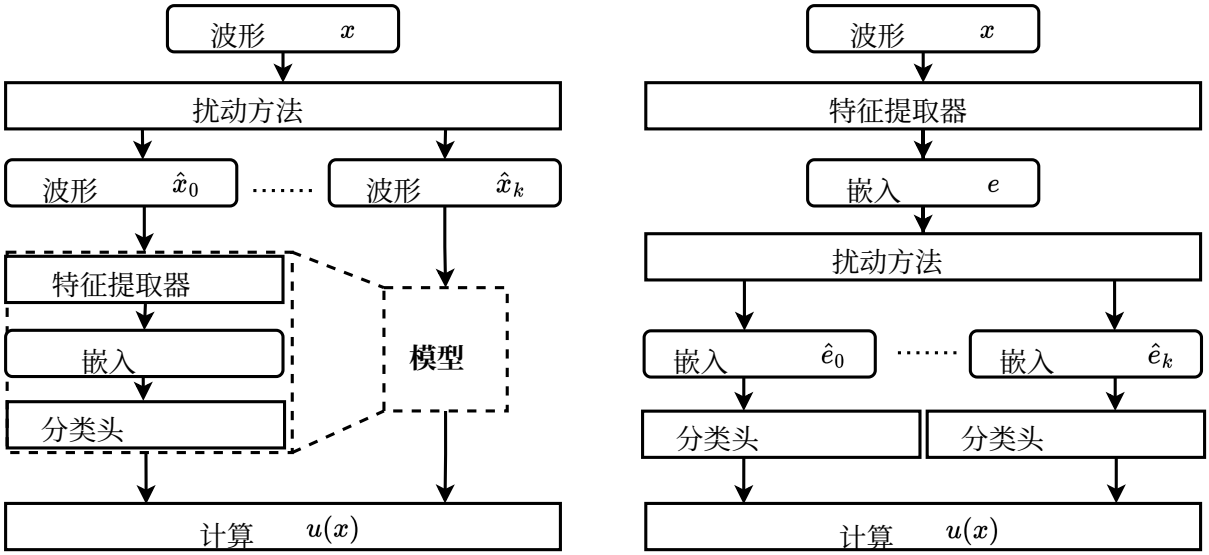
记忆更新算法（MUA）

文献中介绍了四种记忆更新算法。通常，我们假设记忆更新应从前一个任务 τ_{t-1} 的训练数据 $\hat{D}_{t-1}$ 中选取L 个样本，用于训练任务 τ_t。

随机 [96]记忆更新算法从候选样本 $\hat{D}_{t-1}$ 中随机选取L 个新样本 $\{(x_1,y_1),(x_2,y_2)\ldots,(x_L,y_L)\}$，放入回放缓冲区，供下一个任务使用。

蓄水池 [19] 记忆更新算法从 $\hat{D}_{t-1}$中进行均匀采样。具体而言，蓄水池算法初始化回放缓冲区，索引范围为 1 到L，其中包含候选样本中的前L 个样本$\{(x_1, y_1), (x_2, y_2) \ldots, (x_L, y_L)\}$。从候选样本更新回放缓冲区时，对于每个样本，蓄水池算法都会在 $\{1, \ldots, len(\hat{D}_{t-1})\}$ 内均匀生成一个随机数m。如果$m \in \{1, \ldots, L\}$，则用样本 $\hat{D}_{t-1}[m]$替换回放缓冲区中索引为m的样本。

原型 [20] 记忆更新算法从 $\hat{D}_{t-1}$ 中选择样本，其中分类器的嵌入接近其所属类别的嵌入均值。具体而言，该算法首先按唯一类别将 $\hat{D}_{t-1}$划分为子集，表示为 $D_c, c = 1 \ldots N^t$，其中N^t表示 $\hat{D}_{t-1}$ 集中唯一类别的总数。然后，该算法使用当前模型提取每个D_c 中候选样本的嵌入，并通过嵌入计算类别均值，即平均特征向量。对于每个类别，该算法选择候选样本，使回放缓冲区中的平均特征向量



（a）逐波形扰动 **(b) 逐嵌入扰动**

图 4.1： 原生不确定性方法和我们提出的方法的框图。具体而言，朴素方法通过波形向 x 添加扰动，并生成多个波形 $\hat{x}$。我们的方法输入嵌入 e，并生成扰动后的嵌入 $\hat{e}$，这意味着我们只需保存嵌入。模型骨干网络的输出仅计算一次。“计算 $u(x)$”即根据公式 $u(x)$（4.3）进行计算。K 表示通过扰动方法生成的扰动数量。

最接近该类别所有样本的平均特征向量。

不确定性 [21] 记忆更新算法通过分类模型推理得到的样本不确定性来选择样本。具体而言，第一步以与上述原型 算法相同的方式对 $\hat{D}_{t-1}$ 进行分组。第二步估计x 中每个样本的不确定性 D_c。预测似然反映模型对数据的拟合程度，值越大表示模型拟合越好。不确定性得分可由预测似然确定 [92]。根据 [92], 模型给定样本的预测似然的推导，可通过蒙特卡洛（MC）积分 [106] 方法，利用扰动样本的模型输出 [16], 进行近似，定义如下：

$$P(y=c \mid x) = \int p(y=c \mid \hat{x}) p(\hat{x} \mid x) d\hat{x}, \tag{4.3}$$

其中，x、$\hat{x}$、y 分别表示一个类别的音频话语、x的扰动样本以及x的标签。因此，音频话语x 的不确定性定义为$u(x)$：

$$u(x) \approx 1 - \frac{1}{K} \sum_{k=1}^{K} P(y=c \mid \hat{x}_k), \tag{4.4}$$

其中，K 表示由音频偏移 [94],音频 PitchShift [94] 和 Audio Colored Noise [107, 108]等扰动方法生成的扰动数量。$u(x)$越大，表示模型对扰动样本的预测置信度越低。

第三步通过按$u(x)$降序排列，从D_c 中选取L 个样本，步长为$len(D_c) * C/L$，其中L 是回放缓冲区的大小。

先前的研究 [16] 表明，在关键词检测等语音任务中，不确定性记忆更新算法的表现优于其他三种算法。然而，

不确定性随扰动操作次数线性增加。

4.2.2 提出的 MUA 方法（不确定性++）

如图 4.1所示，原生不确定性记忆更新算法需要在线下对每个样本的波形采用扰动方法，首先生成扰动样本。在我们提出的方法中，噪声扰动被添加到样本的分类器前嵌入中，而不是波形中，因此模型骨干网络的输出只需计算一次。具体而言，我们提出了一种逐向量扰动方法，该方法根据分类器嵌入的方差添加不同强度的噪声。我们将分类器嵌入 e 记作 $\hat{e}$，其计算方式如下：

$$\hat{e} = e + U(-\frac{\lambda}{2}, \frac{\lambda}{2}) * std(e), \tag{4.5}$$

其中，$std(\cdot)$ 表示标准差，函数 $U(a,b)$ 表示在 a 到 b 之间均匀分布的噪声，$U(a,b)$ 是一个与 e 形状相同的向量，而 λ 是一个用于控制相对噪声强度的超参数。

通过逐向量扰动方法，我们生成嵌入 e 的扰动嵌入 e。最后，我们将 e 输入模型的最终分类层，并输出 $P(y = c \mid \mathrm{e}e)$，并如公式（4.3）所示，用该输出计算不确定性。估算出不确定性后，我们采用与原始方法相同的方式选择样本进行回放。该方法只需计算一次模型骨干网络的输出，从而节省时间。我们还通过用分类器嵌入替代额外的受扰动原始数据来降低内存占用；与原始数据相比，分类器嵌入的大小要小得多。

4.3 实验与结果

4.3.1 环境声音分类模型

对于设备端环境声音分类模型，我们使用 BC-ResNet-Mod [25]，它是 BC-ResNet [24] 的一种改进版本，在声学场景分类方面取得了更好的结果。BC-ResNet 范式通过反复依次提取频谱特征和时间特征来运行。由于这些频谱特征的维度低于输入维度，该模型的参数比直接处理波形的模型更少。特征提取采用逐通道方式，这两种参数缩减对性能的影响都可以忽略不计 [24]。在我们的实验中，我们使用 BC-ResNet-Mod-4，它会在提取频谱特征和时间特征之前，将输入通道维度增加到 80。

4.3.2 数据集

ESC-50 由 2000 段时长为 5 秒的环境音频录音 [23]组成。数据均衡分布于 50 个类别中，每个类别包含 40 个样本，涵盖动物声音、自然声景、人类声音（非语音）和环境噪声。该数据集已预先划分为五折，用于交叉验证。

DCASE 2019 任务 1 是一项声学场景分类任务，其开发集 [22]由来自 10 种声学场景的 10 秒音频片段组成：机场、室内购物中心、地铁站、步行街、公共广场、中等交通量街道、乘坐有轨电车、乘坐公交车、乘坐地下地铁和城市公园。开发集中分别有 9185 个训练音频片段和 4185 个验证音频片段。

表 4.1：所提出的记忆更新算法的对比研究中的准确率（*ACC*）和反向迁移（*BWT*）。

方法	DCASE 2019 任务 1		ESC-50	
	准确率 ↑	后向迁移 ↑	准确率 ↑	后向迁移 ↑
微调	0.205	-0.276	0.181	-0.307
随机	0.473	-0.115	0.225	-0.231
蓄水池	0.568	-0.096	0.430	-0.121
原型	0.559	-0.089	0.482	**-0.104**
不确定性	0.578	**-0.079**	0.477	-0.111
不确定性++	**0.581**	**-0.079**	**0.500**	-0.121

4.3.3 实验设置

任务设置 为评估所提出方法的性能，我们将数据拆分为五个任务。每个任务都包含 DCASE 19 Task 1 中的 2 个新唯一类别和 ESC-50 中的 10 个新唯一类别，这些类别在先前任务中均未出现。为模拟边缘设备的使用条件，考虑到内存限制，我们将 DCASE 19 Task 1 和 ESC-50 中的样本缓冲区大小 L 分别设为 500 和 100。

实现细节 原始音频片段通过短时傅里叶变换转换为 64 维对数梅尔频谱图，帧长为 1024 个样本，帧移为 320 个样本，并使用汉宁窗。分类网络通过 Adam [109]算法进行优化，学习率为 1×10^{-3}。批次大小设为 32，训练轮数为 50。

4.3.4 评估指标

我们采用准确率和遗忘度量来报告性能。具体而言，准确率(ACC) 表示整个训练结束后，在已学习类别上计算的平均准确率。向后迁移 (BWT) [99] 用于评估学习新任务后所有先前任务的准确率变化，从而反映遗忘程度。为衡量 BWT，我们首先构建矩阵 $R \in \mathbb{R}^{T \times T}$，其中 $R_{i,\ j}$表示模型在观察任务 τ_i的最后一个样本后，在任务 τ_j 上的测试分类准确率。模型完成对每个任务 τ_i的学习后，我们评估其在所有 T 个任务上的 BWT，其定义如下：

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} R_{T,i} - R_{i,i}. \tag{4.6}$$

当学习某项任务导致某个先前任务的性能下降时，就会出现负后向迁移。BWT 值越小，表明灾难性遗忘越严重。

4.3.5 参考基线

我们构建了五种基线方法进行比较。微调训练策略针对每个新任务调整 BC-ResNet-Mod 模型，但不采用任何持续学习策略，作为性能下界基线。基于回放的持续学习所采用的四种既有记忆更新算法（即 随机、蓄水池、原型、不确定性）在第 4.2.1 节中介绍。具体而言，在不确定性的扰动阶段，我们采用两种扰动方法，即“不确定性偏移”，其中包括音频偏移和音频

表 4.2：对所提出的扰动方法进行对比研究时的准确率（*ACC*）和反向迁移（*BWT*）。*K* 指扰动方法生成的扰动数量。

方法	K	DCASE 2019 任务 1		ESC-50	
		准确率 ↑	后向迁移 ↑	准确率 ↑	后向迁移 ↑
不确定性偏移	2	0.557	-0.101	0.461	**-0.111**
	4	0.575	-0.103	0.476	-0.118
	6	0.567	-0.079	0.477	-0.118
不确定性噪声	2	0.560	-0.100	0.465	-0.118
	4	0.535	-0.104	0.473	-0.118
	6	0.578	-0.079	0.458	-0.120
不确定性++	2	0.571	-0.102	**0.500**	-0.121
	4	0.548	-0.103	0.481	-0.114
	6	**0.581**	**-0.079**	0.484	-0.119

音高偏移，以及“不确定性噪声”，其指的是 Audio Colored Noise 扰动方法。

4.3.6 结果

MUA 方法实验

表 1 展示了在 DCASE 2019 任务 1 和 ESC-50 测试集上的准确率（ACC）和后向迁移（BWT）结果。我们将提出的 不确定性++ MUA 方法与五种基线方法进行了比较。结果显示，不确定性MUA 方法的性能优于五种基线方法。与最佳基线不确定性相比，提出的 不确定性++ 方法的分类准确率达到 58.1%，优于现有的 MUA 方法。此外，与其他基线方法相比，微调 方法的准确率和后向迁移表现最差，这表明存在灾难性遗忘问题。

我们进一步分析并总结了所提出的不确定性++方法与不确定性 MUA 方法在不同扰动方法数量下的性能，并根据准确率和后向迁移进行比较，如表 2 所示。*K* 表示由扰动方法生成的扰动数量。即使仅使用两种扰动方法，我们提出的方法仍优于另外两种基线方法。我们还观察到，使用两种扰动时，我们的方法在 ESC-50 测试集上取得了最佳性能。这种性能可能是由于 ESC-50 规模较小，因此对扰动更敏感。

不确定性与不确定性++的计算时间对比实验

我们进一步报告了所提出的方法在扰动数量不断增加时的平均时间。平均时间衡量的是每项任务中相对于训练时间的时间增幅。如表 3 所示，即使有 6 种扰动，不确定性++ 的平均时间仍不到 60 秒。这是因为我们提出的方法能够限制额外训练时间的增长。我们还发现，无论扰动数量是多少，我们提出的方法都优于其他基线方法，这表明该方法在计算方面更加高效。此外，

表 4.3：所提出的不确定性++ 方法的对比研究中的平均时间（秒）。K 表示扰动方法生成的扰动数量。

方法	K	平均时间（秒）↓
不确定性偏移	2	1221.7
	4	2205.1
	6	2926.1
不确定性噪声	2	246.2
	4	390.8
	6	506.3
不确定性++	2	44.0
	4	48.5
	6	55.1

不确定性偏移 的平均时间远长于其他方法。这是因为音频偏移和音频音高偏移这两种扰动比简单地添加噪声更耗时。

4.4 总结

本章介绍了不确定性++，这是一种用于设备端环境声音分类的高效基于回放的持续学习方法。我们通过测量分类器嵌入层上的逐样本分类不确定性，选择用于训练的历史数据。在 DCASE 2019 任务 1 和 ESC-50 数据集上的实验结果表明，所提出的方法在分类准确率和计算效率方面均优于基线持续学习方法。未来，我们计划将该方法应用并调整至其他设备端音频分类任务，例如音频标注和声音事件检测。

第5章

结论与建议

5.1 结论

目前，基于深度学习的音频分类系统通常在类别有限的紧凑型模型上进行训练，以降低计算量并缩小占用空间。因此，当运行时遇到目标域中未见过的类别时，使用源域数据训练的模型性能可能会显著下降。微调这一在领域迁移问题中非常有效的朴素方法，因缺少先前任务的数据而受到限制，导致分类器无法对来自这些任务的数据进行分类。这种先前学会的任务上的性能急剧下降被称为灾难性遗忘 [11]。持续学习旨在防止灾难性遗忘。本学位论文首先探讨音频分类任务——关键词检测和环境声音分类——的最新进展，随后综述持续学习方法的若干类别。针对关键词检测的持续学习，我们提出一种新颖的多样性感知类增量学习方法，名为 Rainbow Keywords（RK）方法，以更少的内存避免灾难性遗忘。实验结果表明，与表现最佳的基线方法相比，所提出的 RK 方法在平均准确率方面实现了 4.2% 的绝对提升。消融研究也表明，所提出的数据增强和知识蒸馏损失在边缘设备上相当有效。另一方面，针对环境声音分类，我们提出不确定性++，这是一种用于设备端环境声音分类的高效基于回放的持续学习方法。我们通过测量分类器嵌入层上的逐样本分类不确定性，选择用于训练的历史数据。在 DCASE 2019 任务 1 和 ESC-50 数据集上的实验结果表明，所提出的方法在分类准确率和计算效率方面均优于基线持续学习方法。

5.2 未来工作

5.2.1 不同信噪比条件下的持续学习

在这项工作之后，我计划研究持续学习在不同信噪比条件下的影响，使关键词检测系统能够在干净和噪声环境中自适应并持续运行。关键的是，我们始终使用涵盖不同信噪比级别的完整数据集进行训练，只改变训练顺序。具体而言，我们将训练过程分为五个难度逐步递增的阶段。开始时，我们使用不含噪声的干净样本对模型进行训练，并采用持续学习方法，让模型保留先前习得的知识。因此，模型不仅在干净条件下表现更佳，在噪声环境中的性能也有所提升。

5.2.2 多语言持续学习

[90] 介绍了一种适用于任意语言的关键词检测少样本迁移学习方法。他们使用 Common Voice 的 [110] 多语言众包语音数据集，对嵌入模型进行关键词分类训练，并通过强制对齐 [111]自动提取九种语言中的 760 个高频词。随后，他们对该嵌入模型进行微调，仅用五条样本语句即可对目标关键词进行分类，即使模型此前从未接触过目标语言也不例外。在 22 种语言的 440 个关键词上，他们实现了 87.4% 的平均流式关键词检出准确率，误接受率为 4.3%，并在关键词搜索方面观察到令人鼓舞的初步结果。

近年来，出现了其他更适用于多语言关键词检出的跨语言学习方法。跨语言学习旨在构建能够利用其他语言数据来提升性能的模型。XLS-R [112] 是一个基于 wav2vec 2.0 的大规模跨语言语音表征学习模型。研究人员使用近 50 万小时、涵盖 128 种语言的公开语音音频数据训练模型，参数量最高达 20 亿；其公开数据量比已知此前最大规模的研究高出一个数量级。此外，他们还表明，只要模型大小足够大，在将英语语音翻译成其他语言时，跨语言预训练的表现可以媲美仅英语预训练，而这种情形有利于单语预训练。XLS-R 有助于提升世界上更多语言的语音处理任务表现，其中也包括关键词检测。

我提议将跨语言预训练模型 XLS-R 与 wav2kws [55] 相结合，用于多语言关键词检出任务。此外，我还提议探索多语言关键词检出的持续学习方法。例如，解决学习一系列语言时的遗忘问题。

参考文献

[1] Haohe Liu、Xubo Liu、Qiuqiang Kong、Wenwu Wang 和 Mark D Plumbley。学习音频分类的声谱图时间分辨率。*arXiv* 预印本：*2210.01719*，2022年。[2] Xubo Liu、Haohe Liu、Qiuqiang Kong、Xinhao Mei、Mark D Plumbley 和 Wenwu Wang。用于高效音频分类的简单池化前端。*arXiv* 预印本：*2210.00943*，2022年。[3] Annamaria Mesaros、Toni Heittola、Tuomas Virtanen 和 Mark D Plumbley。声音事件检测：教程。*IEEE Signal Processing Magazine*，38(5)：67–83，2021年。[4] Iv´an L´opez-Espejo、Zheng-Hua Tan、John Hansen 和 Jesper Jensen。深度口语关键词检出：综述。*IEEE Access*，2021年。[5] Regunathan Radhakrishnan、Ajay Divakaran 和 A Smaragdis。面向监控应用的音频分析。载于*IEEE* 音频与声学信号处理应用研讨会论文集，第158–161页。IEEE，2005年。[6] Ya-Ti Peng、Ching-Yung Lin、Ming-Ting Sun 和 Kun-Cheng Tsai。使用隐马尔可夫模型和分层隐马尔可夫模型的医疗音频事件分类。载于*IEEE* 多媒体与博览会国际会议论文集，第1218–1221页。IEEE，2009年。[7] Serkan Kiranyaz、Ahmad Farooq Qureshi 和 Moncef Gabbouj。用于多媒体索引与检索的通用音频分类与分割方法。*IEEE* 音频、语音与语言处理汇刊，14(3)：1062–1081，2006年。[8] Yong Xu、Qiuqiang Kong、Wenwu Wang 和 Mark D Plumbley。使用门控卷积神经网络进行大规模弱监督音频分类。载于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），第121–125页。IEEE，2018年。[9] Yang Xiao、Xubo Liu、James King、Arshdeep Singh、Eng Siong Chng、Mark D Plumbley 和 Wenwu Wang。设备端环境声音分类的持续学习。*arXiv* 预印本：*2207.07429*，2022年。[10] Yizheng Huang、Nana Hou 和 Nancy F Chen。口语关键词检出的渐进式持续学习。载于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），第7552–7556页。IEEE，2022年。[11] Michael McCloskey 和 Neal J Cohen。联结主义网络中的灾难性干扰：序列学习问题。载于*Psychology oflearning and motivation*，第24卷，第109–165页。爱思唯尔，1989年。[12] Abhijeet Awasthi 和 Sunita Sarawagi。神经网络持续学习：综述。载于*ACM* 印度数据科学与数据管理联合国际会议论文集，第362–365页，2019年。[13] Zheda Mai、Ruiwen Li、Jihwan Jeong、David Quispe、Hyunwoo Kim 和 Scott Sanner。图像分类中的在线持续学习：实证综述。*Neurocomputing*，469：28–51，2022年。

[14] Timoth´ee Lesort、Vincenzo Lomonaco、Andrei Stoian、Davide Maltoni、David Filliat 和 Natalia D´ıaz-Rodr´ıguez。《机器人学中的持续学习：定义、框架、学习策略、机遇与挑战》。*Information Fusion*，58：52–68，2020年。[15] Magdalena Biesialska、Katarzyna Biesialska 和 Marta R Costa-Jussa。《自然语言处理中的持续终身学习：综述》。*arXiv*预印本：*2012.09823*，2020年。[16] Yang Xiao、Nana Hou 和 Eng Siong Chng。《Rainbow关键词：用于在线口语关键词检测的高效增量学习》。*arXiv*预印本：*2203.16361*，2022年。[17] Arshdeep Singh 和 Mark D Plumbley。《一种用于高效声学场景分类的基于被动相似度的CNN滤波器剪枝方法》。*arXiv*预印本：*2203.15751*，2022年。[18] Ameya Prabhu、Philip HS Torr 和 Puneet K Dokania。《GDumb：一种质疑持续学习进展的简单方法》。载于欧洲计算机视觉会议（*ECCV*）论文集，第524–540页。Springer，2020年。[19] Jeffrey S Vitter。《使用蓄水池进行随机采样》。*ACM*数学软件汇刊（*TOMS*），11(1)：37–57，1985年。[20] Sylvestre-Alvise Rebuffi、Alexander Kolesnikov、Georg Sperl 和 Christoph H Lampert。《iCaRL：增量分类器与表征学习》。载于*IEEE/CVF*国际计算机视觉会议论文集，第2001–2010页，2017年。[21] Jihwan Bang、Heesu Kim、YoungJoon Yoo、Jung-Woo Ha 和 Jonghyun Choi。《Rainbow记忆：利用多样化样本记忆实现持续学习》。载于*IEEE/CVF*国际计算机视觉会议论文集，第8218–8227页，2021年。[22] Toni Heittola、Annamaria Mesaros 和 Tuomas Virtanen。《TAU 2019 城市声学场景，开发数据集》。2019年3月。[23] Karol J Piczak。《ESC：用于环境声音分类的数据集》。载于*ACM*国际多媒体会议论文集，第1015–1018页，2015年。[24] Byeonggeun Kim、Simyung Chang、Jinkyu Lee 和 Dooyong Sung。《用于高效关键词检测的广播式残差学习》。*arXiv*预印本：*2106.04140*，2021年。[25] Byeonggeun Kim、Seunghan Yang、Jangho Kim 和 Simyung Chang。《QTI提交至DCASE 2021：通过高效设计实现设备不平衡声学场景分类的残差归一化》。技术报告，DCASE2021挑战赛，2021年6月。[26] Mittal C Darji。《音频信号处理：音频信号分类特征综述》。《国际计算机科学、工程与信息技术研究杂志》，2(3)：227–230，2017年。[27] Yimeng Zhuang、Xuankai Chang、Yanmin Qian 和 Kai Yu。《使用LSTM-CTC进行无限词汇关键词检测》。载于*Interspeech*论文集，第938–942页，2016年。[28] Guoguo Chen、Carolina Parada 和 Georg Heigold。《使用深度神经网络进行小型关键词检测》。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第4087–4091页。IEEE，2014年。[29] Raziel Alvarez 和 Hyun-Jin Park。《端到端流式关键词检测》。载于 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）论文集，第6336–6340页。IEEE，2019年。[30] Preetum Nakkiran、Raziel Alvarez、Rohit Prabhavalkar 和 Carolina Parada。《使用秩约束拓扑压缩深度神经网络》。2015年。

[31] Bruno U Pedroni、Sadique Sheik、Hesham Mostafa、Somnath Paul、Charles Augustine 和 Gert Cauwenberghs。《用于低功耗关键词检测的小型脉冲神经网络》。载于*IEEE*生物医学电路与系统会议（*BioCAS*）论文集，第1–4页。IEEE，2018年。[32] Tara Sainath 和 Carolina Parada。《用于小型关键词检测的卷积神经网络》。2015年。[33] Raphael Tang 和 Jimmy Lin。《用于小型关键词检测的深度残差学习》。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第5484–5488页。IEEE，2018年。[34] Emad A Ibrahim、Jos Huisken、Hamed Fatemi 和 Jose Pineda de Gyvez。《在时序卷积网络中使用时域特征进行关键词检测》。载于*Euromicro*数字系统设计会议（*DSD*）论文集，第313–319页。IEEE，2019年。[35] Menglong Xu 和 Xiao-Lei Zhang。《采用挤压-激励模块的深度可分离卷积ResNet，用于小型关键词检测》。*Interspeech*论文集，第2547–2551页，2020年。[36] Alice Coucke、Mohammed Chlieh、Thibault Gisselbrecht、David Leroy、Mathieu Poumeyrol 和 Thibaut Lavril。《使用扩张卷积和门控实现高效关键词检测》。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第6351–6355页。IEEE，2019年。[37] Seungwoo Choi、Seokjun Seo、Beomjun Shin、Hyeongmin Byun、Martin Kersner、Beomsu Kim、Dongyoung Kim 和 Sungjoo Ha。《面向移动设备实时关键词检测的时间卷积》。*arXiv*预印本：*1904.03814*，2019年。[38] Simon Mittermaier、Ludwig K¨urzinger、Bernd Waschneck 和 Gerhard Rigoll。《使用sinc卷积在原始音频数据上进行小型关键词检测》。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第7454–7458页。IEEE，2020年。[39] Andrew G Howard、Menglong Zhu、Bo Chen、Dmitry Kalenichenko、Weijun Wang、Tobias Weyand、Marco Andreetto 和 Hartwig Adam。《MobileNets：面向移动视觉应用的高效卷积神经网络》。*arXiv*预印本：*1704.04861*，2017年。[40] Rajath Kumar、Vaishnavi Yeruva 和 Sriram Ganapathy。《关于联合唤醒词检测与文本相关说话人验证的卷积长短期记忆网络建模》。载于*Interspeech*论文集，第1121–1125页，2018年。[41] Harshavardhan Sundar、Jill Fain Lehman 和 Rita Singh。《面向儿童的多人语音驱动游戏中的关键词检测》。载于*Interspeech*论文集，2015年。[42] Oleg Rybakov、Natasha Kononenko、Niranjan Subrahmanya、Mirk´o Visontai 和 Stella Laurenzo。《移动设备上的流式关键词检测》。*arXiv*预印本：*2005.06720*，2020年。[43] Sercan O Arik、Markus Kliegl、Rewon Child、Joel Hestness、Andrew Gibiansky、Chris Fougner、Ryan Prenger 和 Adam Coates。《用于小型关键词检测的卷积循环神经网络》。*arXiv*预印本：*1703.05390*，2017年。[44] Ali Madani、Bryan McCann、Nikhil Naik、Nitish Shirish Keskar、Namrata Anand、Raphael R Eguchi、Po-Ssu Huang 和 Richard Socher。《ProGen：用于蛋白质生成的语言建模》。*arXiv*预印本：*2004.03497*，2020年。[45] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。《语言模型是少样本学习者》。神经信息处理系统进展论文集，33：1877–1901，2020年。

[46] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本：*1810.04805*，2018年。[47] Cheng-Zhi Anna Huang、Ashish Vaswani、Jakob Uszkoreit、Noam Shazeer、Ian Simon、Curtis Hawthorne、Andrew M Dai、Matthew D Hoffman、Monica Dinculescu 和 Douglas Eck。音乐变换器。*arXiv* 预印本：*1809.04281*，2018年。[48] Rohit Girdhar、Joao Carreira、Carl Doersch 和 Andrew Zisserman。视频动作变换器网络。载于《*IEEE/CVF*国际计算机视觉会议论文集》，第244–253页，2019年。[49] Chen Sun、Austin Myers、Carl Vondrick、Kevin Murphy 和 Cordelia Schmid。VideoBERT：用于视频与语言表征学习的联合模型。载于《*IEEE/CVF*国际计算机视觉会议论文集》，第7464–7473页，2019年。[50] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过 16×16 个词：用于大规模图像识别的变换器。*arXiv* 预印本：*2010.11929*，2020年。[51] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin变换器：使用移位窗口的层次化视觉变换器。载于《*IEEE/CVF*国际计算机视觉会议论文集》，第10012–10022页，2021年。[52] Axel Berg、Mark O'Connor 和 Miguel Tairum Cruz。关键词 Transformer：用于关键词检测的自注意力模型。*arXiv* 预印本：*2104.00769*，2021年。[53] Wei-Tsung Kao、Yuen-Kwei Wu、Chia Ping Chen、Zhi-Sheng Chen、Yu-Pao Tsai 和 Hung-Yi Lee。论自监督学习与元学习相结合在用户自定义少样本关键词检测中的效率。*arXiv* 预印本：*2204.00352*，2022年。[54] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊，29：3451–3460，2021年。[55] Deokjin Seo、Heung-Seon Oh 和 Yuchul Jung。Wav2KWS：从语音表征迁移学习以实现关键词检测。*IEEEAccess*，9：80682–80691，2021年。[56] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：自监督语音表征学习框架。神经信息处理系统进展，33：12449–12460，2020年。[57] Abdelrahman Mohamed、Hung-yi Lee、Lasse Borgholt、Jakob D Havtorn、Joakim Edin、Christian Igel、Katrin Kirchhoff、Shang-Wen Li、Karen Livescu、Lars Maaløe 等。自监督语音表征学习：综述。*arXiv* 预印本：*2205.10643*，2022年。[58] Pete Warden。Speech Commands：一个用于有限词汇语音识别的数据集。*arXiv* 预印本：*1804.03209*，2018年。[59] Hengshun Zhou、Jun Du、Gongzhen Zou、Zhaoxu Nian、Chin-Hui Lee 和 Sabato Marco。MISP2021挑战赛中的音频-视觉唤醒词检测：数据集发布与深入分析。2022年。[60] Karol J Piczak。使用卷积神经网络进行环境声音分类。载于《*IEEE*论文集》 *25th*《机器学习与信号处理国际研讨会（*MLSP*）论文集》，第1–6页，2015年。

[61] Brian Gygi 和 Valeriy Shafiro。当今的环境声音研究。载于《声学会议第*153*卷论文集》，第1卷，第050002页。美国声学学会，2007年。[62] William W Gaver。我们究竟听到了什么？：听觉事件感知的生态学方法。生态心理学，5(1)：1–29，1993年。[63] Arshdeep Singh、James A King、Xubo Liu、Wenwu Wang 和 Mark D. Plumbley。用于声学场景分类的低复杂度卷积神经网络。技术报告，DCASE2022 挑战赛，2022年6月。[64] Kwanghee Choi、Martin Kersner、Jacob Morton 和 Buru Chang。用于设备端音频分类的时序知识蒸馏。载于《*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集》，第486–490页，2022年。[65] Irene Mart´ín-Morat´o、Francesco Paissan、Alberto Ancilotto、Toni Heittola、Annamaria Mesaros、Elisabetta Farella、Alessio Brutti 和 Tuomas Virtanen。DCASE 2022 挑战赛中的低复杂度声学场景分类。*arXiv* 预印本：*2206.03835*，2022年。[66] Haohe Liu、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Wenwu Wang 和 Mark D Plumbley。萨里系统参加 DCASE 2022 第5项任务：利用片段级度量学习进行小样本生物声学事件检测——技术报告。技术报告，DCASE2022 挑战赛，2022年6月。[67] Annamaria Mesaros、Toni Heittola、Emmanouil Benetos、Peter Foster、Mathieu Lagrange、Tuomas Virtanen 和 Mark D Plumbley。声学场景与事件检测和分类：DCASE 2016 挑战赛成果。*IEEE/ACM*音频、语音与语言处理汇刊，26(2)：379–393，2017年。[68] Annamaria Mesaros、Aleksandr Diment、Benjamin Elizalde、Toni Heittola、Emmanuel Vincent、Bhiksha Raj 和 Tuomas Virtanen。DCASE 2017 挑战赛中的声音事件检测。*IEEE/ACM*音频、语音与语言处理汇刊，27(6)：992–1006，2019年。[69] Zhao Ren、Qiuqiang Kong、Jing Han、Mark D Plumbley 和 Bj¨orn W Schuller。基于注意力的空洞卷积神经网络：声学场景的可视化与理解视角。载于《*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集》，第56–60页。IEEE，2019年。[70] Khaled Koutini、Hamid Eghbal-zadeh、Gerhard Widmer 和 J Kepler。Cp-jku 在 DCASE'19 中的投稿：采用感受野正则化卷积神经网络进行声学场景分类和音频标注。载于《声学场景与事件检测和分类研讨会（*DCASE*）论文集》，第25–26页，2019年。[71] Liping Yang、Xinxing Chen 和 Lianjie Tao。利用多尺度特征进行声学场景分类。载于《声学场景与事件检测和分类研讨会（*DCASE*）论文集》，第29–33页，2018年。[72] Janghoon Cho、Sungrack Yun、Hyoungwoo Park、Jungyun Eum 和 Kyuwoong Hwang。基于大间隔分解卷积神经网络的声学场景分类。*arXiv* 预印本 *arXiv*：*1910.06784*，2019年。[73] Yu-Chi Wu、Pao-Chi Chang、Chien-Yao Wang 和 Jia-Ching Wang。用于声学场景分类的非对称卷积神经网络。载于《*IEEE*消费电子国际研讨会（*ISCE*）论文集》，第11–12页。IEEE，2017年。[74] Ahmet Melih Basbug 和 Mustafa Sert。使用卷积神经网络和空间金字塔池化进行声学场景分类。载于《*IEEE*语义计算国际会议（*ICSC*）论文集》，第128–131页。IEEE，2019年。

[75] Erik Marchi、Dario Tonelli、Xinzhou Xu、Fabien Ringeval、Jun Deng、Stefano Squartini 和 Bjoern Schuller。用于声学场景分类的成对分解、深度神经网络与多尺度核子空间学习。载于声学场景与事件检测和分类研讨会（*DCASE*），第 65–69 页，2016年9月。[76] Victor Bisot、Romain Serizel、Slim Essid 和 Ga¨el Richard。用于声学场景分类的非负特征学习方法。2017年。[77] Gen Takahashi、Takeshi Yamada、Nobutaka Ono 和 Shoji Makino。使用深度神经网络高斯混合模型和帧拼接声学特征进行声学场景分类的性能评估。载于亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）论文集，第 1739–1743 页，2017年。[78] Emre Cakır、Giambattista Parascandolo、Toni Heittola、Heikki Huttunen 和 Tuomas Virtanen。用于多声部声音事件检测的卷积循环神经网络。*IEEE/ACM*音频、语音与语言处理汇刊，25(6):1291–1303，2017年。[79] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley。PANNs：用于音频模式识别的大规模预训练音频神经网络。*IEEE/ACM*音频、语音与语言处理汇刊，28:2880–2894，2020年。[80] Yuan Gong、Yu-An Chung 和 James Glass。PSLA：通过预训练、采样、标注和聚合改进音频标注。*IEEE/ACM*音频、语音与语言处理汇刊，29:3292–3306，2021年。[81] Pengcheng Li、Yan Song、Ian Vince McLoughlin、Wu Guo 和 Li-Rong Dai。一种基于注意力池化的语音情感识别表征学习方法。2018年。[82] Yuan Gong、Yu-An Chung 和 James Glass。AST：音频频谱图 Transformer。*arXiv*预印本：*2104.01778*，2021年。[83] Yuan Gong、Cheng-I Lai、Yu-An Chung 和 James Glass。SSAST：自监督音频频谱图 Transformer。载于*AAAI*人工智能会议论文集，第 36 卷，第 10699–10709 页，2022年。[84] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。音频集：音频事件本体与人工标注数据集。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第 776–780 页。IEEE，2017年。[85] Marc Masana、Xialei Liu、Bartlomiej Twardowski、Mikel Menta、Andrew D Bagdanov 和 Joost van de Weijer。类别增量学习：图像分类综述与性能评估。*arXiv*预印本：*2010.15277*，2020年。[86] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等。克服神经网络中的灾难性遗忘。美国国家科学院院刊，114(13):3521–3526，2017年。[87] Zhizhong Li 和 Derek Hoiem。无遗忘学习。*IEEE*《模式分析与机器智能汇刊》，40(12):2935–2947，2017年。[88] Yue Wu、Yinpeng Chen、Lijuan Wang、Yuancheng Ye、Zicheng Liu、Yandong Guo 和 Yun Fu。大规模增量学习。载于*IEEE/CVF*国际计算机视觉会议论文集，第 374–382 页，2019年。

[89] Abhijeet Awasthi、Kevin Kilgour 和 Hassan Rom。通过有限样本教关键词检测器检测新关键词。*arXiv*预印本：*2106.02443*，2021年。[90] Mark Mazumder、Colby Banbury、Josh Meyer、Pete Warden 和 Vijay Janapa Reddi。适用于任何语言的少样本关键词检测。*Interspeech 2021*，2021年8月。[91] Matthias Delange、Rahaf Aljundi、Marc Masana、Sarah Parisot、Xu Jia、Ales Leonardis、Greg Slabaugh 和 Tinne Tuytelaars。持续学习综述：在分类任务中克服遗忘。*IEEE*《模式分析与机器智能汇刊》，2021年。[92] Yarin Gal 和 Zoubin Ghahramani。将随机失活作为贝叶斯近似：表征深度学习中的模型不确定性。载于国际机器学习会议论文集，第 1050–1059 页。PMLR，2016年。[93] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le。SpecAugment：一种简单的自动语音识别数据增强方法。*arXiv*预印本：*1904.08779*，2019年。[94] Tom Ko、Vijayaditya Peddinti、Daniel Povey 和 Sanjeev Khudanpur。用于语音识别的音频增强。载于*Interspeech*论文集，2015年。[95] Geoffrey Hinton、Oriol Vinyals、Jeff Dean 等。神经网络中的知识蒸馏。*arXiv*预印本：*1503.02531*，2(7)，2015年。[96] Yen-Chang Hsu、Yen-Cheng Liu、Anita Ramasamy 和 Zsolt Kira。重新评估持续学习场景：分类方法及采用强基线的理由。*arXiv*预印本：*1810.12488*，2018年。[97] Arslan Chaudhry、Puneet K Dokania、Thalaiyasingam Ajanthan 和 Philip HS Torr。用于增量学习的黎曼游走：理解遗忘与僵化。载于欧洲计算机视觉会议（*ECCV*）论文集，第 532–547 页，2018年。[98] Friedemann Zenke、Ben Poole 和 Surya Ganguli。通过突触智能实现持续学习。载于国际机器学习会议论文集，第 3987–3995 页。PMLR，2017年。[99] David Lopez-Paz 和 Marc'Aurelio Ranzato。用于持续学习的梯度情景记忆。神经信息处理系统进展论文集，第 30 卷，2017年。[100] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz。Mixup：超越经验风险最小化。*arXiv*预印本：*1710.09412*，2017年。[101] Nancy F Chen、Pharri Van Tung、Haihua Xu、Xiong Xiao、Chongjia Ni、I-Fan Chen、Sunil Sivadas、Chin-Hui Lee、Eng Siong Chng、Bin Ma 等。面向斯瓦希里语低资源口语关键词检索的受样本启发策略。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第 6040–6044 页。IEEE，2016年。[102] Nancy F Chen、Sunil Sivadas、Boon Pang Lim、Hoang Gia Ngo、Haihua Xu、Bin Ma、Haizhou Li 等。越南语关键词检索策略。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第 4121–4125 页。IEEE，2014年。[103] Nancy F Chen、Chongjia Ni、I-Fan Chen、Sunil Sivadas、Haihua Xu、Xiong Xiao、Tze Siong Lau、Su Jun Leow、Boon Pang Lim、Cheung-Chi Leung 等。泰米尔语低资源关键词检索策略。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第 5366–5370 页。IEEE，2015年。

[104] Zhepei Wang、Cem Subakan、Efthymios Tzinis、Paris Smaragdis 和 Laurent Charlin。利用生成式回放持续学习新的声音类别。载于*IEEE* 音频与声学信号处理应用研讨会论文集（*WASPAA*），第 308–312 页，2019年。[105] Friedemann Zenke、Ben Poole 和 Surya Ganguli。通过突触智能实现持续学习。载于国际机器学习会议论文集，第 3987–3995 页，2017年。[106] Tuen Kloek 和 Herman K Van Dijk。方程组参数的贝叶斯估计：蒙特卡罗积分的一项应用。《计量经济学》：计量经济学会期刊，第 1–19 页，1978年。[107] Iver Jordal、Shahul ES、Herv´e BREDIN、Kento Nishi、Francis Lata、Harry Coultas Blum、Pariente˙Manuel、akash raj、Keunwoo Choi、FrenchKrab、Piotr Zelasko、amiasato、Moreno La Quatra 和 Emmanuel Schmidbauer。asteroid-team/torch-audiomentations：v0.11.0。2022年6月。[108] Sunil Kamath 和 Philipos Loizou。一种用于增强受有色噪声干扰语音的多频带谱减法方法。载于*IEEE* 声学、语音与信号处理国际会议论文集（*ICASSP*），2002年。[109] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本：*1412.6980*，2014年。[110] Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber。Common Voice：一个大规模多语言语音语料库。载于语言资源与评估会议论文集，第 4218–4222 页，2020年。[111] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本—语音对齐。载于*Interspeech*，第 498–502 页，2017年。[112] Arun Babu、Changhan Wang、Andros Tjandra、Kushal Lakhotia、Qiantong Xu、Naman Goyal、Kritika Singh、Patrick von Platen、Yatharth Saraf、Juan Pino 等。XLS-R：大规模自监督跨语言语音表征学习。2021年。

附录 A

```python
def uncertainty_sampling(self, samples, num_class):
        """uncertainty based sampling
        Args:
            samples ([list]): [training_list + memory_list]
        """
        self.montecarlo(samples, uncert_metric=self.uncert_metric)

        sample_df = pd.DataFrame(samples)
        mem_per_cls = self.memory_size // num_class  # kc: the number of the samples of each
class
        ret = []
        """
        Sampling class by class
        """
        for i in range(num_class):
            cls_df = sample_df[sample_df["label"] == i]
            if len(cls_df) <= mem_per_cls:
                ret += cls_df.to_dict(orient="records")
            else:
                jump_idx = len(cls_df) // mem_per_cls
                uncertain_samples = cls_df.sort_values(by="uncertainty")[::jump_idx]
                ret += uncertain_samples[:mem_per_cls].to_dict(orient="records")

        num_rest_slots = self.memory_size - len(ret)
        if num_rest_slots > 0:
            logger.warning("Fill the unused slots by breaking the equilibrium.")
            ret += (
                sample_df[~sample_df.file_name.isin(pd.DataFrame(ret).file_name)]
                    .sample(n=num_rest_slots)
                    .to_dict(orient="records")
            )

        num_dups = pd.DataFrame(ret).file_name.duplicated().sum()
        if num_dups > 0:
            logger.warning(f"Duplicated samples in memory: {num_dups}")

        return ret
```

图 5.1： 不确定性采样器的 *Python* 代码。