

基于MAML改进的少样本声音事件检测算法

陈豪杰, 杨锐, 潘善亮

(宁波大学信息科学与工程学院, 宁波 315211)

摘要: 基于深度学习的声音事件检测模型通常需要大量带标注的数据从头进行训练, 但是受到数据访问权限、使用许可以及罕见个例样本稀缺等因素制约, 获取特定任务的数据成本高昂。为应对声音事件检测中的少样本挑战, 本文在与模型无关的元学习 (Model-agnostic meta learning, MAML) 基础上提出一种模型无关且梯度平衡的元学习算法。该算法利用大量 N -way K -shot 任务训练模型, 使其具备快速学习的能力, 仅需少次梯度更新即可在 N -way K -shot 目标任务中识别未见声音事件。在外循环阶段, 多梯度下降算法被用于估计动态损失平衡因子, 促使模型关注训练难度更高的少样本任务, 从而增强模型的共享表示。本文还融入数据增强和标签平滑, 进一步降低少样本引起的过拟合。实验结果表明, 该算法在 ESC50、NSynth 以及 DCASE2020 三个数据集的 5-way 1-shot 设定中分别达到 73.56%、82.86% 以及 57.48% 准确率, 相较于改进前的 MAML 算法相对准确率提升 10% 左右。

关键词: 少样本学习; 声音事件检测; 元学习; 跨任务模型训练

中图分类号: TP391.4

文献标志码: A

Improved Few-Shot Sound Event Detection Algorithm Based on MAML

CHEN Haojie, YANG Rui, PAN Shanliang

(College of Information Science and Engineering, Ningbo University, Ningbo 315211, China)

Abstract: Sound event detection models based on deep learning typically require a substantial amount of labeled data to train from scratch. Access to task-specific data is costly due to restrictions such as data access rights, usage licenses, and the scarcity of rare individual samples. In order to address the challenge of few shot in sound event detection, this paper proposes a model-agnostic and gradient-balanced meta learning algorithm based on model-agnostic meta learning (MAML). This algorithm trains the model with a large quantities of N -way K -shot tasks, enabling it to acquire the ability of rapid learning, accurately discriminating the unheard sound event in the N -way K -shot target task with minimal gradient updates. In the outer loop stage, the multi-gradient descent algorithm is used to estimate the dynamic loss balance factor, encouraging the model to focus on few-shot training tasks that are more difficult to train, thereby enhancing the shared representation of the model. Furthermore, this paper incorporates data augmentation and label smoothing to mitigate the risk of overfitting caused by the scarcity of training samples. Experimental results demonstrate that the algorithm achieves 73.56%, 82.86% and 57.48% accuracies in the 5-way 1-shot setting on the ESC50, NSynth and DCASE2020 datasets, respectively, showing about 10% relative accuracy improvement compared to the previous MAML algorithm.

基金项目: 宁波市重点研发计划(2022Z189)。

收稿日期: 2024-04-28; **修订日期:** 2024-07-17

Key words: few-shot learning; sound event detection; meta-learning; cross-task model training

引言

声音事件检测是指通过分析和处理声音信号来识别和分析特定声音事件的技术,这些声音事件可以包括语音指令、警报声和动物叫声等。声音事件检测可以在许多领域中得以运用,比如异音检测^[1]、智慧城市和家居^[2-4]、公共区域监控^[5-6]以及野生动物监测^[7]等。

截至今日,卷积循环神经网络(Convolutional recurrent neural network, CRNN)^[8-10]混合模型仍是解决声音事件检测问题的主流模型,其中卷积神经网络(Convolutional neural network, CNN)能够提取高级特征,且具有平移不变性,但无法获取长时间的上下文信息;而循环神经网络(Recurrent neural network, RNN)能够学习长时间上下文,但无法捕获频域的不变性,二者相辅相成,一定程度上实现了性能的提升。由于声音事件的非激活时间帧比激活时间帧多,且不同声音事件的持续时间不同,这造成具有固定感受野的模型不太适合进行声音事件检测^[11]。文献[12]将特征金字塔的概念应用于CRNN,以解决固定感受野不足的问题。该方法通过特征金字塔提取低层次到高层次的特征,然后将经过上采样后的高层次特征与低层次特征进行特征融合,促使最终的特征表示既包含粗粒度的全局信息又含有细粒度的局部信息。此外,注意力机制也被广泛运用至声音事件检测中以帮助模型有选择性地关注重要信息,忽略冗余信息。文献[13]提出一种基于时间频率注意力机制和空间注意力机制的CRNN模型,其中时频注意力模块可以有效捕捉关键的时间频率特征,而空间注意力模块为特征的不同维度分配动态可学的权重,用于提取重要的特征空间信息。随着自注意力机制在各领域中展示出不俗的影响力,声音事件检测领域逐渐呈现出将RNN模型替换为Transformer的趋势^[14-16],用于解决RNN模型存在的梯度爆炸、难以建模长序列等问题。

以上所有的声音事件检测模型都需要大量带标注的数据进行从头训练,才能胜任特定的声音事件检测任务。但由于数据访问权限、使用许可及罕见个例样本稀缺等因素,获取这些任务特定的音频数据成本非常高昂。因此,在训练样本稀缺的情况下,让声音事件检测模型能够仅凭少量可学样本快速学习目标声音事件是一项巨大的挑战。

为降低少样本问题对声音事件检测模型的影响,使用数据增强或者迁移学习的方法仍是主流的解决方法。目前,生成对抗网络(Generative adversarial network, GAN)^[17-19]被广泛应用于基于声信号的故障诊断中,其通过生成器生成一些与真实样本近似的模拟数据对训练样本进行扩充。但是基于GAN的方法存在训练缓慢以及计算资源密集的问题。而一些过采样技术,比如合成少数过采样技术(Synthetic minority over-sampling technique, SMOTH)^[20],通过从 k 近邻中随机选择若干样本用于合成新的样本,降低了计算负担,但存在数据分布边缘化的问题。当用于合成的数据处于边界时会造成合成的数据同样处于边界位置,进而造成决策边界模糊的问题。此外,无标签数据^[21-22]通过最大化互信息损失得到有效利用,以此补充数据信息,增强模型性能。而迁移学习^[23-24]则通过共享可重用的声信号特征,将源域中习得的泛化知识转移到样本有限的目标域中,进而提升模型在目标域中的声音事件识别性能。然而,当源域与目标域数据分布存在巨大差距时,迁移学习会产生“负迁移”现象。尽管少样本学习在视觉领域有广泛研究,但在声音事件检测中仍处于起步阶段。文献[25]以及Heggan等^[26]探索了基于度量方法和基于元学习的声学基准表现,为本文在声音事件检测中应用元学习方法解决少样本问题提供了参考。

本文在与模型无关的元学习(Model-agnostic meta learning, MAML)^[27]的基础上提出了一种模型无关且梯度平衡的元学习算法(Model-agnostic and gradient-balanced meta learning, MAML-GB),以解

决声音事件检测领域中的样本稀缺问题。该方法通过大量 N -way K -shot 任务训练模型,使其具备快速学习的能力,从而在少样本目标任务中仅需少次梯度更新就能实现对未见声音事件的准确识别。MAML 在外循环阶段使用所有任务验证损失均值对模型进行梯度更新,但不同的训练任务之间因其难度不同存在着差异,而此种做法摊销了难度较大任务的验证损失,致使模型低估了难度较大的任务。为此,在外循环期间通过多梯度下降算法(Multi-gradient descent algorithm, MGDA)^[28]估计一组动态的损失平衡因子,让模型更专注于难度较大的训练任务,增强模型的共享表示和泛化能力。此外,本文还对支撑集使用丢帧和变速等数据增强方法,扩充训练样本数,缓解样本稀缺问题。同时,所有训练过程均采用带标签平滑的交叉熵损失,进一步降低模型的过拟合风险。

1 模型无关且梯度平衡的元学习算法

1.1 N -way K -shot 任务

在应用声音事件检测模型解决少样本目标任务之前,需要先让模型在与目标任务相似的源域中进行元学习,使其具备快速学习的能力。因此,该算法是一个与模型架构无关的跨任务参数训练算法。为了便于理解该算法,可将该算法不恰当地类比为中学生的学习历程。在学习阶段,通过几道经典的例题学习知识,进而构建整个知识框架;在实践过程中遇到某个全新的知识点时,中学生根据知识框架以及少量的例题快速学习这个全新的知识点。综上所述,测试阶段的声音事件类别与训练阶段的声音事件类别是没有交集的。

因为该算法是一种跨任务的模型训练算法,所以首先需要构建大量 N -way K -shot 任务,其中 N 代表每个任务中的声音事件类别数, K 代表每类用于模型调整的支撑样本数。样本的采集通过 N -Shot 采样器实现,具体过程如下:首先从数据集中任意选取 N 个类别,每个类别中随机挑选 K 条样本构成支撑集;然后在剩余样本中随机抽取 M 条样本用于组成查询集。一个 N -way K -shot 任务 T_i 由 1 个支撑集和 1 个查询集共同构成,其中支撑集用于快速调整模型 f_θ 以产生胜任于该检测任务的模型 f_{θ_i} ,而查询集用于评估模型在该检测任务中的快速学习效果。模型 f_θ 通过学习大量 N -way K -shot 任务,得到对局部参数变化敏感的初始参数 θ ,这使得模型在新的少样本任务中,仅需次数较少的梯度更新即可实现快速适应。

1.2 元学习总流程

模型无关且梯度平衡的元学习算法的学习过程可以看作是在目标任务上最大化损失函数相对于参数的敏感度:当敏感度足够高时,参数较小的局部变化都会引起任务损失大幅下降。这种模型训练算法与模型的形式无关,仅和模型中的参数 θ 相关。此外,还需要损失函数相对于参数 θ 足够平滑,以至于可以使用基于梯度的优化方法对模型进行参数更新。

如算法 1 所示,假定 1 个参数为 θ 的模型用函数 f_θ 表示,当该模型在 1 个新任务 T_i 上使用支撑集进行调适后,其参数由原来的 θ 变为 θ'_i ,其中音频样本 X 通过模型 f 映射为对应的声音事件类别的标签 y ,即 $y \leftarrow f(X)$ 。

算法 1 模型无关且梯度平衡的元学习(MAML-GB)算法

前提: $p(T)$ 为所有任务服从统一的数据分布; α, β 为学习率超参

- (1) 随机初始化 θ
- (2) while 不收敛 do
- (3) 采样小批次的训练任务 $T_i \sim p(T)$
- (4) for 所有 T_i do
- (5) 采样 K 条样本 $D_S^i = \{x^i, y^i\}$ 作为训练任务 T_i 的支撑集

- (6) 使用数据增广策略以增加支撑集样本量
- (7) 使用支撑集 D_i 以及带标签平滑的任务特定损失 L_{T_i} 计算梯度 $\nabla_{\theta} L_{T_i}(f_{\theta})$
- (8) 使用梯度下降算法更新参数: $\theta'_i = \theta - \alpha \nabla_{\theta} L_{T_i}(f_{\theta})$
- (9) 采样 M 条数据组成训练任务 T_i 的查询集 $D_Q^i = \{x^i, y^i\}$ 用于元更新
- (10) end for
- (11) 基于多梯度下降算法进行元更新参数 $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} \eta_i L_{T_i}(f_{\theta'_i})$
- (12) end while

为防止由少样本引起的过拟合问题,在模型训练过程中融入数据增广策略,增加支撑集的样本量。通过数据增广策略,将原始支撑集中的样本数增加了 n 倍,其中 n 表示所有使用数据增广策略的数目。

在某个 N -way K -shot 训练任务 T_i 中,使用增广后的支撑集对元模型 f_{θ} 进行一步或者多步的梯度更新后,得到一套适合于该声音事件检测任务的参数 θ'_i 。以单步梯度更新为例,有

$$\theta'_i = \theta - \alpha \nabla_{\theta} L_{T_i}(f_{\theta}) \quad (1)$$

式中 α 为一个固定的学习率超参用于更新任务内部的模型参数 θ'_i ,而这个任务内部的更新被称为内循环。

对于参数 θ 的优化则通过模型 $f_{\theta'_i}$ 在查询集 D_Q^i 上的验证损失 $L_{T_i}(f_{\theta'_i})$ 完成,而这个验证损失反映了模型在该任务中的快速学习的结果。因此,元学习的目标可以定义为

$$\min_{\theta} \sum_{T_i \sim p(T)} \eta_i L_{T_i}(f_{\theta'_i}) = \sum_{T_i \sim p(T)} \eta_i L_{T_i}(f_{\theta} - \alpha \nabla_{\theta} L_{T_i}(f_{\theta})) \quad (2)$$

式中 η_i 为一组动态的损失平衡因子。

再次强调,元学习是在大量 N -way K -shot 任务 T_i 上训练,以便得到一套对局部参数变化敏感的初始参数 θ ,因此需要算法通过带权累和各训练任务上的验证损失 $L_{T_i}(f_{\theta'_i})$,更新模型 f_{θ} 的参数 θ 。对于模型参数 θ 的优化可表示为

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} \eta_i L_{T_i}(f_{\theta'_i}) \quad (3)$$

式中 β 代表了元学习的学习率,用于任务外部的训练,而这次梯度更新被称为外循环。

由于整个训练过程涉及2次梯度更新,为了进行区别,本文将这个任务外部的更新定义为元更新。算法1中所涉及到的所有训练过程都采用带标签平滑的交叉熵损失函数,以更新计算模型参数,其可表示为

$$L(y, X) = - \sum_{i=1}^N y_i \log_2 \frac{\exp(X_i)}{\sum_{j=1, j \neq i}^N \exp(X_j)} \quad (4)$$

式中独热编码的硬标签 y_i 经标签平滑后转为了软标签,以降低过拟合风险

$$y_i = \begin{cases} 1 - \epsilon & y = i \\ \epsilon / (N - 1) & \text{其他} \end{cases} \quad (5)$$

式中 ϵ 代表平滑因子。标签平滑通过施加正则的方法将独热标签转变为更平滑的概率分布,降低了模型对某个样本的置信度,从而让模型能够更鲁棒地处理噪声或错误标签。

遵照常规的模型训练方法,选取一些 N -way K -shot 训练任务以组成1次小批量的训练,而1个小批量的训练在少样本学习领域则通常被称为一个episode的学习。模型被训练至收敛后,将其迁移到 N -way K -Shot 的目标任务中使用支撑集进行少次梯度更新以适应该目标任务的检测。图1为基于MAML改进的元学习算法流程图。

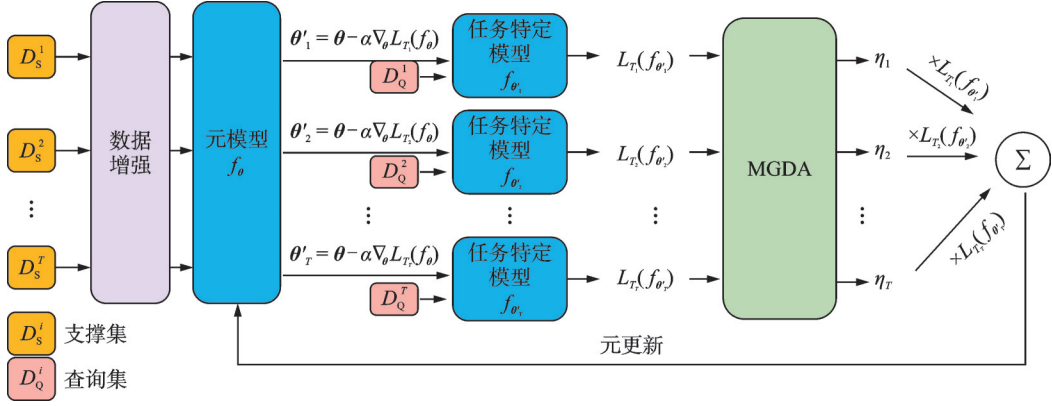


图1 基于MAML改进的元学习算法流程图

Fig.1 Flowchart of an improved meta-learning based on MAML

1.3 多梯度下降算法

MAML算法在进行元更新时,使用所有训练任务的损失均值对参数 θ 进行梯度更新。然而不同训练任务其难易程度是不尽相同的(由验证损失 $L_{T_i}(f_{\theta'_i})$ 体现),因而此种做法推销了某个训练任务上较大的验证损失,致使模型低估了难度较大的训练任务。而多梯度下降算法的使用克服了这一缺陷:整个模型训练的过程可以不完全恰当地类比为中学生的刷题历程,在学习初期需要通过大量的训练任务以构建整个知识框架,而在学习的后期需要将更多精力放在错题集上,即验证损失 $L_{T_i}(f_{\theta'_i})$ 较大的训练任务上,以提升学习效率。因此,本文通过多梯度下降算法估计1组动态损失平衡因子 $\eta = \{\eta_1, \eta_2, \dots, \eta_i | \sum_{i=1}^T \eta_i = 1\}$,让模型能够更专注于难度较高的训练任务,进而构建更具代表性的共享任务表示,提升模型在各种任务上的泛化性能。

MGDA首先需要以Karush-Kuhn-Tucker(KKT)作为必要条件,即: u 存在一组平衡因子 $\eta_1, \eta_2, \dots, \eta_i \geq 0$,使得 $\sum_{i=1}^T \eta_i = 1$ 且 $\nabla_{\theta} \sum_{i=1}^T \eta_i L_{T_i}(f_{\theta'_i}) = 0$; u 对于所有训练任务 $T_i \sim p(T)$ 而言,满足 $\nabla_{\theta} L_{T_i}(f_{\theta'_i}) = 0$ 。

显然,任意一套解集 η 都能使基于深度学习的声音事件检测模型习得一套最优的参数 θ^* ,使得在所有训练任务上的任务特定损失 $L_{T_i}(f_{\theta'_i})$ 达到最小。

如式(2)所示,元学习的目标足够平滑,因此自然而然地可以使用基于梯度优化的方法寻找1个距离最优参数 θ^* 最近的局部最优解。对于 η 而言,式(2)是一个多元的线性约束求解问题。由于多元问题的求解较为复杂,因此首先考虑两个训练任务之间平衡问题,进而实现从特殊到一般的推广,而两个训练任务间的优化问题可定义为

$$\min_{\theta} \left\| \eta \nabla_{\theta} L_{T_1}(f_{\theta'_1}) + (1 - \eta) \nabla_{\theta} L_{T_2}(f_{\theta'_2}) \right\|_2^2 \quad \eta \in [0, 1] \quad (6)$$

对于平衡因子 η 而言,这是一个一元二次函数求最小值的问题,因此对 η 求偏导,并令偏导数为0,可求得 η 的解为

$$\hat{\eta} = \left[\frac{(\nabla_{\theta} L_{T_2}(f_{\theta'_2}) - \nabla_{\theta} L_{T_1}(f_{\theta'_1}))^T \nabla_{\theta} L_{T_2}(f_{\theta'_2})}{\left\| \nabla_{\theta} L_{T_1}(f_{\theta'_1}) - \nabla_{\theta} L_{T_2}(f_{\theta'_2}) \right\|_2^2} \right]_{+1} \quad (7)$$

其中 $[\cdot]_{+1}$ 代表使用 $\max(\min(\hat{\eta}, 1), 0)$ 将 $\hat{\eta}$ 的值限制在 $[0, 1]$ 之间。

算法 2 寻找 θ'_1 和 θ'_2 两个梯度平衡的方向: $\min_{\gamma \in [0,1]} \|\gamma \theta'_1 + (1-\gamma) \theta'_2\|_2^2$ (γ 为平衡因子)

(1) if $\theta_1^T \theta'_2 \geq \theta_1^T \theta'_1$ then

(2) $\gamma = 1$

(3) else if $\theta_1^T \theta'_2 \geq \theta_2^T \theta'_2$ then

(4) $\gamma = 0$

(5) else

(6) $\gamma = \frac{(\theta'_2 - \theta'_1)^T \theta'_2}{\|\theta'_2 - \theta'_1\|_2^2}$

end if

算法 3 MGDA 的更新程式

(1) $\eta_1, \eta_2, \dots, \eta_T = \text{FrankWolfeSolver}(\theta)$ // 寻找一个共同的梯度方向

(2) $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} \eta_i L_{T_i}(f_{\theta'_i})$ // 元更新 θ

(3) FrankWolfeSolver(θ) 的程式

初始化 $\eta = (\eta_1, \eta_2, \dots, \eta_T) = \left(\frac{1}{T}, \frac{1}{T}, \dots, \frac{1}{T}\right)$

预计算梯度矩阵 M , 即 $M_{i,j} = (\nabla_{\theta} L_{T_i}(f_{\theta'_i}))^T \nabla_{\theta} L_{T_j}(f_{\theta'_j})$

循环

(4) $\hat{t} = \underset{r}{\operatorname{argmin}} \sum_t \eta_t M_{r,t}$

(5) $\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} ((1-\gamma)\eta + \gamma e_i)^T M ((1-\gamma)\eta + \gamma e_i)$ // 使用算法 2

(6) $\eta = (1-\hat{\gamma})\eta + \hat{\gamma} e_i$ // e 为单位矩阵

(7) 直至 $\hat{\gamma} \sim 0$ 或 迭代次数耗尽

(8) 返回 $\eta_1, \eta_2, \dots, \eta_T$

(9) 结束程式

FrankWolfe 算法被广泛用于求解多元线性约束的问题, 因此本文将式(6)的结果式(7)作为 FrankWolfe 算法的判定子程序, 并通过 FrankWolfe 算法的行搜索, 求解多元的损失平衡因子 η 。在预计算任务间的梯度矩阵 M 时, 涉及到梯度的二阶求导 $\frac{\partial L_{T_i}(f_{\theta'_i})}{\partial \theta'_i} \frac{\partial \theta'_i}{\partial \theta}$, 为了降低计算的复杂度, 本文遵循 MAML 的做法使用一阶导进行近似。

2 实验条件

2.1 数据集

为了验证本文所提出方法的广泛性以及可重复性, 选取 ESC50、NSynth 以及 DCASE202 Task5 三个数据集进行同数据集内元训练与元测试, 使用 UrbanSound8K 数据集进行跨数据集验证, 数据集的详细信息如表 1 所示。所有实验均采用为 5-way 1-shot 设置, 并使用固定的 7/1/2 类别划分, 以保证实验的可重复性。在数据预处理方面, 本文使用固定的 16 kHz 采样率以及 Fbank 超参 (64 ms 窗长、32 ms 的帧移以及 128 个 Mel 滤波器)。此外, 为了增加模型训练的稳定性, 对输入样本进行了全局规范化。

表 1 数据集的详细信息
Table 1 Detail information of datasets

数据集	类型	类别数/个	样本数/个	音频长度/s	用途
ESC50	环境音	50	2 000	5	元学习/元测试
NSynth	乐器	1 006	305 978	4	元学习/元测试
DCASE2020_task5	野生动物	47	7 574	≤ 1	元学习/元测试
UrbanSound8K	环境音	10	8 732	≤ 4	元测试

2.2 基准方法

为验证提出方法的有效性,本文选择了一些经典的少样本学习方法作为baseline进行对比:

(1)基于元学习的方法。(a)MAML^[27]:通过跨任务训练让模型掌握快速学习的能力,从而在新任务中仅需少量调整以适应该任务。(b)Meta-Curvature^[29]:在MAML的基础上使用曲率信息更准确地估计参数的更新方向,避免在陡峭的损失函数曲面上出现震荡或错过最佳参数值的情况。

(2)基于距离度量的方法。(a)原型网络^[30]:将样本映射至统一的特征空间后,在该特征空间中使用距离函数或者相似度对样本进行归类。(b)SimpleShot^[31]:在原型网络基础上使用中心L2范数准则(Centering L2-normalization, CL2N)对样本特征向量进行归一化,使得各特征向量在更加公平可靠的特征空间内进行比较。(c)Meta-Baseline^[32]:先对模型进行预训练,让其在掌握大量的通用知识的基础上再进行少样本学习。

2.3 实验超参设定

在所有实验中,均采用Adam优化器更新模型参数,其中,元学习率 α 的初始值被设为0.001,而任务内部学习率 β 被设定为0.4。此外,使用余弦退火算法,以衰减元学习率 α 。使用200个周期的迭代,保证元学习收敛。在验证与测试期间,本文还使用速度变换和丢帧对支撑集进行数据增广,使得支撑集的样本量增加了2倍。此外,所有训练过程均采用平滑因子为0.1的交叉熵损失对模型进行训练。

本文主要研究少样本方法的有效性,为降低模型之间差异的影响以及计算资源开销,所有实验均采用轻量级的CRNN神经网络架构。该网络由4层卷积层以及1层单向RNN组成,隐空间大小统一为64。最后全连接层的输出设置为 N -way,而对于度量学习而言,最后的输出大小为64。

2.4 评价指标

在本文实验中,采用了两种指标对声音事件检测效果进行评价,分别为准确率(Accuracy, ACC)以及均值平均精确率(mean Average precision, mAP)。其中,准确率计算公式为

$$\text{ACC} = \frac{1}{T} \sum_{i=1}^T \frac{N_T^i}{N \times M} \quad (8)$$

式中: N_T^i 表示测试集的任务 i 中预测正确的样本数, N 表示类别数, M 表示每类中查询样本数, T 表示总的测试任务数。

仅仅使用准确率作为评价指标是不科学的,因为该指标无法区分声音事件检测的错误类型。而在多分类任务时,使用mAP能够综合评估模型在多个类别上的性能,而不是仅仅关注单个类别的准确率。mAP计算公式为

$$\text{mAP} = \frac{1}{N \times T} \sum_{i=1}^T \sum_{j=1}^N \frac{N_{\text{TP}}^j}{(N_{\text{TP}}^j + N_{\text{FP}}^j)} \quad (9)$$

式中: N_{TP}^j 表示第 j 类中预测正确的正例数, N_{FP}^j 表示第 j 类的假阳性样本数。

3 实验结果与分析

3.1 与其他少样本学习方法的对比实验

本文分别在ESC50、NSynth以及DCASE2020任务5三个数据集上,与经典的少样本学习方法进行对比,验证MAML-GB解决少样本问题的有效性与广泛性。为保证实验的公平性,所有数据集均采用固定的类别划分,并且所有方法都是在同数据分布内进行元学习和元测试。观察比较分析表2,可以得到:

(1)所有少样本学习的方法在NSynth数据集上的表现普遍优于ESC50数据集上的表现。出现这种情况的原因是ESC50数据集相比于NSynth数据集有着更强的背景噪音,这一定程度上会对模型表现产生影响。

(2)在DCASE2020任务5数据集中,所有的少样本学习的表现都产生了较大的下降。产生这一现象的原因一方面是类内的相似性,同一属类的野生动物之间的声音样本往往具有较高的类内相似性,这需要模型捕捉一些更细粒度的声学特征才能进行区分;另一方面是同一物种的叫声受到环境、情绪等因素的影响,存在着较大差异,这增加了模型学习和泛化的难度。

(3)本文在MAML的基础上进行改进,在外循环阶段使用多梯度下降算法估计一组动态损失平衡因子,促使模型更专注于难度较大的训练任务,进而增强模型的共享任务表示,提升模型在各种少样本任务中的泛化性能。此外,还融入了数据增强以及标签平滑策略,进一步降低过拟合风险。通过在众多数据集上与MAML进行比较,证明改进方法的有效性。

表2 同数据集内测试以及跨数据集验证
Table 2 Test within and across datasets %

阶段	数据集	指标	MAML	Meta-curvature	Prototypical network	Simpleshot (CL2N)	Meta-baseline	MAML-GB (Ours)
Train	ESC50	ACC	70.14	72.17	66.48	69.56	70.10	73.56
		mAP	85.82	86.31	85.78	86.57	85.71	87.46
	NSynth	ACC	81.46	82.20	94.29	85.31	79.60	82.86
		mAP	92.85	92.77	98.70	94.57	92.45	93.34
	DCASE2020 Task5	ACC	52.46	54.16	55.50	52.12	51.86	57.48
		mAP	72.62	73.27	77.25	76.77	75.07	76.60
Cross	Urban	ACC	56.80	57.40	55.80	51.20	54.00	59.20
	Sound8K	mAP	76.60	75.23	76.31	70.85	73.23	77.90

3.2 跨数据集验证

如表2所示,本文还拓展元测试的数据集范围,以评估模型在不同数据分布上的泛化能力。通过观察分析表中的结果,可以发现将在源域中训练的模型迁移至不同数据分布的目标域后,模型的表现都会不同程度地下降。产生这种现象的原因可归结为:(a)源域与目标域数据分布差异大。当模型在源域中进行元学习时,由于源域数据多样性不足或者源域数据分布与目标域之间数据分布差距大,导致模型在迁移至少样本目标域中表现下降。(b)模型容量不够。实验中采用的是轻量级的CRNN网络结构,这种简单神经网络在数据多样性强的源域难以进行适应,致使提取的特征表示不够强。

3.3 标签平滑因子调优

标签平滑是一种正则化的技术,用于防止模型过拟合训练数据。该技术将真实标签为独热编码的硬标签转为了软标签,降低了模型对于过于确定标签的置信度以及噪声标签的敏感性,极大地提高了模型的泛化性,缓解了过拟合问题。然而,过小的标签平滑因子无法起到很好的正则作用,而过大的平滑因子则会影响模型的检测性能。因此,调优一个合适的标签平滑因子至关重要。本节在0.05~0.3范围内对标签平滑因子进行超参调优实验。如图2所示,当标签平滑因子设置为0.1时检测效果最好。

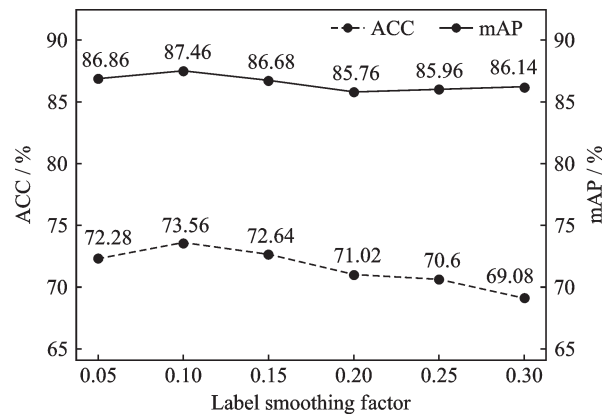


图2 标签平滑因子调优

Fig.2 Tuning of label smoothing factor

3.4 数据增广添加范围的影响

数据增广是一种常用且有效的策略,用于缓解样本稀缺问题。在元学习中,本文探索了不同的数据增广添加范围对元学习效果的影响。具体而言,数据增广的添加范围可以分为3种:仅在元训练期间添加、仅在元测试期间添加以及二者都添加。观察分析图3可得到:

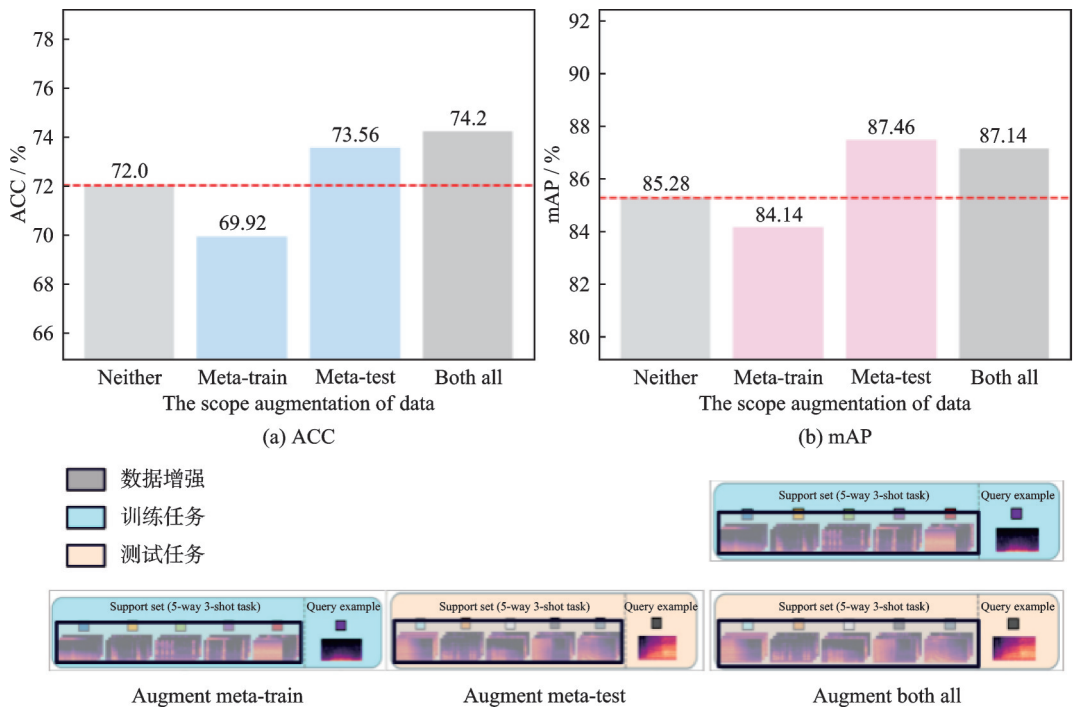


图3 数据增广添加范围的影响

Fig.3 Impact of scope of data augmentation

(1)根据直觉,仅在元测试期间进行数据增广可以让模型在经过相对严苛的训练后,在训练条件相对宽泛的元测试环境中表现更出色。如图3上的Meta-test的结果所示,可以发现这种添加方式可以缓解元测试期间的数据稀缺问题。

(2)本文也探索了在元训练和元测试期间对支撑集进行数据增强的影响。从图3中Both all的表现来看,相较于仅在Meta-test上增强,这种做法有轻微的性能提升。然而,在元训练期间需要频繁地对数据进行在线增广,导致训练时间延长。鉴于收益较低,最终本文选择仅在元测试期间对支撑集数据进行增广。

(3)本文还探索了仅在元训练期间进行数据增广。由于元学习对源域与目标域数据分布的一致性有严重依赖,这种数据增广添加方式会对模型在元测试期间的表现造成损害。在元训练期间引入增广后的样本致使源域数据分布出现轻微偏移。当元模型在与目标域数据分布有偏的源域上训练后,其在迁移到与源域数据分布不一致的目标域时自然而然地会出现性能衰退。

3.5 不同改进策略的消融实验

如图4所示,其横坐标w/o L.S.表示不使用标签平滑;w/o D.A.表示不使用数据增强;w/o MGDA表示不使用多梯度下降算法。通过观察图4可知,在MAML-GB的基础上通过不断的变化所使用的改进策略,对模型的表现产生了不同程度的影响。其中,作为主要的改进策略MGDA对模型的表现影响最为显著,这从侧面证明了在元训练期间建立各任务间的联系的重要性,其有助于元模型构建更具代表性的共享任务表示。

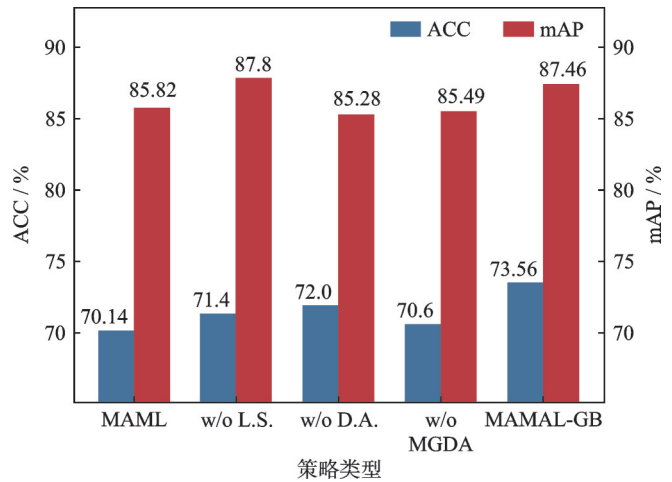


图4 不同改进策略对模型表现的影响

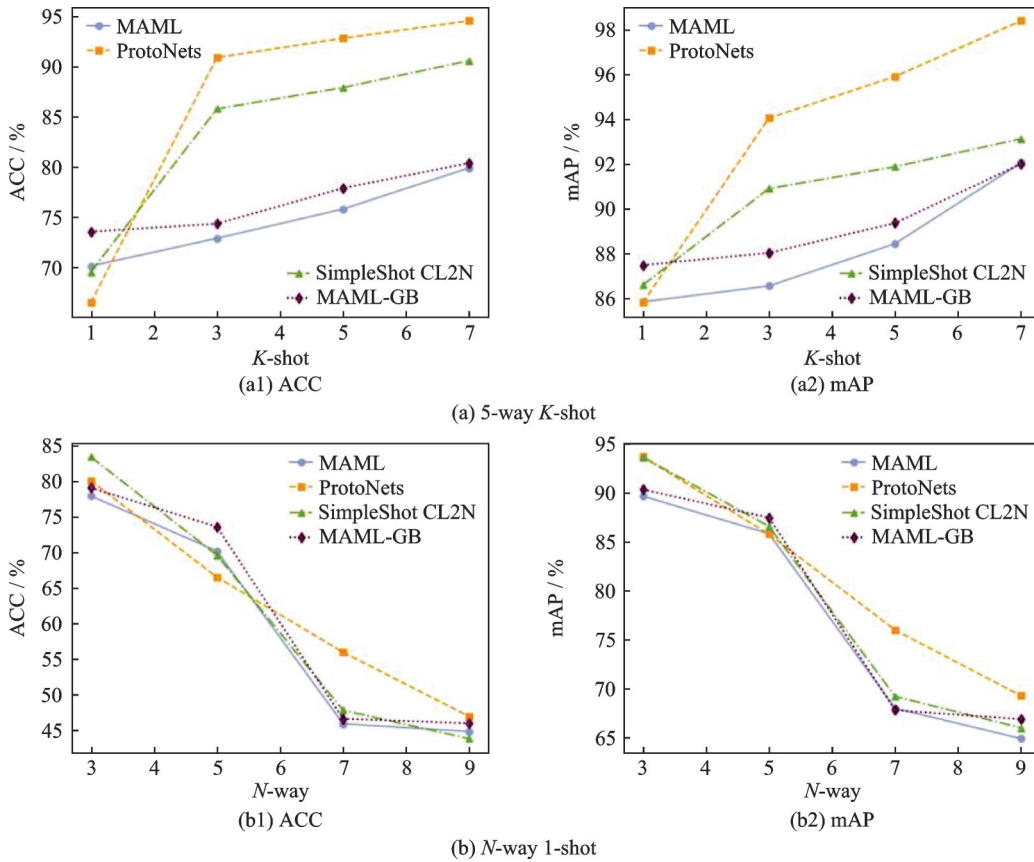
Fig.4 Impact of different improvement strategies on model's performance

3.6 N-way K-shot 分析

之前的实验中均在5-way 1-shot的设定

下进行,但本文也研究了更多的支撑样本数以及类别数对少样本学习方法的影响。因此,本小节在不同的N-way K-shot设定下对模型的表现进行了实验,实验结果如图5所示。在图5(a)中,本文固定类别数为5,并从1~7变化支撑样本数。结果显示,所有少样本学习方法在支撑集样本数增加时都有不同程度的性能提升。同时,当支撑样本数达到3张以上时,基于度量的少样本学习方法趋于收敛,而基于元学习的少样本学习方法仍呈缓慢上升趋势。此外,在样本数极度稀缺的情况下(1-shot),本文所提出的MAML-GB方法表现更出色,但随着样本数的扩充,这种优势逐渐被基于度量的方式所抹平。对此,本文认为是因为随着样本数的增加,基于度量的少样本学习方法能提取更具代表性的原型向量;相反,元学习方法对于支撑集训练样本量的大小并不敏感,因为元模型在进行元训练期间是针对少样本子任务进行的,其目的在于构建任务间通用的共享表示。

同样地,在图5(b)中,本文固定支撑集样本数为1,并从3~9变化类别数,以观察不同类别数下对少样本学习方法的影响。结果显示,随着类别数不断地增多,所有少样本学习方法的表现都有不同程度的下降。分析其成因:一方面是因为随着类别数的增加,类别间的差异性以及相似性也会增加,这要求模型能够捕捉更细粒度的特征以进行更好的判别;另一方面,随着类别数的增加,要求模型需要更高的容量来适应更复杂的类别结构,而在实验中本文采用是轻量级的CRNN网络,这种简单结构的神经网络由于模型本身容量有限而难以充分捕捉类别间的差异。

图5 在ESC50数据集上的 N -way K -shot分析Fig.5 N -way K -shot analysis on the ESC50 dataset

4 结束语

本文提出了一种基于MAML改进的元学习算法MAML-GB,旨在解决声音事件检测中的少样本问题。该算法利用大量的 N -way K -shot任务对模型进行训练,使其具备快速学习的能力。与传统的MAML算法相比,本文的MAML-GB算法在训练过程中通过多梯度下降算法估计一组动态损失平衡因子,使模型更关注训练难度较大的任务,增强模型的共享表示。此外,还融入数据增强和标签平滑策略,进一步降低过拟合风险。本文内容为元学习提供了新的视角,增强了模型在少样本任务中的适应能力。在应用场景中,该算法在少样本声音事件检测领域具有重要价值,比如野生动物监测、异音诊断等。在众多数据集上的对比结果表明,MAML-GB算法在处理少样本声音事件检测任务时表现出更高的准确率、鲁棒性以及普适性。具体而言,在ESC50、NSynth和DCASE2020数据集的5-way 1-shot设定中分别达到了73.56%、82.86%和57.48%的准确率,相较于传统的MAML算法约有10%的相对准确率提升。尽管该算法取得了显著的成果,本文仍存在一些未解决的问题。例如,元学习要求源域和目标域任务相似性高。但当二者的数据分布差异较大时,直接将训练好后的模型迁移到少样本目标域可能会产生“负迁移”问题,而解决这一问题需要使用域适应等技术。

参考文献:

- [1] KOIZUMI Y, SAITO S, UEMATSU H, et al. Unsupervised detection of anomalous sound based on deep learning and the

- Neyman-Pearson Lemma[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, 27(1): 212-224.
- [2] TAN E L, KARNAPI F A, NG L J, et al. Extracting urban sound information for residential areas in smart cities using an end-to-end IoT system[J]. *IEEE Internet of Things Journal*, 2021, 8(18): 14308-14321.
- [3] PANDYA S, GHAYVAT H. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence[J]. *Advanced Engineering Informatics*, 2021, 47: 101238.
- [4] 李玲俐. 家庭保健监测系统中环境声音事件的识别[J]. *重庆师范大学学报(自然科学版)*, 2016, 33(4): 118-122.
LI Lingli. Environmental sound recognition for a family health monitoring system[J]. *Journal of Chongqing Normal University (Natural Science)*, 2016, 33(4): 118-122.
- [5] 张丽君. 公共场所异常声音识别算法设计与研究[D]. 重庆: 重庆大学, 2017.
ZHANG Lijun. Algorithm design and research on abnormal sound recognition in public places[D]. Chongqing: Chongqing University, 2017.
- [6] ARSLAN Y, CANBOLAT H. Sound based alarming based video surveillance system design[J]. *Multimedia Tools and Applications*, 2022, 81(6): 7969-7991.
- [7] TANG T T, LIANG Y H, LONG Y H. Two improved architectures based on prototype network for few-shot bioacoustic event detection[EB/OL]. (2021-06-10). https://dcase.community/documents/challenge2021/technical_reports/DCASE2021_Tang_54_task5.pdf.
- [8] CAKIR E, PARASCANDOLO G, HEITTOLA T, et al. Convolutional recurrent neural networks for polyphonic sound event detection[J]. *IEEE / ACM Transactions on Audio, Speech, and Language Processing*, 2017, 25(6): 1291-1303.
- [9] WANG Y B, ZHAO G H, XIONG K, et al. Multi-scale and single-scale fully convolutional networks for sound event detection[J]. *Neurocomputing*, 2021, 421: 51-65.
- [10] WANG Y B, ZHAO G H, XIONG K, et al. MSFF-Net: Multi-scale feature fusing networks with dilated mixed convolution and cascaded parallel framework for sound event detection[J]. *Digital Signal Processing*, 2022, 122: 103319.
- [11] KOH C Y, CHEN Y S, LIU Y W, et al. Sound event detection by consistency training and pseudo-labeling with feature pyramid convolutional recurrent neural networks[C]//*Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway, NJ: IEEE, 2021: 376-380.
- [12] IMOTO K, MISHIMA S, ARAI Y, et al. Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance [J]. *Applied Acoustics*, 2022, 196: 108882.
- [13] IN Y, WANG M, LUO L Y, et al. Polyphonic sound event detection using temporal-frequency attention and feature space attention[J]. *Sensors*, 2022, 22(18): 6818.
- [14] MIYAZAKI K, KOMATSU T, HAYASHI T, et al. Weakly-supervised sound event detection with self-attention[C]//*Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway, NJ: IEEE, 2020: 6670.
- [15] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented Transformer for speech recognition[C]//*Proceedings of Interspeech*. [S.l.]: [s.n.], 2020: 5036-5040.
- [16] KOUTINI K, SCHLÜTER J, EGHBAL-ZADEH H, et al. Efficient training of audio transformers with patchout[C]//*Proceedings of Interspeech*. [S.l.]: [s.n.], 2022: 2753-2757.
- [17] COOPER C, ZHANG J J, GAO R X, et al. Anomaly detection in milling tools using acoustic signals and generative adversarial networks[J]. *Procedia Manufacturing*, 2020, 48: 372-378.
- [18] HATANAKA S, NISHI H. Efficient GAN-based unsupervised anomaly sound detection for refrigeration units[C]//*Proceedings of the 30th IEEE International Symposium on Industrial Electronics*. Kyoto, Japan: IEEE, 2021.
- [19] TAGAWA Y, MASKELIŪNAS R, DAMAŠEVIČIUS R. Acoustic anomaly detection of mechanical failures in noisy real-life factory environments[J]. *Electronics*, 2021, 10(19): 2329.
- [20] MARTIN-DIAZ I, MORINIGO-SOTELO D, DUQUE-PEREZ O, et al. Early fault detection in induction motors using AdaBoost with imbalanced small data and optimized sampling[J]. *IEEE Transactions on Industry Applications*, 53(3): 3066-3075.
- [21] YANG D, WANG H, YE Z, et al. Few-shot bioacoustic event detection: A good transductive inference is all you need[EB/

- OL]. [2024-03-16]. https://dcase.community/documents/challenge2021/technical_reports/DCASE2021_Zou_22_task5.pdf.
- [22] BANOORI F, IJAZ N, SHI J, et al. Few-shot bioacoustics event detection using transductive inference with data augmentation [J]. IEEE Sensors Letters, 2024, 8(3): 1-4.
- [23] LIU W S, CHEN Z, ZHENG M H. An audio-based fault diagnosis method for quadrotors using convolutional neural network and transfer learning[C]//Proceedings of the 2020 American Control Conference (ACC). Denver, USA: IEEE, 2020: 1367-1372.
- [24] HASAN M J, ISLAM M M M, KIM J M. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions[J]. Measurement, 2019, 138: 620-631.
- [25] SHI B, SUN M, PUVVADA K C, et al. Few-shot acoustic event detection via meta learning[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 76-80.
- [26] HEGGAN C, BUDGETT S, HOSPEDALES T, et al. MetaAudio: A few-shot audio classification benchmark[C]//Proceedings of Artificial Neural Networks and Machine Learning—ICANN 2022. Cham: Springer International Publishing, 2022: 219-230.
- [27] FINN C, ABBEEL P, LEVINE S, 2017. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, NSW, Australia: JMLR, 1126-1135.
- [28] SENER O, KOLTUN V. Multi-task learning as multi-objective optimization[C]//Proceedings of the 31rd International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2018: 525-536.
- [29] PARK E, OLIVA J B. Meta-curvature[C]//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc., 2019: 3314-3324.
- [30] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2017.
- [31] WANG Y, CHAO W L, WEINBERGER K Q, et al. SimpleShot: Revisiting nearest-neighbor classification for few shot learning[EB/OL]. (2019-05-23)[2024-03-16]. <http://abs/1911.04623>.
- [32] CHEN Y, LIU Z, XU H, et al. Meta-baseline: Exploring simple meta-learning for few-shot learning[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 9042-9051.

作者简介:



陈豪杰(1999-),男,硕士研究生,研究方向:语音信号处理,E-mail:765835937@qq.com。



杨锐(1999-),男,硕士研究生,研究方向:语音信号处理,E-mail:yangraycn@163.com。



潘善亮(1970-),通信作者,男,教授,研究方向:服务计算、大数据处理、自然语言处理,E-mail:panshanliang@nbu.edu.cn。

(编辑:刘彦东)