arXiv:2305.03391v1 [cs.SD] 5 May 2023

基于图中心性的音频卷积神经网络滤波器剪枝

James A King[+], *Arshdeep Singh*[+], *Mark D. Plumbley*,

萨里大学视觉、语音与信号处理中心（CVSSP），英国。

摘要

卷积神经网络（CNN）常见于音频分类等许多现实问题的高性能解决方案中。卷积神经网络拥有大量参数和滤波器，其中一些对性能的影响比其他滤波器更大。这意味着网络中可能包含许多不必要的滤波器，它们会增加卷积神经网络的计算和内存需求，却只能带来有限的性能收益。为提高卷积神经网络的效率，我们提出了一种剪枝框架，用于移除“共性”最高的滤波器。我们使用图论概念中心性来衡量这种共性。我们假设，中心性高的滤波器应当被移除，因为它代表了共性，可以由其他滤波器替代，而不会对网络性能产生太大影响。我们在声学场景分类和音频标签分类任务上对所提出的框架进行了实验评估。在 DCASE 2021 任务1A 基准网络上，与原始网络相比，我们的所提方法在准确率下降不到两个百分点的情况下，将每次推理的计算量减少了 71%，参数数量减少了 50%。对于专为音频标签分类设计的 PANNs 等大规模卷积神经网络，我们的方法在性能略有提升的同时，将每次推理的计算量减少了 24%，参数数量减少了 41%。

索引词— 卷积神经网络、剪枝、音频分类、PANNs、DCASE。

1. 引言

卷积神经网络（CNN）在多种音频任务中展现出了良好的效果，包括语音识别[1], 音乐分析 [2]和音频分类 [3, 4]。通常，卷积神经网络包含许多层，例如卷积层和池化层。每一层都包含一些参数，包括权重、偏置和滤波器，这些参数都是针对特定问题通过优化过程学习得到的。卷积神经网络通常包含大量参数，因此需要大量内存存储。此外，推理过程中的大部分计算都由卷积运算完成；卷积运算通过在输入数据上滑动一组滤波器来生成特征图。当输入数据量大且滤波器数量多时，这种计算尤其耗时。尽管卷积神经网络在解决非线性复杂任务方面非常有效 [5], ，但推理过程中对大量内存和繁重计算的需求，成为其部署到移动电话或物联网（IoT）设备等资源受限设备上的瓶颈 [6]。此外，机器学习（ML）模型长期需要大量计算，这会显著增加 CO_2 排放，使得规模较大的卷积神经网络不利于环境。例如，用于训练机器学习模型的现代图形处理器（如 NVIDIA RTX-2080 Ti 图形处理器）运行 48 小时所产生的二氧化碳排放量，相当于一辆普通汽车行驶 13 英里[l]所排放的二氧化碳[2]。因此，减少卷积神经网络的计算量和内存需求这一问题，已引起研究界的广泛关注。

近期压缩卷积神经网络的研究涉及滤波器剪枝方法 [7], ，这些方法会移除对性能贡献最小的不重要滤波器。这些方法通过主动或被动方法来衡量滤波器的重要性。主动方法 [8, 9] 使用数据集生成滤波器的特征图，然后利用熵、秩或特征图上的零值平均百分比等多种指标衡量滤波器的重要性。一些主动方法甚至会在卷积神经网络训练期间识别重要滤波器：为每个滤波器添加软掩码等额外参数，然后联合优化卷积神经网络参数和软掩码 [10, 11]。相较之下，被动方法无需数据，仅使用滤波器本身来量化其重要性。因此，与主动滤波器剪枝相比，被动方法更易于应用，且所需存储空间显著更少；这对较大型模型尤为重要。

被动滤波器剪枝方法可分为基于范数的方法 [12],（计算滤波器的 l_1 或 l_2 范数以确定其重要性），以及基于相似度的方法 [13], （移除相似的滤波器）。基于范数的方法遵循“范数越小、重要性越低”的准则，会移除范数较小的滤波器。然而，移除范数较小的滤波器可能会忽略网络中学到的多样性，以及高范数滤波器中的冗余。基于相似度的剪枝方法能够捕捉这种多样性，并根据滤波器之间的成对相似性移除冗余滤波器[13]。与基于范数的方法相比，这类基于相似度的方法性能更好。然而，成对相似性方法仅考虑成对滤波器之间的相似性来移除冗余滤波器，而最相似的滤波器对可能有所不同。忽略这类滤波器可能会减少网络中学到的有用多样化信息。

本文提出一种被动滤波器剪枝方法：如果其他滤波器可以替代某个滤波器，则认为该滤波器存在冗余。为衡量其冗余程度，我们将滤波器视为图中的节点，并确定每个节点的中心性。节点中心性较高，表示该节点与其他任意两个节点之间具有较高的共性。通过对中心性进行排序，我们能够更好地了解移除这类共性滤波器对网络的影响。相比之下，先前的方法 [13],仅衡量最相似滤波器对之间的共性，而不考虑它们与其他滤波器之间的共性。

本文其余部分安排如下。第 2 节介绍剪枝方法的背景。第 3 节描述所提出的用于识别相似滤波器的方法。接下来，第 4 节介绍实验设置。结果与分析见第 5 节。最后，第 6 节和第 7 节分别给出讨论与结论。

[+]两位作者贡献相同。[l]机器学习

[2] CO_2 估计器

2. 背景与相关工作

2.1. 剪枝问题

滤波器剪枝问题是通过消除部分滤波器、同时保持性能来缩减网络规模的主要方法之一。给定一个具有 L 个卷积层的卷积神经网络（CNN），其中每层都包含一组滤波器 $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$，目标是找到一个剪枝后的卷积神经网络，使层 L' 具有缩减后的滤波器集合 $\mathcal{F}' \subseteq \mathcal{F}$，从而移除 p% 的滤波器。

$$\min_{\mathcal{F}' \subseteq \mathcal{F}, |\mathcal{F}'| = \lceil (1-p)|\mathcal{F}| \rceil} \mathcal{C}(L') \quad \text{s.t.} \quad \mathcal{P}(L') \gtrapprox (1-\epsilon)\mathcal{P}(L) \tag{1}$$

其中，ϵ 是一个较小的容差值。因此，$|\mathcal{F}'| = \lceil (1-p)|\mathcal{F}| \rceil$，性能指标 $\mathcal{P}(L')$ 仍接近于 $\mathcal{P}(L)$。此外，ϵ 取负值的情况也并不少见，因此剪枝能够提升性能。

为了选出少量重要的滤波器，$\mathcal{F}'$，使用主动或被动滤波器剪枝方法计算滤波器的重要性。在获得剪枝网络后，会执行微调过程，以恢复大部分性能。该微调过程会使用原始数据重新训练剪枝网络。

2.2. 计算卷积神经网络滤波器重要性的方法

主动方法： 主动滤波器剪枝方法是数据驱动的，因此可以利用 $\mathcal{P}(L')$的评估结果来决定剪除或保留哪些滤波器。例如，先前的研究会在训练过程中，通过联合学习卷积神经网络及额外参数（如与每个卷积神经网络参数相关联的软掩码）来计算滤波器的重要性 [11, 14]。然而，这些方法会使训练时间增加最多 10 倍 [15, 16]，且计算成本高。其他主动滤波器剪枝方法会生成与一组样本相对应的特征，然后应用秩 [8], 能量 [9], 零值平均百分比[17] 等指标来量化滤波器的重要性，或在特征图上应用聚类 [18] 等相似性度量，以消除与冗余特征图相对应的滤波器。然而，这些方法需要额外的内存资源来获取特征图，并且需要复杂的训练流程来优化软掩码等额外参数，尤其是在将大规模预训练网络用于下游任务时。

被动方法： 被动方法无需数据，也不会在滤波器选择过程中评估 $\mathcal{P}(L')$。因此，与主动方法相比，被动方法具有更好的可扩展性和更高的效率。例如，Li 等人 [12] 计算滤波器的 l_1-范数（即滤波器参数的绝对值之和），以量化滤波器的重要性，并剔除低范数滤波器，从而获得剪枝后的网络。He 等人 [19] 计算滤波器的几何中位数，并消除与所有滤波器的几何中位数距离较近的冗余滤波器。然而，以往的方法基于“范数越小、重要性越低”的准则，即如果某个滤波器相对于原点或滤波器的几何中位数具有较低的 l_1/l_2-范数，就认为该滤波器的重要性较低，而忽略了高范数滤波器中存在的冗余。此外，这些方法也忽略了滤波器选择的多样性，因为只有高范数滤波器才会被视为重要。其他方法则利用基于相似性的度量来捕捉多样性。例如，Kim 等人 [20] 对滤波器进行聚类，从每个簇中选取一个滤波器作为重要滤波器，并消除其余滤波器。Singh 等人 [13] 则测量滤波器之间的相似性。

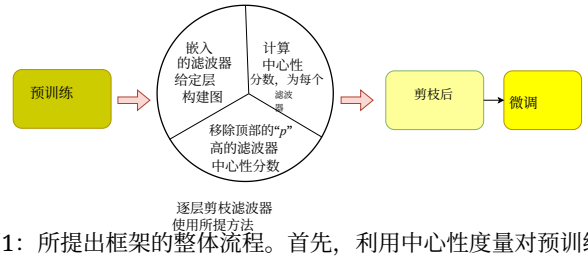

图1：所提出框架的整体流程。首先，利用中心性度量对预训练卷积神经网络进行逐层剪枝，随后通过微调过程恢复大部分性能。

通过计算所有滤波器之间的两两余弦距离，并从每一对相似的滤波器中剔除一个。

2.3. 最优剪枝启发式方法

在求解被动剪枝问题时，我们需要一个能够表明解的质量的良好启发式。我们采用的启发式是：移除的滤波器应与保留的滤波器最不相似，这样才能消除最多的冗余。为了获得具有 $|\mathcal{F}_{\text{pruned}}| = \lceil (1-p)|\mathcal{F}| \rceil$ 的最优滤波器集合 $\mathcal{F}_{\text{pruned}}$：

$$\mathcal{F}_{\text{pruned}} = \operatorname*{argmin}_{\mathcal{F}' \subseteq \mathcal{F}, |\mathcal{F}'| = \lceil (1-p)|\mathcal{F}| \rceil} \sum_{i=1} \sum_{j=1} sim(\mathcal{F}')_{i,j} \tag{2}$$

其中，sim 用于衡量两个滤波器的相似程度。这样，我们便在一个矩阵中对当前的总体相似度进行了建模。要实现一个能够精确计算此函数的算法，需要遍历 $\mathcal{F}$ 的所有可能子集，因此即使滤波器数量相对较少，也难以实现。

3. 所提方法

所提出的方法逐层剪枝卷积神经网络（CNN）。每一层包含一组 n 滤波器 $\mathcal{F}$，其中每个 $F \in \mathcal{F}$ 都是一个形状为（$w \times h \times c$）的三维张量。我们首先将每个滤波器 F展平为形状为（$wh \times c$）的 F^{flat}，其中 F展平为（$i-1$）$_{h+j、k} = F_{i、j、k}$。然后，我们对每个展平后的滤波器 F^{flat}执行奇异值分解（SVD），以找到其在 Frobenius 范数下的最佳秩-1近似；由此得到最显著的奇异值 σ_1 以及对应的左右奇异向量 l_1、r_1。现在，我们可以使用 $\hat{F} = (\sigma_1 l_1 r_1^T)_{i\,j、}$（$(\ )_{k、j})^2$来近似原始滤波器。接下来，我们从 $\sqrt{\sum_{k=1}^{wh} \sigma_1 l_1 r_1^T}$ $\hat{F}$中任选一列作为该滤波器的代表 f。我们可以任选一列，因为秩-1近似意味着所有列都相同。最后，我们计算该层中各滤波器代表 f_n 之间的余弦相似度，从而得到一个形状为 W $\in \mathbb{R}^{n \times n}$的相似度矩阵。利用 W，我们执行以下步骤，以获得给定层的重要性分数。完成这一预处理后，还需要执行另外三个步骤。

1. 将滤波器嵌入图中： 我们构建一个图 $G = (\mathcal{F}, E)$，其中 $\mathcal{F}$ 是表示滤波器的顶点集合，E 是满足属性 $w(e) = \mathbf{W}_{u,v}$ 的边集合，该属性为每条边 $e = (u, v) \in E$分配一个权重。令 $S_N(G)$ 表示 G 中所有大小为 N的完全子图的集合。因此：

$$S_N(G) = \{H \mid H \subseteq G, |\mathcal{F}_H| = N, \text{ and } E_H = \{(u,v) \in E \mid u \in \mathcal{F}_H \text{ and } v \in \mathcal{F}_H\}\} \tag{3}$$

因此，对于 $H \in S_{\lceil (1-p)\ |\mathcal{F}|\rceil}(G)$，$H = (\mathcal{F}_H,\ E_H)$ 是 G 的一个子图，其中 $|\mathcal{F}_H| = \lceil (1-p)|\mathcal{F}| \rceil$。因此，子图 H_{pruned} 可定义为：

$$H_{\text{pruned}} = (\mathcal{F}_{\text{pruned}}, E_{\text{pruned}}) = \text{argmin}_{H \in S_{\lceil (1-p)|\mathcal{F}| \rceil}(G)} \sum_{e \in E_H} w(e) \tag{4}$$

其中，$\mathcal{F}_{\text{pruned}}$ 的计算等同于公式（2）。

2. 计算每个滤波器的中心性分数：我们利用这一图结构设置和两种中心性算法来计算每个滤波器的重要性。

加权度中心性（WDC）[21]：我们的第一种方法采用加权度中心性（WDC）来评估每个节点的重要性。随后，我们保留与得分最低的 $\lceil (1-p)|\mathcal{F}| \rceil$ 个节点相对应的滤波器，因为得分最高的滤波器与大多数其他滤波器相似。我们使用以下方法计算这一重要性。

$$C(V) = \sum_{e \in \text{Edges of V}} w(e) \tag{5}$$

从而近似表示公式 (4)，并且】【。

$$\mathcal{F}_{pruned} = \{F_i \in \mathcal{F} \mid v_i \text{ is one of the } \lceil (1-p)|\mathcal{F}| \rceil \text{ largest elements in } \{C(F) \mid F \in \mathcal{F}\}\} \tag{6}$$

这个方程更容易求解，可以轻松地通过多项式时间近似求解，因此能够在性能损失最小的情况下完成计算。

介数中心性(中介中心性)[21]：我们还探讨了使用介数中心性（中介中心性）执行剪枝的想法。中介中心性根据节点对网络连通性的重要程度，为网络中的每个节点赋予一个分数。我们通过测量每个节点之间的最短路径，并统计每个节点出现在此类路径中的次数，来量化这一指标。因此，经过某个节点的最短路径越多，该节点的中心性就越高，移除它导致网络总体相似度下降的可能性也就越大。为进行这些实验，我们使用了与公式（6）相同的近似器，但使用以下指标对节点进行评分

$$C(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

其中，$\sigma(s,\ t)$ 是从节点 s 到节点 t的最短路径总数，而 $\sigma(s,\ t|v)$ 是经过节点 v的这些最短路径的数量。求和遍历图中所有节点对 s 和 t。

3. 获取剪枝网络并进行微调：获得给定层中每个滤波器的中心性分数后，我们剪除该层中中心性分数较高的 p 个滤波器。然后，对其他层重复相同的步骤，即步骤（1）和（2）。从各个卷积层移除滤波器后，即可得到剪枝网络。最后，对剪枝网络重新训练，以恢复因移除部分滤波器而损失的大部分性能。图1展示了所提方法的整体流程。

4. 实验设置

我们在为声学场景分类（ASC）和音频标签分类设计的卷积神经网络上评估所提出的剪枝框架。未剪枝卷积神经网络的概览如下：

(a) DCASE21_Net：我们采用一个公开可用、专为声学场景分类设计的 DCASE 2021 任务1A 基线网络，对 10 种不同的声学场景进行分类 [22]，并将其称为“DCASE21 Net”。DCASE21 Net 由三个卷积层（称为 C1 至 C3）和一个全连接层组成。网络的输入为尺寸为 (40 × 500) 的对数梅尔频谱图，对应一个 10 秒的音频片段；网络使用 Adam优化器训练 200 次迭代。该网络包含 46,246 个参数，推理时每个输入约需 287M 次乘加运算（MACs），准确率为 48.58%。

(b) PANNs_CNN14：PANNs [4] 是专为音频标签分类设计的大规模预训练音频神经网络。PANNs 使用包含 200 多万个带标签音频事件的数据集 AudioSet [23], 进行训练，其中涵盖 527 种不同的声音类别。在我们的实验中，我们使用一个 PANNs 模型 CNN14，该模型由 12 个卷积层组成（记为 C1 至 C12），并将该网络称为“PANNs CNN14”。

PANNs CNN 14 × 64 14 14 0 431 1000 以大小为 ( ) 的对数梅尔频谱图作为输入。CNN 在 AudioSet 评估数据集上取得了 mean average precision (mAPs) 和 0.973 的曲线下面积 (AUC)。CNN14 有 81M 个参数和 21G 次乘加运算次数[2]，对应于一个 10 秒长的音频片段，其采样率为 32KHz，窗口大小为 1024 个采样点，跳跃大小为 320 个采样点。PANNs CNN14 使用 Mixup 和 SpecAugment 等数据增强技术训练了 600k 次迭代。

剪枝和微调：使用所提出的基于中心性的剪枝方法获得各卷积层中的重要性滤波器后，我们从部分卷积层中消除排名前 $p \in \{25\%、\ 50\%、75\%\}$ 的不重要滤波器，以获得剪枝网络。获得剪枝网络后，我们在与训练未剪枝网络时相似的条件下对剪枝网络进行微调，例如损失函数、批量大小，但迭代次数更少。

对于 DCASE21_Net，我们考虑所有卷积层进行剪枝，并进行 100 次迭代的微调。对于 PANNs CNN14，我们提供初步分析，并仅剪枝 C7 至 C12 层，因为这些层包含约 99% 的参数；随后对剪枝网络进行 180k 次迭代的微调。**用于比较的其他方法：**我们将所提出的剪枝方法与以下方法进行比较：(a) l_1-norm [12], (b) 几何中位数（GM）方法 [19] 以及 (c) 成对相似度方法（CS） [13]。我们还采用主动滤波器剪枝方法，包括 HRank [8] 和能量感知剪枝 [9]，后者使用特征图进行剪枝。剪枝时，我们随机选择 500 个训练样本，为每个滤波器生成对应的特征图。随后，我们按照其他方法中使用的相同微调过程进行微调。

5. 性能分析

DCASE21_Net：图2展示了使用 BC 度量对 DCASE21 Net 的不同 p 下，对卷积层的不同子集进行剪枝后得到的准确率、参数数量和乘加运算次数。对 C3层中的 25% 滤波器进行剪枝，可使参数数量和乘加运算次数均减少 20%，而准确率仅下降约 1 个百分点。我们发现，与未剪枝网络相比，在各个层的不同子集中剪枝 25% 的滤波器，可使准确率下降小于 3 个百分点，同时参数减少约 60%，乘加运算次数减少约 40%。随着各层中被剪枝的滤波器数量

[2]MACs 计算 PyTorch 软件包。

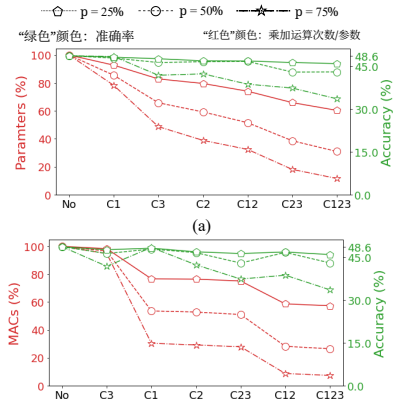

图2：在不同剪枝比例下，使用介数中心性（BC）度量对DCASE21 Net 的各个中间卷积层进行剪枝后得到的准确率。(a) 展示准确率与参数的关系，(b) 展示准确率与乘加运算次数的关系。

Tab表 1：DCASE21 Net 与其他剪枝方法的比较　　m方法。

剪枝方法	活跃	滤波器或特征图存储	准确率（%）	参数	MACs
基线（未剪枝）	-	-	48.58	46246	287M
HRank [8]	✓	1.26GB	47.24	24056	139M
能量感知 [9]	✓	1.26GB	47.00	—"—	—"—
l_1-范数 [12]	×	0.15MB	44.42	—"—	—"—
基于相似度 [13]	×	0.15MB	45.54	—"—	—"—
早上好，[19]	×	0.15MB	45.84	—"—	—"—
所提出的（中介中心性）	×	0.15MB	45.84	—"—	—"—
所提出的（WDC）	×	0.15MB	46.91	—"—	—"—

从 25% 增加到 75%，剪枝网络的准确率下降幅度、减少的参数数量和乘加运算次数均增加，但 C1层例外：在不同 p下，其对准确率下降和乘加运算次数表现出最低的敏感性。

接下来，我们在表1中比较了采用各种剪枝框架所获得的准确率。为了确保比较的公平性，我们按照成对相似度方法 [13]所采用的剪枝比例，获得剪枝后的网络。我们发现，采用中心性方法所获得的准确率等于或高于采用l_1-范数、GM和成对相似度方法所获得的准确率。与现有的主动剪枝方法相比，所提出的WDC剪枝方法在剪枝过程中无需占用大量内存，也仅需使用500个样本，即可获得近似相同的准确率。

PANNs CNN14： 图 3 展示了使用 BC 度量时，在不同剪枝比例下对剪枝网络进行微调所获得的平均精度均值。我们发现，在 C7 至 C12 层剪除 25% 的滤波器，可减少 41% 的参数和 24% 的乘加运算次数，同时性能略有提升，平均精度均值为 0.434，曲线下面积为 0.974，优于未剪枝网络。剪除 50% 的滤波器后，平均精度均值为 0.426，曲线下面积为 0.974，参数减少 70%，乘加运算次数减少 36%。剪除 75% 的滤波器后，平均精度均值为 0.399，曲线下面积为 0.973，参数减少 78%，乘加运算次数减少 46%。

接下来，我们在图 4 中比较所提剪枝方法与现有被动剪枝方法，比较时在 C7 至 C12 层剪除 25% 的滤波器。当微调迭代次数少于 15k 时，所提方法的平均精度均值优于其他方法。对剪枝网络进行 180k 次迭代微调后，所提方法略有提升

与其他方法相比的性能。总体而言，所提出的WDC剪枝方法得到的剪枝网络，其性能

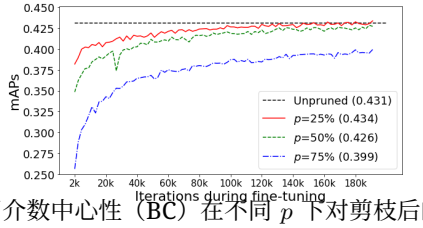

图 3：使用介数中心性（BC）在不同 p 下对剪枝后的 PANNs CNN14 进行微调期间获得的平均精度均值。同时，微调过程中获得的最大平均精度均值显示在圆括号中。

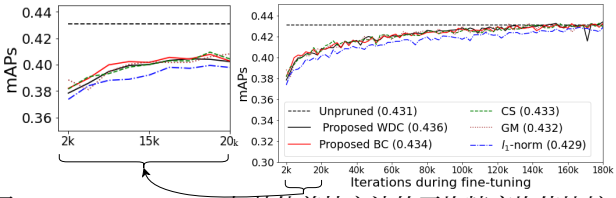

图 4：PANNs CNN14 与其他剪枝方法的平均精度均值比较。同时，每种方法获得的最大平均精度均值显示在圆括号中。

优于未剪枝的PANNs CNN14，同时还具有乘加运算次数和参数数量更少的优势。

6. 讨论

本文利用卷积神经网络中滤波器的图中心性来定义其冗余程度并进行剪枝。我们发现，所提出的基于加权度中心性的被动滤波器剪枝方法优于现有的成对相似度方法和基于范数的方法。对于 DCASE21 Net，我们的实验表明，所提出的被动剪枝方法无需涉及任何特征图，即可达到与主动滤波器剪枝方法相近的准确率。对于 PANNs CNN14，我们发现，剪枝网络的性能略优于未剪枝网络，同时训练迭代次数约减少为未剪枝网络所用次数的三分之一。这表明，可以先应用所提出的被动滤波器剪枝，获得规模更小的剪枝网络，然后仅需进行少量微调迭代（少于未剪枝网络所需的迭代次数），即可实现相近的性能，从而高效地利用现有的大规模预训练网络。因此，底层计算资源也能够得到有效利用。

7. 结论

本文提出了一种被动滤波器剪枝方法，通过探索滤波器的图中心性来降低卷积神经网络的计算复杂度和内存存储开销。与现有的基于范数的方法和成对相似性方法相比，所提出的剪枝方法取得了相近或更好的性能，体现了利用基于图的中心性度量来定义滤波器冗余的优势。与主动滤波器剪枝方法相比，所提出的被动剪枝方法无需在剪枝过程中涉及特征图，也能取得相近的性能。

未来，我们希望通过设计更优的中心性度量，并进一步降低微调过程的开销，来提升所提出的剪枝方法的性能。

8. 致谢

本研究部分得到了工程与物理科学研究理事会（EPSRC）博士培养合作项目 EP/T518050/1 提供的博士生奖学金，以及“声音人工智能（AI4S）”项目 EP/T019751/1 的资助。为实现开放获取，作者已将知识共享署名（CC BY）许可应用于由此产生的任何作者接受稿版本。

9. 参考文献

[1] Y.-H. Tu、J. Du 和 C.-H. Lee，“基于改进语音存在概率的师生深度学习，用于噪声鲁棒语音识别的语音增强”，*IEEE/ACM*音频、语音与语言处理汇刊，第27卷，第12期，第2080–2091页，2019年。[2] Y. Yu、S. Luo、S. Liu、H. Qiao、Y. Liu 和 L. Feng，“基于深度注意力的音乐流派分类”，神经计算，第372卷，第84–91页，2020年。[3] H. Purwins、B. Li、T. Virtanen、J. Schl¨uter、S.-Y. Chang 和 T. Sainath，“用于音频信号处理的深度学习”，*IEEE*信号处理精选主题期刊，第13卷，第2期，第206–219页，2019年。[4] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. 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，第28卷，第2880–2894页，2020年。[5] E. L. Denton、W. Zaremba、J. Bruna、Y. LeCun 和 R. Fergus，“利用卷积网络中的线性结构实现高效评估”，神经信息处理系统进展，第1269–1277页，2014年。[6] T. Heittola、A. Mesaros 和 T. Virtanen，“DCASE 2020挑战赛中的声学场景分类：跨设备泛化与低复杂度解决方案”，*2020*年声学场景与事件检测和分类研讨会论文集，2020年。[7] T. Liang、J. Glossner、L. Wang、S. Shi 和 X. Zhang，“深度神经网络加速的剪枝与量化：综述”，神经计算，第461卷，第370–403页，2021年。[8] M. Lin、R. Ji、Y. Wang、Y. Zhang、B. Zhang、Y. Tian 和 L. Shao，“HRank：利用高秩特征图进行滤波器剪枝”，*IEEE/CVF*计算机视觉与模式识别会议论文集，第1529–1538页，2020年。[9] S.-K. Yeom、K.-H. Shim 和 J.-H. Hwang，“通过能量感知剪枝构建紧凑型深度神经网络”，*arXiv*预印本 *arXiv:2103.10858*，2021年。[10] Y. He、G. Kang、X. Dong、Y. Fu 和 Y. Yang，“用于加速深度卷积神经网络的软滤波器剪枝”，*arXiv*预印本 *arXiv:1808.06866*，2018年。[11] S. Lin、R. Ji、C. Yan、B. Zhang、L. Cao、Q. Ye、F. Huang 和 D. Doermann，“通过生成对抗学习实现最优结构化CNN剪枝”，*IEEE*计算机视觉与模式识别会议论文集，第2790–2799页，2019年。[12]H. Li、A. Kadav、I. Durdanovic、H. Samet 和 H. P. Graf，“为高效卷积网络剪枝滤波器”，国际学习表征会议，2017年。[13] A. Singh 和 M. D. Plumbley，“一种基于被动相似性的CNN滤波器剪枝方法，用于高效声学场景分类”，*Interspeech*（*arXiv*预印本 *arXiv:2203.15751*），2022年。[14] Z. Liu、J. Li、Z. Shen、G. Huang、S. Yan 和 C. Zhang，“通过网络瘦身学习高效卷积网络”，载于*IEEE*国际计算机视觉会议论文集，2017年，第2736–2744页。[15]F. Lagunas、E. Charlaix、V. Sanh 和 A. M. Rush，“用于加速Transformer模型的块剪枝”，*arXiv*预印本 *arXiv:2109.04838*，2021年。[16]M. Xia、Z. Zhong 和 D. Chen，“结构化剪枝可学习紧凑且准确的模型”，载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），2022年，第1513–1528页。[17] H. Hu、R. Peng、Y.-W. Tai 和 C.-K. Tang，“网络裁剪：一种数据驱动的神经元剪枝方法，旨在构建高效深度架构”，*arXiv*预印本 *arXiv:1607.03250*，2016年。[18] Z. Wu、F. Li、Y. Zhu、K. Lu、M. Wu 和 C. Zhang，“一种基于特征图聚类的CNN模型滤波器剪枝方法”，《应用科学》，第12卷，第9期，第4541页，2022年。[19]Y. He、P. Liu、Z. Wang、Z. Hu 和 Y. Yang，“通过几何中位数进行滤波器剪枝，以加速深度卷积神经网络”，*IEEE*计算机视觉与模式识别会议论文集，第4340–4349页，2019年。[20] M. Park、W. Kim 和 S. Kim，“REPrune：通过代表选举进行滤波器剪枝”，*arXiv*预印本 *arXiv:2007 06932*，2020。。[21]A. Singh、R. R. Singh 和 S. R. S. Iyengar，“节点加权中心性：一种中心性混合的新方法”，《计算社会网络》，第7卷，第1期，2020年11月。[在线]。网址：https: // doi.org/10.1186/s40649-020-00081-w[22] I. Mart´ın-Morat´o、T. Heittola、A. Mesaros 和 T. Virtanen，“面向多设备音频的低复杂度声学场景分类：DCASE 2021挑战赛系统分析”，*DCASE*研讨会，2021年。[23] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，“Audio Set：用于音频事件的本体与人工标注数据集”，载于*2017*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集。IEEE，2017年，第776–780页。