doi:10.3969/j.issn.1672-1292.2023.04.004

新型轻量化神经网络结构范式的剪枝研究

宋　晨,魏子重,姜　凯,李　锐,段　强

(山东浪潮科学研究院有限公司,山东 济南 250014)

[**摘要**]　随着深度学习技术的推广,图像处理中的目标检测任务取得了蓬勃发展. 伴随着大模型的流行发展,深度学习模型精度在不断的上升. 而这些大模型却难以部署在日益发展的边缘设备上. 针对目前边缘端的目标检测任务,提出了一个 MobileOne-S0 和 SSD 相结合的网络结构,该网络结构经重参数化后,形成了 VGG 形式的网络结构用于推理过程. 随后采用了非结构化的权重剪枝,结构化的 BN 剪枝和泰勒剪枝这 3 种不同的剪枝标准进行了剪枝. 结果显示权重剪枝效果最差,而两种结构化剪枝对 FLOPs 和参数量随稀疏度上升下降速率几乎一致,但 BN 剪枝精度下降较泰勒剪枝缓慢,而泰勒剪枝对峰值内存大小的剪枝效果最好. 在模型精度下降约 10%时,BN 剪枝可以压缩 22.3 倍的参数量,9.4 倍的 FLOPs 和 2.5 倍的内存占用峰值大小. 最终模型大小仅为 123.88 kB,使模型更容易部署在 TinyML 适用、MCU 级别的低功耗端侧设备上.

[**关键词**]　MobileOne,SSD,深度可分离卷积,剪枝,TinyML
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1672-1292(2023)04-0029-08

Pruning Research on New Lightweight Neural Network Structures Paradigm

Song Chen,Wei Zizhong,Jiang Kai,Li Rui,Duan Qiang

(Inspur Academy of Science and Technology,Jinan 250014,China)

Abstract:With the widespread adoption of deep learning technology,the object detection task in image processing has made vigorous progress. Along with the popularity and development of large models,the accuracy of deep learning models continuously improves. However,these large models are difficult to deploy on edge devices that are increasingly developing. To address the current object detection tasks at the edge-side,a network structure combining MobileOne-S0 and SSD is proposed. This network structure is reparameterized to form a VGG-like network structure for the inference process. Then,three different pruning criteria are used,including unstructured weight pruning,structured BN pruning,and Taylor pruning. The results show that weight pruning has the worst effect,while the two structured pruning methods have almost the same decrease rate for FLOPs and parameter quantity with the increase of sparsity. However,the accuracy drop of BN pruning is slower than that of Taylor pruning while Taylor pruning has the best pruning effect on peak memory size. When the model precision decreases by about 10%,BN pruning can compress the parameter quantity by 22.3 times,FLOPs by 9.4 times,and peak memory usage by 2.5 times. The final model size is only 123.88 kB,making it easier to deploy on TinyML-suitable,MCU-level,low-power end-side devices.

Key words:MobileOne,SSD,deep separable convolution,pruning,TinyML

近年来,深度神经网络(deep neural network,DNN)在自然语言处理和图像领域取得了巨大的成功. 尤其是目标检测任务一直是图像领域的热点之一,得益于卷积神经网络的发展,目标检测任务的算法一直在进行不断的创新. 得益于更深和更宽的网络结构,深度学习模型的精度也在不断地上升. 然而这些大型模型都需要通过巨大的内存消耗和高计算量来实现强大的性能,难以部署在有限资源的硬件平台上. 而目前越来越多的数据出现在边缘任务端,这些图像或自然语言处理任务对延迟有非常高的要求(例如,自动驾驶),而且有些数据则对隐私非常敏感(例如,医疗保健),所以不能将这些任务和数据委托给云服务器.

收稿日期:2023-04-24.
通讯作者:李锐,博士,正高级工程师,研究方向:数据挖掘和机器学习. E-mail:lirui01@inspur.com

为了将目标网络布置在日益增多的边缘设备中,研究人员通过人工设计或者神经网络搜索(neural architecture search,NAS)搜索出专门为移动端设计的高效网络,还可以对现有网络通过剪枝和量化的方法进行高效的压缩. 这种方法需要我们将模型的大小、延迟、精度、内存的占用达到完美的平衡. 尤其在目前将人工智能和物联网技术融合的微型机器学习(tiny machine learning,TinyML)领域,要求布置在微控制器(microcontroller unit,MCU)设备上的模型的占用内存在几百 kB 以下,推理时间 ms 级别,功耗 mW 级别. 这种设备就需要研发更小,更快的深度学习模型.

针对 MCU 设备中目标人物检测的 MobileOne-S0 和 SSD 相结合的网络. 其中 SSD 结构采用了 Liu 等[1]提出的网络结构,而 MobileOne-S0 则采用了 Pavan 等[2]提出的网络结构. 本文主要研究以下内容:

(1)设计了针对 MCU 设备的 MobileOne-S0 结合 SSD 的网络结构. 该结构是一种训练时采用多分支结构而推理时将多分支结构合并形成 VGG 式的链式网络结构. 该网络结构经过重参数化的过程,使得训练时和推理时输出相同的结果.

(2)设计了专门用于深度可分离卷积的结构化剪枝方法. 该方法能有效地缩小模型的大小,利用深度卷积层中的对位原则,解决了剪枝后无法直接布置模型问题,从而使模型在剪枝完后即可以迅速的部署到 MCU 设备中.

(3)对上述网络采用了 3 种不同的剪枝标准,分别是将每个通道中权重的绝对值的大小,每个通道批归一化层(batch normalization,BN)的权重参数 γ 的绝对值的大小和每层通道泰勒一阶项的大小作为剪枝的标准. 其中第一种按权重的绝对值大小为非结构化剪枝,后两种为结构化剪枝. 实验并分析了 3 种剪枝在 MobileOne-S0-SSD 网络结构剪枝的效果. 其中 BN 剪枝效果最好. 在模型精度下降约 10%时,BN 剪枝可以压缩 22.3 倍的参数量,9.4 倍的 FLOPs 和 2.5 倍的内存占用峰值大小,最终模型大小仅为 123.88kB.

1 相关工作

1.1 目标检测

目标检测是深度学习计算机视觉领域中的一个应用方向,其目的是从自然图像中的定位大量的目标实例. 目前的算法主要包含 TwoStage 目标检测算法和 OneStage 目标检测算法. 前者包括 R-CNN[3]、SPP-Net[4]、Fast R-CNN[5]、Faster R-CNN[6]等,后者包括 YOLO 系列[7-8]、SSD[1]等. 由于 TwoStage 目标检测算法需要先在图像中提出候选区域,再对候选区域进行分类和定位. 而 OneStage 目标检测算法则直接将上述两个过程合并在了一起. 所以在实时检测任务中,OneStage 目标检测算法一直是任务的首选项. 而其中的 SSD 算法的结构尤为简单. 当使用 VGG16 作为主干网络后,整个网络结构为链式结构,没有任何分支结构,所以网络运行速度快. 而且 SSD 的结构具有可替换性,主干网络可以替换成更为先进快速的轻量化网络结构,以便于我们重新设计.

1.2 高效的神经网络设计

前人通过人工设计和 NAS 搜索针对移动设备的高效网络已经取得了很大的进展[9-16]. 但有些高效的网络在提升精度的同时过多地占用了内存(如 Inception 的分支结构[13])或提升了推理时的延迟(如 ShuffleNet 中需要调整通道[14]),降低了整体的效率. 现要将网络部署在内存和延迟都限制很小的边缘设备上,只能考虑具有简单链式结构的卷积神经网络. MobileOne-S0 的网络结构是针对移动网络改进的一种网络架构[17]. MobileOne-S0 推理时的网络结构由深度可分离卷积构成,训练时由深度卷积层和逐点卷积层这两种并列结构组成. 在推理时则通过重参数化算法将上述的并列结构参数合并为深度可分离卷积,这时整个网络结构就变成了全为深度可分离卷积构成的链式结构. 所以这个结构在兼顾了速度的条件下还最大程度上提取了图像的特征.

1.3 剪枝进展

剪枝是减少模型占用内存大小和带宽的一项重要的技术. 对剪枝后的网络进行重新训练,有可能会取得更高的精度[18]. 剪枝通过不同粒度可分为结构性剪枝和非结构性剪枝. 结构性剪枝相比于非结构性剪枝缺点是含有较少的稀疏度,优点是剪完后模型可以直接压缩,通过硬件进行加速. 而非结构化剪枝会导致参数结构变成稀疏矩阵,但很多硬件并不支持稀疏矩阵的运算,导致修剪后的网络实际很难得到加速. 剪枝就是剪去模型中过多的参数或结构,也就是网络冗余,剪枝不会显著损失精度. 如何定义网络冗

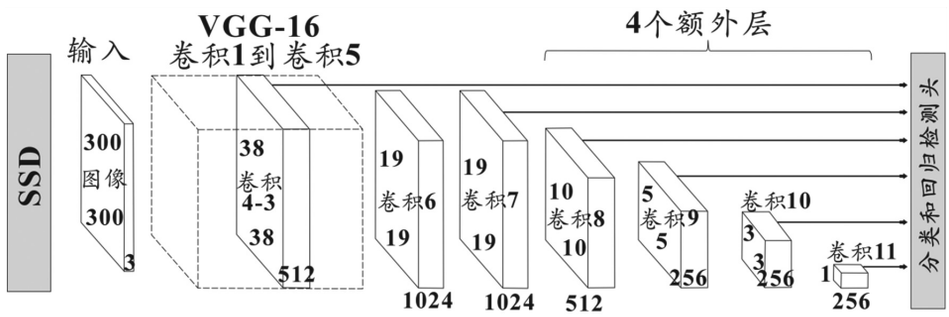
余通常是剪枝最为重要的部分,可以决定剪枝最终模型的大小.

早在 1989 年,Lecun 等[19]提出了最优脑损伤剪枝(optimal brain damage,OBD)来剪去不必要的权重,但由于这种方法需要计算损失函数的二阶导数海森矩阵(hessian matrix),而由于权重数量很多,计算海森矩阵需要很大的算力,导致这种剪枝并未得到推广. Han 等[20]提出基于幅度的剪枝,将低于某个阈值的权重全部剪掉,之后再进行微调直到精度达到预期. 在微调过程中加入 L2 的正则化过程使后面的权重更倾向于两极分化. 通过这种迭代式的剪枝方法最终获得高稀疏度的网络. Liu 等[21]通过提取在 BN 层中的缩放因子 γ 来确定通道的重要性. 将低于某个阈值的通道全部剪掉,之后也是微调并在微调过程中增加一个关于 γ 的正则项. Molchanov 等[22]介绍了一种通过求解损失函数的泰勒公式的一阶导数来判断每个通道的重要性,每次剪枝需要运行一次训练集以便于提取每层的激活值和梯度. 由于这种方法的计算在每层卷积中会有不同的标准,所以作者在所有层计算后会将每一层内应用 L2 归一化. 这 3 种不同的定义网络冗余的标准中,第一种为非结构化剪枝,后两种皆为结构化剪枝. 选取了这 3 种比较具有代表性的剪枝标准,并在我们的高效的网络上应用了上述剪枝标准.

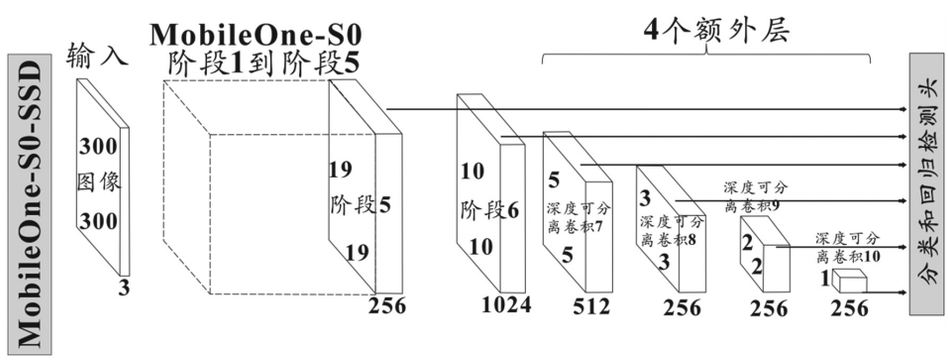
2　目标检测网络构建和剪枝标准

2.1　目标检测网络的构建

SSD 的网络主要由 VGG16 作为主干网络,再加上 4 个额外的卷积层(如图 1(a)所示). 其中分类和回归预测总共提取 SSD 网络结构中的 6 层. 浅层特征图用于提取小目标,而深层的特征图则用于提取较大的目标. 首先我们将整个主干网络 VGG16 替换为 MobileOne-S0 结构. MobileOne-S0 中截取到前面的 stage6 为止. 另外额外的 4 个卷积层也替换为 4 个深度可分离卷积(如图 2 所示). 模型应用于边缘端的目标检测任务,摄像头检测的对象均为大目标. 针对网络的特殊性,分类和回归预测提取的 6 层中,取消了 38×38 的特征输出图,改由 19×19 作为特征输出层的第一层. 再将 MobileOne-S0 中 stage6 的通道由 512 改为 1024. 依次获取 MobileOne-S0 的 stage5,stage6 和 4 个额外层的输出特征图(如图 1(b)). SSD 和 MobileOne-S0 的结合可以将速度保持到最快,又兼顾了保留图像最大的特征提取.



(a) 原始的SSD模型结构



(b) 修改的MobileOne-S0-SSD模型结构

图 1　改进的 SSD 模型结构

Fig. 1　Improved SSD model structure

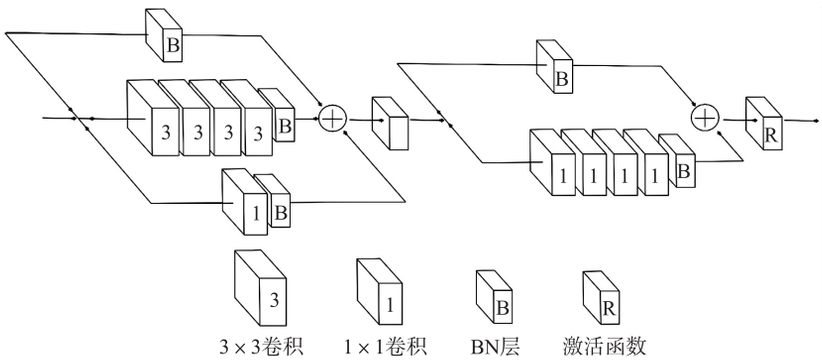


图 2　MobileOne-S0 的深度可分离卷积的多分支

Fig. 2　MobileOne-S0 multibranch structure with deeply separable convolution

2.2　3 种不同的剪枝标准

采用 3 种不同的剪枝标准. 第一种是非结构化的权重剪枝,剪枝思想参考文献[20]. 操作过程为提取已经训练或微调好的模型,将权重参数全部展平绝对值化后提取到一个向量中. 直接设置或计算出剪枝的阈值,将小于阈值的掩码部分置 0,其余为 1,之后每次微调时只需参数乘以相对应的掩码即可. 另外每层人为设置最小剪枝数目. 在微调过程中加入 L2 正则化,使得权重更易剪枝. 第二种剪枝为结构化的 BN 剪枝,剪枝思想参考文献[21]. 操作过程需要提取 BN 层中的缩放因子 γ,即 BN 层的权重参数. 后续过程和第一种相似,将权重变成缩放因子 γ,掩码变成针对通道维度的掩码. 另外需要在目标函数上加上缩放因子 γ 的惩罚项 L1 范数,使得 γ 实现稀疏性. 第三种剪枝标准需要计算泰勒公式的一阶项,剪枝思想参考文献[22]. 计算泰勒标准的公式为

$$\Theta_{\mathrm{TE}}(h_i)=\left|\frac{\delta C}{\delta h_i}h_i\right|.$$

其中,h_i 为激活值,$\dfrac{\delta C}{\delta h_i}$为激活值的梯度,在计算泰勒标准时需要反向传播一次. 计算出公式的结果后再通过通道维度进行平均,计算出每个通道的标准,标准小的进行剪枝. 后续的剪枝过程和第一种相似.

2.3　剪枝后新模型的构建

剪枝后部署新模型时需要直接构建保留通道的新模型. 由于模型中采用了深度可分离卷积,其中的深度卷积层的输入和输出通道必须一致,而逐点卷积层则可以不一致. 所以深度卷积层需要实行对位原则,即剪枝只可以剪去输入通道和输出通道都剪去的相同位置的通道. 例如图 3(a)中输入通道可以剪去 1,3,4,6,7 通道,而输出通道可以剪去 2,3,5,6,7 通道. 在深度卷积层中只可以剪去两者共同的 3,6,7 通道. 1,2,4,5 通道由于对位通道不需要剪枝,所以需要保留这些通道(如图 3(b)所示).

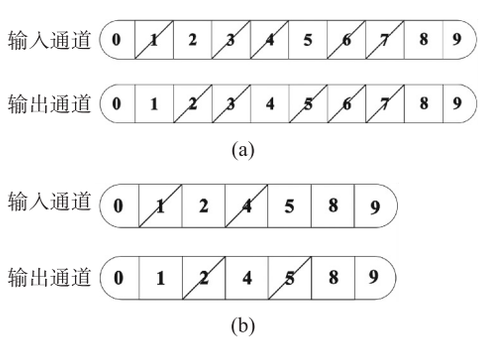


图 3　深度可分离卷积层中深度卷积层的剪枝示意图

Fig. 3　Schematic diagram of the pruning of the deep convolutional layer in deep separable convolution

3　剪枝实验结果和分析

3.1　实验参数

使用 MobileOne-S0 在 ImageNet-1K 上预先训练好的权重作为初始化的权重. 后续再针对训练集 VOC2007 和 VOC2012 中“人”部分进行训练. 具体使用的参数:数据读取的批次大小为 16,权重衰减系数为 0.000 1,随机梯度下降的动量设置为 0.9. 初始学习率设置为 0.01,使用余弦下降来调整学习率. 训练迭代轮次为 900 轮,于是余弦下降 T_max 设置为 900. 训练完成后在 VOC2007 测试集上的精度为 0.715 2.

实验使用 Pytorch 框架,在 Ubuntu18.04 操作系统下训练,显卡采用的是 8 张 Nvidia Tesla V100.

3.2　3 种不同的剪枝结果

剪枝的过程会将精度 AP50 从 0.715 2 降到 0.6 左右,然后看 3 种剪枝标准的结果. 由于剪枝完成后的模型的大小和在硬件中的内存占用峰值较为重要,故将 FLOPs、参数量、模型在硬件中的内存占用峰值大小和模型在硬件中最终占用的大小 4 个指标计算出评估 3 种剪枝标准. 最后一个指标为第二个指标和第三个指标之和.

首先,使用第一种非结构化的权重剪枝. 剪枝阈值设置为 1×10^{-2},剪枝后进行微调,微调参数和训练的参数除了 epochs 和 T_max 参数设置为 500 外都一样. 每 5 个 epochs 保留一次模型,最终取 loss 最低的 40 个模型进行测试集测试,选出精度最高的一个模型作为下次剪枝的初始模型. 权重剪枝到稀疏度 96.6%时根据稀疏度设置阈值,每次提高 0.5%稀疏度,直到稀疏度为 98.6%,剪枝完成. 最终的精度为 0.622 3. 详细的结果参见表 1.

表 1　MobileOne-S0-SSD 权重剪枝结果

Table 1　MobileOne-S0-SSD weight pruning results

稀疏度/%	FLOPs/Mmac	参数量/kB	内存占用峰值大小/kB	最终硬件占用内存大小/kB	精度
0	610.08	2 764.80	1 054.69	3 819.49	0.715 2
62.3	569.32	2 672.64	747.07	2 673.37	0.710 0
93	424.33	2 068.48	527.34	2 068.99	0.697 2
98.6	358.60	1 976.32	483.40	1 976.79	0.622 3
99.1	357.92	1 976.32	483.40	1 976.79	0.549 8

从表 1 可知,权重剪枝剪到稀疏度为 93%时还保持有 0.697 2 的精度,但剪枝过程中参数量和 FLOPs 下降的很慢(和后两种剪枝标准相比). 当稀疏度为 93%时,FLOPs 减少了 30.4%,参数量减少了 25.2%,内存占用峰值大小减少了 50.0%,所以权重剪枝对于模型的压缩很不友好. 虽然最终稀疏度可以到 98.6%,但模型的 FLOPs 仅减少了 41.2%,参数量仅减少了 28.5%,内存占用峰值大小减少了 54.2%.

第二种剪枝为结构化的 BN 剪枝,微调和第一种参数设置相同. 剪枝的过程为先设置 BN 层的缩放因子 γ 的阈值为 0.001,一直剪枝到稀疏度为 74%,随后每次剪 1%的稀疏度直到 84%. 再按每次剪 0.8%的稀疏度到 87.2%结束剪枝,最终的精度为 0.598 5. 详细的结果参见表 2.

表 2　MobileOne-S0-SSD BN 剪枝结果

Table 2　MobileOne-S0-SSD BN pruning results

稀疏度/%	FLOPs/Mmac	参数量/kB	内存占用峰值大小/kB	最终硬件占用内存大小/kB	精度
0	610.08	2 764.80	1 054.69	3 819.49	0.715 2
45	287.47	980.22	615.23	1 595.45	0.702 8
62	197.31	582.15	593.26	1 175.41	0.698 5
68.9	153.27	443.42	527.34	970.76	0.685 2
87.2	65.10	123.88	461.43	699.27	0.598 5

从表 2 可知,BN 剪枝剪到稀疏度为 62%时精度为 0.698 5,而 FLOPs 减少了 67.7%,参数量减少了 78.9%,内存占用峰值减少了 43.8%. 相比于第一种剪枝,BN 剪枝的模型压缩效率远远提升. 到最终剪枝完毕时参数仅为初始模型的 4.48%,压缩了近 22.3 倍. 虽然精度下降了 0.117,但模型的缩小使它能够运行在严重限制模型大小的边缘设备上.

第三种剪枝为结构化的泰勒剪枝. 泰勒剪枝第一次剪去 10%的稀疏度,然后每次剪 5%的稀疏度,最终剪枝稀疏度为 70%. 详细结果见表 3. 泰勒剪枝在稀疏度为 35%时,精度能保持在 0.696 1,剪枝 FLOPs 减少了 43.5%,参数量减少了 50.0%,内存占用峰值大小减少了 64.6%. 结果优于权重剪枝,但弱于 BN 剪枝. 最终剪枝完毕精度为 0.591 9,模型 FLOPs 减少了 81.4%,参数量减少了 85.2%,内存占用峰值减少了 81.3%.

3 种剪枝标准的最终模型的压缩倍数统计列如表 4 所示. 从表 4 可见,权重剪枝的稀疏度最高,但该方法实际压缩的倍数最小. BN 剪枝是模型压缩最大的剪枝标准,参数量可以压缩 22.3 倍,FLOPs 压缩 9.4 倍. 而内存占用峰值大小压缩最大的是泰勒剪枝,可以压缩 5.3 倍. 最终在硬件中的占用内存大小,BN 剪

枝最高可压缩 7.1 倍.

表 3　MobileOne-S0-SSD 泰勒剪枝结果

Table 3　MobileOne-S0-SSD Taylor pruning results

稀疏度/%	FLOPs/Mmac	参数量/kB	内存占用峰值大小/kB	最终硬件占用内存大小/kB	精度
0	610.08	2 764.80	1 054.69	3 819.49	0.715 2
10	550.93	2 590.72	769.04	3 359.76	0.710 7
35	344.89	1 382.40	373.54	1 755.94	0.696 1
50	242.33	868.63	285.64	1 154.27	0.663 0
70	113.23	409.09	197.75	606.84	0.591 9

表 4　3 种剪枝标准的模型压缩结果

Table 4　Model compression results of three standard pruning

剪枝后 剪枝模型	稀疏度/%	缩减比例 FLOPs/Mmac	参数量/kB	内存占用峰值 大小/kB	最终硬件占用内存 大小/kB	精度
权重剪枝	98.6	1.7	1.4	2.2	1.9	0.622 3
BN 剪枝	87.2	9.4	22.3	2.5	7.1	0.598 5
泰勒剪枝	70	5.4	6.8	5.3	6.3	0.591 9

图 4 中展示的是 3 种剪枝过程中 4 种指标随稀疏度的变化. 从图 4(a)和图 4(b)可见,结构化剪枝的 BN 剪枝和泰勒剪枝的下降的斜率几乎一致. 说明结构化剪枝剪的深度可分离卷积中包含 3×3 卷积层和 1×1 卷积层,但相同剪枝稀疏度下,两种剪枝标准却有相似的两种卷积层比例. 而内存占用峰值大小却完全不相同(如图 4(c)所示),因为这个指标由网络结构中最大的激活值的大小确定. 在这个网络结构中,输入图片被设定为 300×300 大小,计算得出最大的激活值是第一层网络的输出,未剪枝前是 48×150×150. 泰勒剪枝对这层的剪枝效果非常好,所以峰值内存大小下降很快. 从精度随剪枝稀疏度下降的图 4(d)中看出,存在一个阈值,只要在这个阈值之下剪枝,模型的精度损失会非常小. 而稀疏度超过这个阈值后,微调后的模型精度才会比较快速的下降[21]. 在本文中这个阈值约为 0.697.

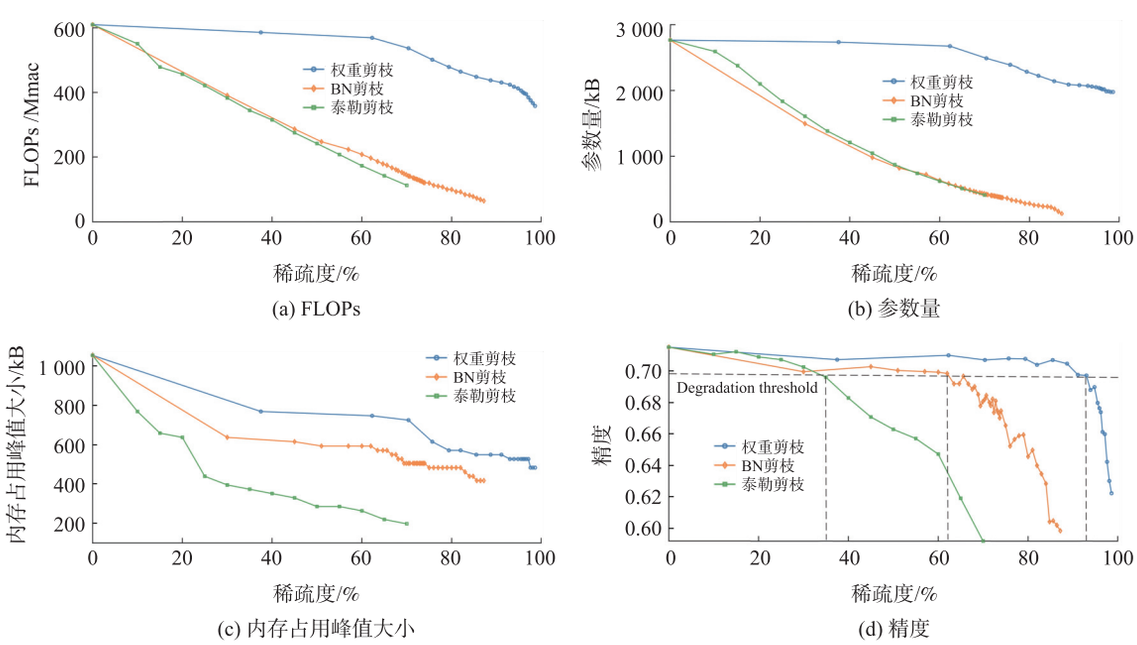


图 4　3 种剪枝标准随着稀疏度上升 3 种指标的变化

Fig. 4　Variation of three pruning standards with increasing sparsity for three metrics

表 5 展示了模型在 3 种不同的剪枝标准下精度 0.7 和 0.6 附近实际压缩的比例. 模型精度在 0.7 附近可以认为模型几乎没有什么损失. 在这种情况下,3 种剪枝标准可以压缩最多 4.7 倍的参数量,3.1 倍的 FLOPs 和 2.8 倍的内存占用峰值大小,说明模型内部确实存在很多网络冗余. 但去掉这些网络冗余后,再进行剪枝,模型精度就会较快速地下降. 从精度 0.7 附近到 0.6 附近,FLOPs 最多下降 3 倍,参数量最多下

降 4.7 倍,内存占用峰值大小最多下降 1.9 倍.

表 5　3 种不同的剪枝标准在精度变化下 FLOPs、参数量和内存占用峰值大小压缩的倍数

Table 5　Three different pruning criteria result in a compression factor for FLOPs, parameter count, and peak memory size when pruned

	精度	FLOPs 压缩倍数	参数量压缩倍数	内存占用峰值大小压缩倍数
原始模型	0.715 2			
权重剪枝	0.697 2	1.4	1.3	2
	0.622 3	1.2	1.1	1.1
Bn 剪枝	0.698 5	3.1	4.7	1.8
	0.598 5	3	4.7	1.4
泰勒剪枝	0.696 1	1.8	2	2.8
	0.591 9	3	3.4	1.9

3.3　剪枝结果分析

经分析可知,BN 剪枝是最优的剪枝标准. 3 种剪枝标准中权重剪枝是评估卷积层中单个权重的大小,BN 剪枝是评估整个特征图的值的大小,泰勒剪枝是评估特征图的值和它的梯度相乘后再平均的大小. BN 剪枝是最优的剪枝结果说明特征图的值越小对精度的影响就越小. 而且 BN 层参数中的缩放因子 γ 本身是对整个特征图的乘积,所以缩放因子 γ 影响的是整个特征图的大小. 而权重剪枝中通道可能残留重要的权重(如果按通道中参数的绝对值总和剪通道),泰勒剪枝计算的最后一步需要平均后得到通道的判别标准值,所以也存在重要的一些特征图被剪枝. 另外在 BN 剪枝过程中加入的 L1 范数的惩罚项实施的稀疏化对缩放因子 γ 的效果也非常好,使得 γ 很容易向零靠近.

4　结论

本文提出了一种用于 MCU 设备中目标人物检测的 MobileOne-S0 和 SSD 相结合的网络结构. 网络针对人的识别率在 VOC 测试集上可达到 0.715 2 的精度. 针对全由深度可分离卷积层构成的网络结构,发明了对位通道剪枝的方法. 非结构化剪枝(权重剪枝)对实际模型的缩小效果最差,两种结构化剪枝对 FLOPs 和参数量随稀疏度升高而下降的斜率几乎一致,但 BN 剪枝精度下降得比泰勒剪枝要缓慢. 所以在几乎无损的条件下 BN 剪枝表现最好,可以压缩 4.7 倍的参数量,3.1 倍的 FLOPs 和 1.8 倍的内存占用峰值大小. 在模型精度下降约 0.1 精度时,BN 剪枝可以压缩 22.3 倍的参数量,9.4 倍的 FLOPs 和 2.5 倍的内存占用峰值大小. 最终模型大小仅为 123.88 kB,使得实时进行图像目标检测的模型对内存容量和延迟的要求更小,从而更容易在 MCU 设备上部署.

[参考文献](References)

[1] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: single shot multibox detector[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[2] PAVAN K A V, GABRIEL J, ZHU J, et al. An improved one millisecond mobile backbone[J/OL]. arXiv Preprint arXiv: 2206.04040, 2022.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, USA: IEEE, 2014: 580-587.

[4] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[5] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1440-1448.

[6] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016: 779-788.

[8] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time

object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada:IEEE,2023:7464-7475.

[9] IANDOLA F N,HAN S,MOSKEWICZ M W,et al. Squeezenet:Alexnet-level accuracy with 50x fewer parameters and<0.5 MB model size[J/OL]. arXiv Preprint arXiv:1602.07360,2016.

[10] HOWARD A G,ZHU M,CHEN B,et al. Mobilenets:efficient convolutional neural networks for mobile vision applications[J/OL]. arXiv Preprint arXiv:1704.04861,2017.

[11] SANDLER M,HOWARD A,ZHU M,et al. MobileNetV2:inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,Utah,USA:IEEE,2018:4510-4520.

[12] HOWARD A,SANDLER M,CHU G,et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach,California,USA:IEEE,2019:1314-1324.

[13] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,Nevada,USA:IEEE,2016:2818-2826.

[14] ZHANG X,ZHOU X,LIN M,et al. Shufflenet:An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,Utah,USA:IEEE,2018:6848-6856.

[15] CAI H,ZHU L,HAN S. Proxylessnas:Direct neural architecture search on target task and hardware[J/OL]. arXiv Preprint arXiv:1812.00332,2018.

[16] CAI H,GAN C,WANG T,et al. Once-for-all:train one network and specialize it for efficient deployment[J/OL]. arXiv Preprint arXiv:1908.09791,2019.

[17] DING X,ZHANG X,MA N,et al. Repvgg:making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual:IEEE,2021:13733-13742.

[18] FRANKLE J,CARBIN M. The lottery ticket hypothesis:finding sparse,trainable neural networks[J/OL]. arXiv Preprint arXiv:1803.03635,2018.

[19] LECUN Y,DENKER J,SOLLA S. Optimal brain damage[J]. Advances in Neural Information Processing Systems,1990,2(279):598-605.

[20] HAN S,POOL J,TRAN J,et al. Learning both weights and connections for efficient neural network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal,Canada:NIPS,2015:1135-1143.

[21] LIU Z,LI J,SHEN Z,et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice,Italy:IEEE,2017:2736-2744.

[22] MOLCHANOV P,TYREE S,KARRAS T,et al. Pruning convolutional neural networks for resource efficient inference[J/OL]. arXiv Preprint arXiv:1611.06440,2017.

[责任编辑:陈 庆]