

面向轻量化神经网络的模型压缩与结构搜索

梁峰¹, 董名², 田志超¹, 张国和¹, 成舒婷¹

(1. 西安交通大学微电子学院, 710049, 西安; 2. 西安交通大学软件学院, 710049, 西安)

摘要: 针对轻量化神经网络中大量 1×1 卷积操作限制网络模型压缩的问题,提出了轻量化神经网络模型压缩算法,并对网络结构进行了搜索。根据有限长单位冲激响应滤波器线性相位特性思想,设计了具有线性相位约束的 1×1 卷积滤波器,将其应用在 MobileNet 网络中验证了有效性。采用遗传算法对 MobileNet 网络中具有线性相位约束的 1×1 卷积滤波器比例进行搜索,使用权重共享算法对遗传算法搜索过程进行加速,使用相对适应度指导算法进化方向。实验证明:具有线性相位约束滤波器的 MobileNet 网络在 Cifar10 数据集上的网络参数量降低为原始 MobileNet 网络的 51.24%,网络准确率下降 0.38%;在 ImageNet 数据集上的网络参数量降低为原始 MobileNet 网络的 62.88%,网络 top5 准确率下降 1.44%,具有线性相位约束的 1×1 卷积滤波器可以对网络模型进行有效压缩;遗传算法搜索出的最优结构网络准确率与 MobileNet 网络的相仿,网络参数量下降为原始网络的 83.54%,网络模型更小,性能更优。

关键词: 轻量化神经网络;模型压缩;遗传算法;权重共享

中图分类号: TP311 **文献标志码:** A

DOI: 10.7652/xjtuxb202011013 **文章编号:** 0253-987X(2020)11-0106-07



OSID 码

Model Compression and Structure Search for Lightweight Neural Network

LIANG Feng¹, DONG Ming², TIAN Zhichao¹, ZHANG Guohe¹, CHENG Shuting¹

(1. School of Microelectronics, Xi'an Jiaotong University, Xi'an 710049, China;

2. School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: The 1×1 convolutional filters are widely used in lightweight neural networks and accounted for a large part of the overall network. Lots of 1×1 convolutional filters have limited the compression ratio of the network model. A lightweight neural network model compression method is proposed, and the network structure is searched. Inspired by the linear phase characteristic of the FIR filter, the proposed 1×1 convolution filters also have a linear phase. We perform experiments on MobileNet to verify the effectiveness of our proposed method. The genetic algorithm is used to search the optimal ratio of 1×1 convolution filters with a linear phase constraint of MobileNet. Then, we adopt the method of weight sharing to accelerate the search process of genetic algorithms. In addition, the relative fitness is utilized to guide the evolution of the algorithm. The experiments result show that, in the Cifar10 dataset, after adding linear phase constraint to MobileNet, the amount of parameter is reduced to 51.24%, while the accuracy rate only dropped by 0.38%. In the ImageNet dataset, the network parameter amount is reduced to 62.88%, the network top5 accuracy rate is only dropped by 1.44%. The 1×1 convolution filters with the linear phase constraints can further compress the network. The

收稿日期: 2020-05-18。 作者简介: 梁峰(1976—),男,副教授;张国和(通信作者),男,教授。 基金项目: 陕西省自然科学基金资助项目(2020JM-006);航空科学基金资助项目(20184370012)。

网络出版时间: 2020-07-03

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1069.T.20200702.1145.004.html>

network with optimal structure searched by genetic algorithm achieves a higher accuracy rate and the amount of network parameters is reduced to 83.54%. The proposed network requires smaller storage space and has better performance.

Keywords: lightweight neural network; model compression; genetic algorithm; weight sharing

轻量化神经网络模型是专门针对嵌入式视觉应用终端设计的轻量且高效的神经网络模型。相较于传统神经网络,该网络结构更加简单,学习和推理需要的计算资源需求量较少,可以有效提升所有以卷积神经网络为基础的计算机视觉模型的性能和效率,经济价值潜力巨大。网络模型性能优劣将直接影响目标检测、人脸识别、语义分割等目前主流计算机视觉系统的性能。

2017年5月,谷歌推出自动机器学习,越来越多的自动优化神经网络结构算法被提出,旨在设计网络模型更小、性能更优的神经网络。网络结构搜索通过算法自动搜索神经网络结构,分为3个步骤:①定义目标搜索空间;②根据算法中选择的搜索策略产生新网络结构;③评估搜索出网络结构的优劣并反馈。目前,主流的搜索神经网络结构算法包括基于梯度的算法、强化学习算法^[1]、进化算法^[2]等。

设计轻量化神经网络时,大量采用的 1×1 卷积操作参数量占整个网络的主要部分^[3],限制了网络模型压缩。本文根据有限冲激响应(FIR)滤波器线性相位约束思想,设计具有线性相位约束的 1×1 卷积滤波器,将 1×1 滤波器在深度方向上进行线性相位约束处理。具有线性相位约束的 1×1 卷积滤波器参数量减少一半,有利于网络模型压缩,在设计轻量化神经网络时具有较大优势。

网络结构搜索算法可以根据需求自动搜索网络结构,优化网络性能。本文使用遗传算法,对 MobileNet 网络中每层包含的具有线性相位约束的 1×1 对称滤波器比例进行搜索,可以减少人工调参时间,获得性能更佳、模型更小的网络。

1 神经网络模型压缩与结构搜索

1.1 网络模型压缩

2016年,从轻量化神经网络 SqueezeNet 提出开始,越来越多无需高昂 GPU 资源的优秀轻量化神经网络模型被设计出来,大幅度提高了神经网络的经济价值。Iandola 等提出先利用 1×1 卷积操作对输入特征压缩,再使用 1×1 和 3×3 卷积对特征进行扩张的 fire module,使用 fire module 构建的 SqueezeNet 在 ImageNet 数据集上实现了和 Alex-

Net 相同的准确率,参数量只有 AlexNet 的 $1/50$ ^[4]。MobileNet 网络不仅采用了深度可分离卷积结构,将原始 3×3 卷积进行了拆分,大幅度降低了网络参数量,还采用了宽度乘子和分辨率乘子,对网络模型进行进一步压缩^[5]。MobileNet-v2 采用先升维再降维的方式,增强网络中间深度卷积层的表现能力,并移除了降维层激活函数。使用反向残差单元的 MobileNet-v2 在多种视觉任务数据集上获得了更高的准确率,并且需要的计算资源大幅减少。Zhang 等提出了 ShuffleNet,为了解决网络结构中大量密集 1×1 卷积导致的网络性能低下问题,提出了逐点分组卷积思想,使用两个 1×1 group 卷积将通道间学习分解为两步^[6]。同时,采用通道混合算法促进各通道间信息融合,使 group 卷积能够获得不同 group 的输入数据,将输入特征和输出特征联系起来,提高网络特征提取能力,搭建更强有力的网络结构^[6]。谷歌推出 NasNet 网络,其结构主要由 normal layer 和 reduction layer 两种不同 cell 堆叠而成,该网络可以针对不同数据集,采用人工堆叠方式灵活调整 normal layer 和 reduction layer 两种 cell 的堆叠方式,使网络应用范围更加广泛^[7]。

以上这些经典的轻量化神经网络模型压缩算法被广泛应用于计算机视觉中,摆脱了高昂的计算资源限制,实现了神经网络从实验室走向工业,提高了神经网络应用范围和经济价值。

1.2 网络结构搜索

神经网络在处理非线性问题(如图像识别)时表现良好,但是在设计时存在网络结构难以确定、网络训练参数难以选取、网络权值难以设定等问题。这些问题导致设计神经网络时需要大量人工干预,只有小部分专业科研人员可以参与网络设计,限制了缺乏经验的科研人员参与网络设计,不利于神经网络的发展以及推广。神经网络结构搜索中,搜索空间的参数为定义网络结构的参数,具体为网络层数、卷积核大小、每层选取的算子等参数。网络结构搜索为近年来深度学习领域的研究热点,通过高效经济的搜索算法,自动获得泛化能力强、硬件友好的神经网络结构,节省了人工设计网络时间,并且在多种场景下击败了早期人工设计的神经网络。

设计 RNN 控制器对新产生的网络结构进行评估。文献[8]采用强化学习算法,将网络准确率作为反馈返回给 RNN 控制器,RNN 控制器继续指导生成符合期望的更优神经网络;文献[9]是基于梯度算法的代表 DARTS,DARTS 将网络结构设计为一个有向无环图,图的顶点代表网络中的卷积操作,图的边代表卷积层之间的连接与否与连接方式,将候选操作使用 softmax 函数进行混合,网络结构变得可微分后,使用基于梯度算法的优化算法寻找最优结构;文献[10]使用进化算法自动搜索神经网络结构,将初始模型学习率设置为 0.1,搜索共耗时 256.2 h,最终获得期望的网络结构。

1.3 结构搜索加速

在搜索策略基本已经成熟的情况下,网络结构搜索研究转向减少算法所需时间和计算资源,进化深度网络^[11]等越来越多对计算资源要求较低的搜索神经网络算法被提出。文献[12]提出权重共享思想,重用训练好网络中的权重,改变初代网络结构搜索每个模型从头开始训练的算法,采用在大图中搜索子图的思想;文献[13]采用代理模型思想,设计代理模型 LSTM 来预测生成网络结构在目标数据集上的准确性,指导模型搜索。

2 网络模型压缩

2.1 具有线性相位约束的 1×1 卷积

FIR 滤波器的单位冲激响应为 $h(n), 0 \leq n \leq N-1$ ^[14],系统函数为 $H(z) = \sum_{n=0}^{N-1} h(n)z^{-n}$ 。可以看出,FIR 滤波器共有 4 种类型: $h(n)$ 偶对称, N 为偶数; $h(n)$ 偶对称, N 为奇数; $h(n)$ 奇对称, N 为偶数; $h(n)$ 奇对称, N 为奇数^[15]。4 种 FIR 滤波器的相位特性只取决于 $h(n)$ 的对称性,与 $h(n)$ 的取值无关。

本文根据 FIR 滤波器的特性,设计具有线性相位约束的对称滤波器,将权重在深度方向上对称或者反对称的卷积滤波器定义为广义具有线性相位约束的对称滤波器。将线性相位约束思想应用在 1×1 卷积滤波器上,设计 4 种具有线性相位约束的 1×1 对称滤波器,如图 1 所示,将 4 种 1×1 对称滤波器命名为 I、II、III、IV 型具有线性相位约束的对称滤波器。

输入特征图尺寸为 $D_K \times D_K \times M$ 时:常规 N 个 1×1 卷积核计算时的参数量为 $1 \times 1 \times M \times N$,计算量为 $1 \times 1 \times M \times N \times D_K \times D_K$;具有线性相位约束

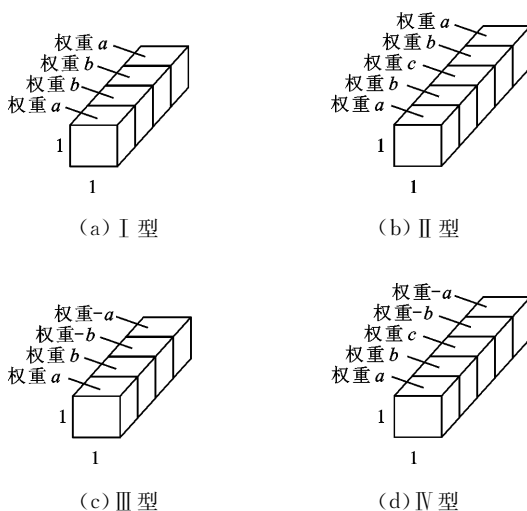


图1 具有线性相位约束的 1×1 对称滤波器结构

的 1×1 对称滤波器计算时的参数量为 $1 \times 1 \times M/2 \times N$,计算量为 $1 \times 1 \times M/2 \times N \times D_K \times D_K$ 。可以看出,具有线性相位约束的对称滤波器与标准 1×1 卷积相比,节省了一半参数量和计算量。I 型具有线性相位约束的 1×1 对称滤波器对输入特征图的具体操作如图 2 所示,其他 3 种具有线性相位约束的对称滤波器与此相似。

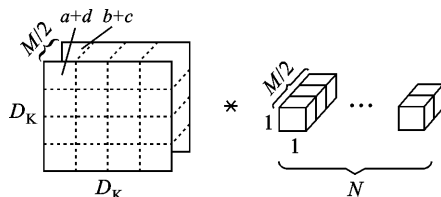
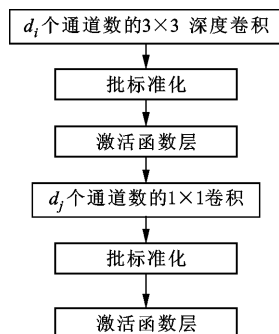


图2 线性相位约束的 1×1 卷积操作示意

2.2 具有线性相位约束的 MobileNet

MobileNet 结构中的深度可分离卷积基础结构如图 3 所示。由于 MobileNet 网络中含有大量的 1×1 卷积操作,所以 MobileNet 网络为基础,验证本文设计的具有线性相位约束的对称滤波器在网络模型压缩方面的性能。



d_i — 3×3 卷积滤波器数; d_j — 1×1 卷积滤波器数。

图3 深度可分离卷积基础结构

对 MobileNet 网络中的 1×1 卷积滤波器均做线性相位约束,对尺寸为 $224\times 224\times 3$ 的输入图像进行处理。添加具有线性相位约束对称滤波器后的 MobileNet 网络结构如表 1 所示。

表 1 具有线性相位约束对称滤波器的 MobileNet 网络结构

网络层	步长	输出图像尺寸	具有线性相位约束 对称滤波器的比例
输入图像		$224\times 224\times 3$	
conv_1	2	$112\times 112\times 32$	
block_1	1	$112\times 112\times 64$	a_1
block_2	2	$56\times 56\times 128$	a_2
block_3	1	$56\times 56\times 128$	a_3
block_4	2	$28\times 28\times 256$	a_4
block_5	1	$28\times 28\times 256$	a_5
block_6	2	$14\times 14\times 512$	a_6
block_7~ block_11	1	$14\times 14\times 512$	$a_7\sim a_{11}$
block_12	2	$7\times 7\times 1\,024$	a_{12}
block_13	1	$7\times 7\times 1\,024$	a_{13}
全局池化层	2	$1\times 1\times 1\,024$	
全连接层		$1\times 1\times 200$	
softmax 层		200	

3 网络结构搜索及加速

3.1 遗传算法搜索网络结构

将网络中的卷积滤波器均做线性相位约束处理后,网络参数量大幅减少,网络准确率也损失较大。本小节研究对网络中的部分滤波器进行 I 型线性相位约束处理。

对网络深度可分离卷积中的 1×1 卷积进行线性相位约束处理。 1×1 逐点卷积层中包括 $(1-a')d_j$ 个标准 1×1 卷积滤波器, $a'd_j$ 个具有线性相位约束的 1×1 卷积滤波器。 1×1 逐点卷积层最后的操作为将这两种不同滤波器计算后的输出结果拼接在一起,再进行 ReLU 函数非线性化处理。线性相位约束的深度可分离卷积结构示意图如图 4 所示。

遗传算法主要流程如图 5 所示。本文使用遗传算法优化搜索网络结构的具体步骤如下。

(1) 编码和解码。本文的搜索参数为 MobileNet 网络 13 个 1×1 卷积层中,每层包含的具有线性相位约束的滤波器比例。采用二进制编码方式^[17],编码遵守完备性、健全性、非冗余性准则。

遗传算法中,染色体的每 3 个基因点数值表示

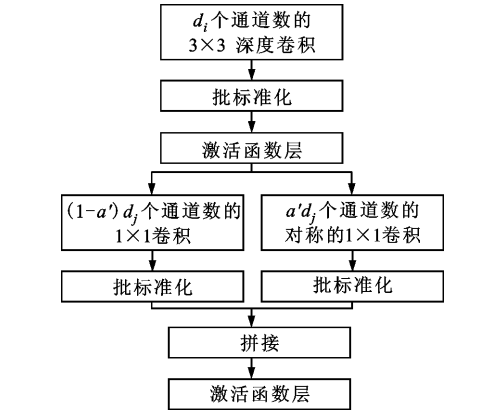


图 4 线性相位约束的深度可分离卷积结构示意图

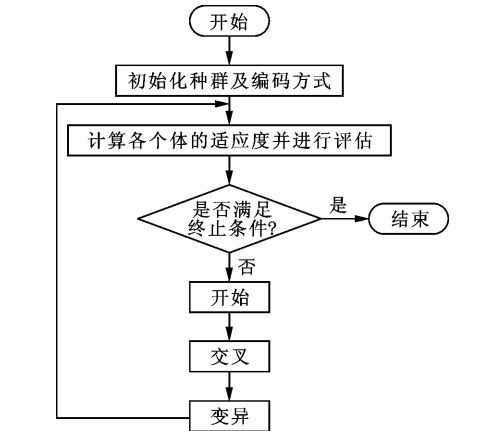


图 5 遗传算法流程^[16]

MobileNet 结构中每层包含具有线性相位约束的对称滤波器比例,13 层卷积层对应的染色体长度为 39 位二进制数。染色体结构如图 6 所示。

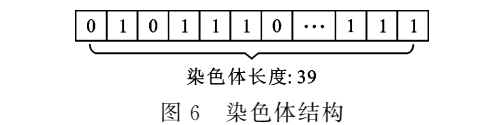


图 6 染色体结构

(2) 种群初始化。采用随机数生成函数随机初始化个体初始基因,设定种群中个体数为 400,种群进化代数为 500。

(3) 种群评估。将不同的具有线性相位约束的对称滤波器比例的网络结构在 Cifar10 数据集下的准确率作为个体适应度,指导遗传算法下一步方向。

(4) 选择操作。采用轮盘赌算法进行个体选择,以与适应度成正比的概率来确定每个个体遗传到下一代种群的数量。

(5) 交叉操作。采用单点交叉方式,互换种群个体中染色体上一个点的基因,产生新染色体。

(6) 变异操作。采用二进制编码方式,由于染色体上信息只有 0、1 两种,变异操作对需要变异的点

进行取反,本文采用单点变异的变异方式。

(7)终止判断。在遗传算法进化到一定代数时,种群中的大多数个体基因进化到大部分一致。若种群在固定范围代数内不再继续进化,则认为遗传算法搜索到目标函数的全局最优解。将此时种群中的染色体进行解码,获得本文需要的具有线性相位约束的滤波器的最优 MobileNet 网络结构。

3.2 权重共享加速结构搜索

在网络结构搜索过程中,遗传算法搜索出的大部分网络子模型结构相似。许多模型中的 $W_{i,j}$ (第 i 个节点与第 j 个节点的权重矩阵)可以重复利用,将前期训练好的结构模型权重矩阵赋予新生成的网络结构,直接测试新生成网络的准确率。遗传算法无需获得网络的最终准确率,可以根据同等条件下的相对准确率判断网络性能的优劣。本文采用相对准确率作为适应度来判断网络结构的优劣^[18]。

采用权重共享算法缩短遗传算法搜索时间,该算法采用在大图中搜索子图的思想。先设计将所有子网络结构都包括在内的大网络结构,称之为大图。算法对大图网络结构进行一次完整训练测试后保存权重矩阵,新生成的子网络结构均重用大图中保存的权重矩阵。

MobileNet 网络中,每层 1×1 卷积滤波器添加线性相位约束的比例范围为 $0\sim 100\%$ 。本文选择牺牲搜索空间来缩短搜索时间。首先将 MobileNet 网络中每层具有线性相位约束的滤波器比例固定设置在 30% 、 50% 、 70% ,然后进行搜索。在此种比例设定下搜索出最优网络结构后,再对网络结构中添加具有线性相位约束的对称滤波器比例进行细化。

本文共设计 3 种比例划分的搜索结构,添加具有线性相位约束的滤波器的比例分别为:① 30% 、 50% 、 70% ,设计的大图和子图结构分别如图 7 和图 8 所示;② 20% 、 40% 、 60% 、 80% ;③ 10% 、 20% 、 30% 、 40% 、 50% 、 60% 、 70% 、 80% 、 90% 。

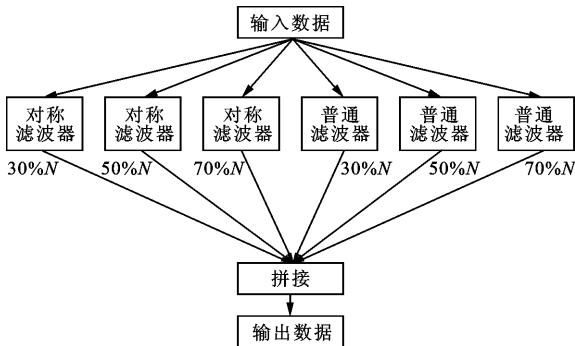


图 7 大图结构

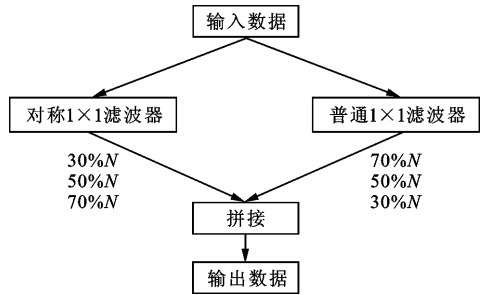


图 8 子图结构

4 实验与分析

4.1 实验环境

实验使用 Python 语言,在开源深度学习框架 Pytorch0.4 上进行实验,采用 GPU 进行加速,GPU 配置为 4 块 NVIDIA GeForce GTX 1080 显卡。在 Cifar10 和 ImageNet 数据集上验证具有线性相位约束的滤波器的网络在网络压缩方面的有效性。在 Cifar10 数据集上对网络结构进行优化搜索。

4.2 实验结果

4.2.1 网络模型压缩效果 将 4 种具有线性相位约束的滤波器的新 MobileNet 网络与未进行修改的 MobileNet 网络在 Cifar10、ImageNet 数据集上进行对比,准确率以及参数量如表 2 和表 3 所示。本文实验结果中的 MobileNet 网络准确率与主流文献中 MobileNet 网络的准确率^[19]一致。

表 2 Cifar10 数据集上网络的准确率及参数量对比

添加滤波器类型	准确率/%	参数量
无	86.29	3.22×10^6
I 型具有线性相位约束的滤波器	85.91	1.65×10^6
II 型具有线性相位约束的滤波器	85.54	1.65×10^6
III 型具有线性相位约束的滤波器	85.50	1.65×10^6
IV 型具有线性相位约束的滤波器	85.55	1.65×10^6

从表 2 和表 3 可以看出:对于 I 型具有线性相位约束的滤波器的 MobileNet 网络,在 Cifar10 数据集上的参数量降低为原始 MobileNet 网络的 51.24% ,准确率下降了 0.38% ,在 ImageNet 数据集上的参数量降低为原始 MobileNet 网络的 62.88% ,top5 准确率下降了 1.44% 。实验证明,4 种具有线性相位约束的 1×1 滤波器均有网络模型压缩的效果,在参数量大幅度减少的前提下,准确率仅有小幅度下降,4 种滤波器在网络模型压缩方面表现持平。

表 3 ImageNet 数据集上网络的准确率及参数量对比

添加滤波器 类型	top1 准确率 /%	top5 准确率 /%	参数量
无	70.62	89.74	4.23×10^6
I 型具有线性相位 约束的滤波器	68.41	88.30	2.66×10^6
II 型具有线性相位 约束的滤波器	68.15	88.40	2.66×10^6
III 型具有线性相位 约束的滤波器	68.09	88.28	2.66×10^6
IV 型具有线性相位 约束的滤波器	68.00	88.21	2.66×10^6

4.2.2 网络模型压缩对比 MobileNet 引入宽度因子 α 控制模型通道数,进一步压缩了网络模型。当 $\alpha=0.75$ 时,网络准确率为 85.91%,参数量为 1.82×10^6 。结合表 2 可知,添加了 I 型具有线性相位约束的滤波器的 MobileNet 网络的准确率与宽度因子 $\alpha=0.75$ 时的原始 MobileNet 网络的准确率相当,但是具有线性相位约束的对称滤波器的网络参数量更少。

4.2.3 网络结构搜索结果 遗传算法搜索出的结构如下。

MobileNet-(1): $a_1 = 70\%, a_2 = 70\%, a_3 = 30\%, a_4 = 30\%, a_5 = 30\%, a_6 = 50\%, a_7 = 50\%, a_8 = 70\%, a_9 = 50\%, a_{10} = 50\%, a_{11} = 30\%, a_{12} = 50\%, a_{13} = 50\%$ 。

MobileNet-(2): $a_1 = 20\%, a_2 = 40\%, a_3 = 20\%, a_4 = 80\%, a_5 = 80\%, a_6 = 20\%, a_7 = 60\%, a_8 = 40\%, a_9 = 40\%, a_{10} = 20\%, a_{11} = 80\%, a_{12} = 20\%, a_{13} = 20\%$ 。

MobileNet-(3): $a_1 = 10\%, a_2 = 40\%, a_3 = 30\%, a_4 = 20\%, a_5 = 30\%, a_6 = 50\%, a_7 = 60\%, a_8 = 30\%, a_9 = 70\%, a_{10} = 20\%, a_{11} = 20\%, a_{12} = 60\%, a_{13} = 80\%$ 。

在 Cifar10 数据集上,遗传算法搜索出的网络准确率及参数量如表 4 所示。

通过遗传算法指导网络结构搜索,可以合理优化 MobileNet 网络中添加具有线性相位约束的卷积滤波器的比例。该算法可以通过调整具有线性相位约束对称滤波器组合的方式,降低网络参数量,提高网络准确率。

表 4 遗传算法搜索出的网络参数量及准确率对比

网络类型	准确率/%	参数量
原始网络	86.29	3.22×10^6
MobileNet-(1)	86.12	2.44×10^6
MobileNet-(2)	86.31	2.69×10^6
MobileNet-(3)	86.25	2.32×10^6

5 结 论

本文根据 FIR 滤波器特性,针对轻量化神经网络中占据网络模型大部分运算量和参数量的 1×1 卷积,设计了 4 种具有线性相位约束滤波器,用于压缩网络模型。在包含大量 1×1 卷积的 MobileNet 网络上进行了实验验证,证实了 4 种具有线性相位约束的对称滤波器均有网络模型压缩效果,且网络性能得以保证。4 种滤波器在网络模型压缩方面表现持平,I 型具有线性相位约束的滤波器在 Cifar10 数据集上,网络参数量降低为原始网络的 51.24%,网络准确率下降 0.38%,在 ImageNet 数据集上,网络参数量降低为原始网络的 62.88%,网络 top5 准确率下降 1.44%。

采用网络结构搜索算法,设定网络中每层具有线性相位约束的对称滤波器的比例为优化目标,对网络结构进行自动搜索调优。遗传算法搜索出的最优结构网络准确率与 MobileNet 网络的相仿,参数量下降为原始网络的 83.54%。具有线性相位约束的 1×1 卷积滤波器适用于存储资源有限的硬件,易于部署到存储和计算资源有限的嵌入式设备上。

参考文献:

[1] BALAPRAKASH P, EGELE R, SALIM M, et al. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research [C] // Proceedings of the 2009 International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ, USA: IEEE, 2019: a37.

[2] SUGANUMA M, SHIRAKAWA S, NAGAO T. A genetic programming approach to designing convolutional neural network architectures [C] // Proceedings of the 2017 Genetic and Evolutionary Computation Conference. New York, USA: ACM, 2017: 497-504.

[3] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recogni-

- tion (CVPR). Piscataway, NJ, USA: IEEE, 2015: 1-9.
- [4] IANDOLA F N, HAN Song, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1602.07360>.
- [5] HOWARD A G, ZHU Menglong, CHEN Bo, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1704.04861>.
- [6] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1707.01083>.
- [7] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 8697-8710.
- [8] ZOPH B, LE Q V. Neural architecture search with reinforcement learning [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1611.01578>.
- [9] LIU Hanxiao, KAREN S, YANG Yiming. DARTS: differentiable architecture search [C/OL]// Proceedings of the 7th International Conference on Learning Representations. London, UK: ICLR, 2019. [2020-03-01]. <https://arxiv.org/abs/1806.09055>.
- [10] REAL E, MOORE S, SELLE A, et al. Large-scale evolution of image classifiers [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1703.01041>.
- [11] DUFOURQ E, BASSETT B A. EDEN: evolutionary deep networks for efficient machine learning [C]// Proceedings of the 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). Piscataway, NJ, USA: IEEE, 2017: 110-115.
- [12] PHAM H, GUAN M Y, ZOPH B, et al. Efficient neural architecture search via parameter sharing [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1802.03268>.
- [13] LIU Chenxi, ZOPH B, NEUMANN M, et al. Progressive neural architecture search [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/1712.00559>.
- [14] 郑南宁, 程洪. 数字信号处理 [M]. 北京: 清华大学出版社, 2007: 120-132.
- [15] DENG Libao, SUN Haili, ZHANG Lili. A new algorithm (ESA-DE) for designing FIR digital filters [C]// Proceedings of the 2019 International Conference on Wireless and Satellite Systems. Cham, Germany: Springer, 2019: 640-652.
- [16] GOLDBERG D E. Genetic algorithm in search, optimization, and machine learning [EB/OL]. [2020-03-01]. https://www.researchgate.net/publication/30870312_Genetic_Algorithms_in_Search_Optimization_and_Machine_Learning.
- [17] YANG Yu, LI Hongzhi, YAO Mingyu, et al. Optimizing the size of a printed circuit heat exchanger by multi-objective genetic algorithm [J]. Applied Thermal Engineering, 2020, 167: 114811.
- [18] JIN Haifeng, SONG Qingquan, HU Xia. Auto-Keras: an efficient neural architecture search system [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2019: 1946-1956.
- [19] BETHGE J, BARTZ C, YANG HAOJIN, et al. MeliusNet: can binary neural networks achieve MobileNet-level accuracy? [EB/OL]. [2020-03-01]. <https://arxiv.org/abs/2001.05936>.

(编辑 陶晴)