DOI：

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方法，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，验证了其有效性。采用遗传算法对 MobileNet 网络中具有线性相位约束 1×1 卷积滤波器比例进行搜索，使用权重共享方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，此时网络参数量下降为原始网络 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

Model Compression and Structure Search for Lightweight Neural Network

LIANG Feng[1], DONG Ming[2], TIAN Zhichao[1], ZHANG Guohe[1], CHENG Shuting[1]

（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 1x1 convolutional filters have limited the compression ratio of the network model. In this paper, 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 performed 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 adopted 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 data set, 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 data set, the network parameter amount is reduced to 62.88%, the network Top5 accuracy rate only dropped by 1.44%. The 1×1 convolution filters with the linear phase constraints can further compress the network. The 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 needs 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（Lightweight Neural Network Model），是专门针对嵌入式视觉应用终端设计的轻量且高效神经网络模型。相较于传统神经网络主要优点有：网络结构更加简单，网络学习和推理需要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性能上限。

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：国家自然科学基金资助项目（61474093）；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）

2017 年 5 月，谷歌推出自动机器学习(AutoML)，越来越多自动优化神经网络结构方法被提出，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(Neural Architecture Search, NAS)，通过算法自动搜索神经网络结构，从过程上可以分为三个步骤：定义目标搜索空间、根据算法中选择的搜索策略产生新网络结构、评估搜索出网络结构优劣并反馈。目前主流搜索神经网络结构算法包括：基于梯度方法、强化学习[1]、进化算法[2]等。

本文主要分为设计网络模型压缩方法部分和优化搜索网络结构部分。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主要部分[3]，限制网络模型压缩。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 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，有利于对网络模型进行压缩，在设计轻量化神经网络时有很大的优势。

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 网络中具有线性相位约束 1×1 卷积滤波器比例进行搜索，该方法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，大幅度提高神经网络经济价值。Iandola 等人提出先利用 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 AlexNet 相同准确率，参数量只有 AlexNet 的 1/50[4]。MobileNet 网络不仅提出了深度可分离卷积结构，将原始 3×3 卷积进行了拆分，大幅度降低网络参数量，还提出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 采用先升维再降维方式增强网络中间深度卷积层表现能力，并移除降维层激活函数。使用反向残差单元的 MobileNet-v2 在多种视觉任务数据集上获得更高准确率，并且计算需要计算资源大幅减少。Zhang 等人提出 ShuffleNet[6]，为了解决网络结构中大量密集 1×1 卷积导致网络性能低下问题，提出了逐点分组卷积(pointwise group convolutions)思想，使用两个 1×1 组卷积将通道间学习分解为两步。同时采用通道混合(channel shuffle)方法促进各通道间信息融合，group 卷积能够获得不同 group 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，实现搭建更强有力网络结构。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 堆叠方式[7]，使网络应用范围更加广泛。

上述经典轻量化神经网络模型压缩方法被广泛应用于计算机视觉中，摆脱高昂计算资源限制，实现神经网络从实验室走向工业，提高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，例如图像识别上有着非常好的表现，但是在设计神经网络时存在网络结构难以确定、网络训练参数难以选取、网络权值较难设定等难题。以上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设计网络，不利于神经网络发展以及推广。神经网络结构搜索空间面向参数为定义网络结构参数，具体为网络层数、卷积核大小、每层选取算子等参数。NAS 为近年来深度学习领域研究热点，通过高效经济搜索方法，自动获得泛化能力强、硬件友好神经网络结构，节省人工设计网络时间，并且在多种场景下击败了早期人工设计神经网络。

设计 RNN 控制器对新产生网络结构进行评估。文献[8]采用了强化学习方法，将网络 accuracy 作为 reward 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]基于梯度方法的代表 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卷积操作，图的边代表卷积层之间连接方式与连接与否。将候选操作使用 softmax 函数进行混合，网络结构变得可微后，使用基于梯度算法优化方法找寻最优结构。文献[10]使用了进化算法自动搜索神经网络结构，该论文从学习率为 0.1 初始模型进行搜索，耗时 256.2h 完成实验，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情况下，目前 NAS 主要研究转向为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消耗要求低的搜

索神经网络方法被提出。文献[12]提出权重共享思想，重用训练好网络中权重，改变初代 NAS 每个模型从头开始训练方法，采用在大图中搜索子图思想。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 1×1 卷积

FIR 滤波器[14]的单位冲激响应为：

$$h(n)\quad 0\le n\le N-1 \tag{1}$$

系统函数为：

$$H(z)=\sum_{n=0}^{N-1}h(n)z^{-n} \tag{2}$$

由 FIR 滤波器频率响应公式中参数不同取值可知共有四种类型线性相位 FIR 滤波器[15]：$h(n)$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，四种 FIR 数字滤波器相位特性只取决于 $h(n)$的对称性，与 $h(n)$的值无关。

本文根据 FIR 滤波器的特性，设计具有线性相位约束对称滤波器，将权重在深度方向上对称或者反对称卷积滤波器定义为广义具有线性相位约束的对称滤波器。本文将线性相位约束思想应用在 1×1 卷积滤波器上，可设计四种 1×1 对称滤波器，具体示意图如图 1 所示。本文将四种 1×1 对称滤波器命名为 I、II、III、IV型具有线性相位约束对称滤波器。

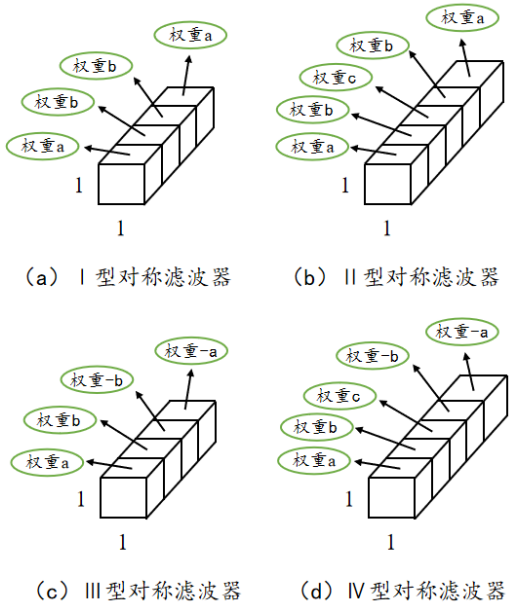

图 1 线性相位约束 1×1 卷积结构

输入特征图尺寸为 $D_K\times D_K\times M$ 时，常规 N 个 1×1 卷积核计算时参数量为：$1\times1\times M\times N$，计算量为：$1\times1\times M\times N\times D_K\times D_K$。具有线性相位约束的 1×1 对称滤波器计算时参数量为：$1\times1\times M/2\times N$，计算量为 $1\times1\times M/2\times N\times D_K\times D_K$。具有线性相位约束对称滤波器与标准 1×1 卷积相比，节省了 1/2 参数量和计算量。I型 1×1 对称滤波器对输入特征图的具体操作如图 2 所示，其余三种对称滤波器也与此相似。

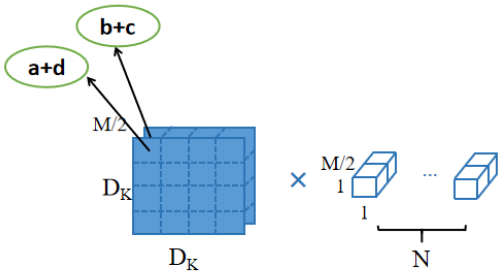

图 2 线性相位约束 1×1 卷积操作示意图

2.2 具有线性相位约束的 MobileNet

MobileNet 结构中深度可分离卷积基础结构如图 3(a)所示，由于 MobileNet 网络中含有大量的 1×1 卷积操作，本文以 MobileNet 网络为基础，验证本文提出的设计具有线性相位约束特性的对称滤波器在网络模型压缩方面的性能。

将 MobileNet 网络中的 1×1 卷积滤波器均做线性相位约束处理后，添加具有线性相位约束后的 MobileNet 相较于原始的 MobileNet 对于输入图片尺寸为 224×224×3 处理整个网络结构如表 1。

表 1 网络结构

网络层	步长	输出图片尺寸	比例
输入图片	-	224×224×3	-
Conv_1	2	112×112×32	-
Block_1	1	112×112×64	a_1
Block_2	2	56×56×128	a_2
Block_3	1	56×56×128	a_3
Block_4	2	28×28×256	a_4
Block_5	1	28×28×256	a_5
Block_6	2	14×14×512	a_6
Block_7-11	1	14×14×512	a_7, a_8, a_9, a_{10}, a_{11}
Block_12	2	7×7×1024	a_{12}
Block_13	1	7×7×1024	a_{13}
GlobalPool	2	1×1×1024	-
FC	-	1×1×200	-
Softmax	-	200	-

3 网络结构搜索及加速

3.1 遗传算法搜索网络结构

将网络中卷积滤波器均做线性相位约束处理后，网络参数量大幅度减少，网络准确率也损失较多。本章研究对网络中部分滤波器进行I型线性相位约束处理。

对网络深度可分离卷积中 1×1 卷积进行线性相位约束处理，a 代表具有线性相位约束滤波器个数占该层总滤波器数 d_j 比例。1×1 逐点卷积层中包括$(1-a)\times d_j$ 个标准 1×1 卷积滤波器，和 $a\times d_j$ 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，对于一个深度可分离卷积层的网络结构示意图如图 3(b)所示。

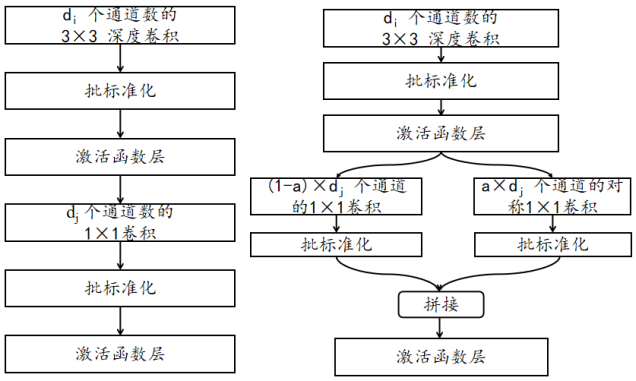

图 3(a)深度可分离卷积　(b)线性相位约束深度可分离卷积

遗传算法主要流程[16]如图所示。

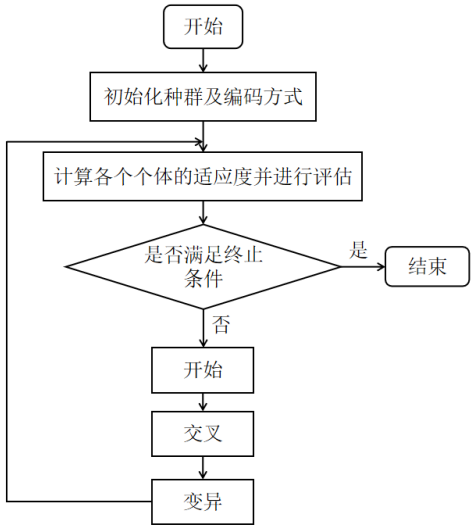

图 4 遗传算法流程

本文中使用遗传算法优化搜索网络结构的具体步骤为：

（1）编码和解码。本文实验中，搜索空间为 MobileNet 网络结构的 13 层 1×1 卷积层中，每层具有线性相位约束滤波器比例。实验采用二进制编码方式[17]，以每层对称滤波器比例按 20%、40%、60%、80%划分的网络结构举例详述本文编码方式，其余比例划分均采用此方法。

MobileNet结构中每层对称滤波器比例用三个基因点来表示，13 层卷积层对应的染色体长度为 39 位二进制数。具体染色体结构为：

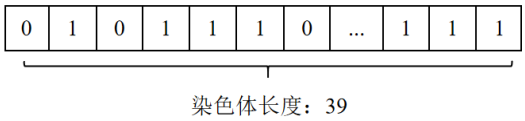

图 5 染色体结构

本实验中的编码遵守完备性、健全性、非冗余性准则。

（2）种群初始化。采用随机数生成函数，随机初始化本实验中个体初始基因，设定种群中个体数为：400，选择种群进化代数为：500。

（3）种群评估。将不同对称滤波器比例的网络结构在 Cifar10 数据集准确率高低作为个体适应度，指导遗传算法下一步方向。

（4）选择操作。采用轮盘赌(Rollete wheel)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数量。算法随机生成一个 0 到 1 之间的随机数 r，如果 $q_i>r$，则 x_i 被选择。轮盘赌算法中个体被选中的概率为：

$$p(x_i)=\frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)} \tag{3}$$

（5）交叉操作。本实验中采用单点交叉方式，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，本文采用了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，种群中大多数个体基因进化到大部分一致。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滤波器最优 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。很多模型中 $W_{i,j}$（第 i 个节点与第 j 个节点权重矩阵）可以重复利用，采用将前期训练好的结构模型权重矩阵赋予新生成网络结构的方式，直接测试新生成网络的准确率。遗传算法无需获得网络最终准确率，可以根据同等条件下相对准确率判断网络性能优劣。本文用相对准确率高低作为适应度值判断网络结构优劣[18]。

本文采用权重共享方法缩短遗传算法搜索时间，该方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。算法中只需要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比例范围为 0%—100%之间。本文选择牺牲搜索空间方法缩短搜索时间，算法最初将添加线性相位约束滤波器比例固定设置在 30%、50%、70%中进行搜索。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对称滤波器比例进行细化。

本文实验中共设计三种比例划分搜索结构：

（1）添加线性相位约束处理滤波器比例：30%、50%、70%，实验设计的大图和子图结构如图所示；

（2）添加线性相位约束处理滤波器比例：20%、40%、60%、80%；

（3）添加线性相位约束处理滤波器比例：10%、20%、30%、40%、50%、60%、70%、80%、90%；

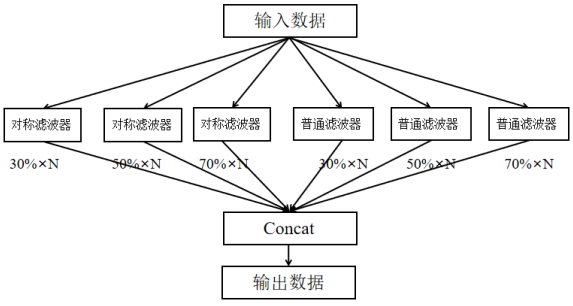

图 6 大图结构

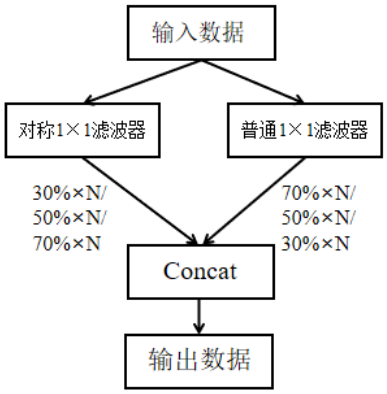

图 7 子图结构

4 实验与分析

4.1 实验环境

本文实验使用了 Python 语言，在开源深度学习 Pytorch0.4 框架上进行实验，采用 GPU 进行加速，GPU 配置为 4 块 NVIDIA GeForce GTX 1080。在 Cifar10 和 ImageNet 数据集上验证将具有线性相位约束滤波器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 四种类型具有线性相位约束滤波器的新 MobileNet 网络在 Cifar10、ImageNet 数据集上实验准确率以及网络参数量对比如表 2、表 3 所示。表 2、表 3、表 5 中的原始网络均指原始的 MobileNet 网络，即未进行修改的 MobileNet 网络,本文的 MobileNet 网络准确率与主流文献一致[19]。

表 2 Cifar10 数据集网络准确率及参数量对比

滤波器类型	网络准确率(%)	参数量
原始网络	86.29	3.22×10^6
I型网络	85.91	1.65×10^6
II型网络	85.54	1.65×10^6
III型网络	85.50	1.65×10^6
IV型网络	85.55	1.65×10^6

表 3 ImageNet 数据集网络准确率及参数量对比

滤波器类型	网络准确率 Top1(%)	Top5(%)	参数量
原始网络	70.62	89.74	4.23×10^6
I型网络	68.41	88.30	2.66×10^6
II型网络	68.15	88.40	2.66×10^6
III型网络	68.09	88.28	2.66×10^6
IV型网络	68.00	88.21	2.66×10^6

由表 2、表 3 可知，I型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%。实验证明，四种类型具有线性相位约束 1×1 滤波器均有网络模型压缩效果，在网络参数量大幅度减少的前提下，网络准确率仅有小幅度下降，四种类型滤波器在网络模型压缩方面表现持平。

4.2.2 模型压缩对比　MobileNet 引入宽度因子 α 控制模型通道数量，进一步压缩网络模型。当 α=0.75 时，宽度因子方法和本文提出的四种滤波器网络模型压缩对比如表 4。

表 4 Cifar10 数据集网络模型压缩效果对比

	网络准确率(%)	参数量
α=0.75	85.91	1.82×10^6
I型网络	85.91	1.65×10^6
II型网络	85.54	1.65×10^6
III型网络	85.50	1.65×10^6
IV型网络	85.55	1.65×10^6

具有线性相位约束的I型网络准确率与宽度因子 α=0.75 的网络相似，但是具有线性相位约束对称滤波器网络参数量更少。

4.2.3 网络结构搜索结果　遗传算法优化网络结构搜索出的结构为:

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%;

遗传算法搜索出的网络结构在 Cifar10 数据集上准确率及参数量，如表 5 所示。

表 5 网络参数量及准确率对比

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

通过遗传算法指导网络结构搜索，可以合理优化 MobileNet 网络中添加线性相位约束的卷积滤波器比例。该方法可以通过调整对称滤波器组合方式，降低网络参数量，提高网络准确率。

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，设计四种具有线性相位约束滤波器用于压缩网络模型。本文在包含大量 1×1 卷积 MobileNet 网络上进行实验验证，证实四种对称滤波器均有网络模型压缩效果，且网络性能得以保证。四种类型滤波器在网络模型压缩方面表现持平，其中I型具有线性相位约束滤波器实现了在 Cifar10 数据集可以达到网络参数量降低为原来 51.24%，网络准确率仅下降 0.38%；在 ImageNet 数据集网络参数量降低为原来 62.88%，网络 Top5 准确率仅下降 1.44%。

同时可以采用网络结构搜索方法，设定优化目标，对具有线性相位约束滤波器网络结构进行自动搜索调优。遗传算法搜索出最优结构可达到网络准确率与 MobileNet 网络相仿情况下，网络参数量下降为原始网络 83.54%，实现在应用对称滤波器对网络结构进行优化时，自动设计参数量更少、性能更好的网络。具有线性相位约束的 1×1 卷积滤波器，对于存储资源有限的硬件十分友好，且易于部署到存储资源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 International Conference for High Performance Computing, Networking, Storage and Analysis. New York, NY, USA: ACM, 2019: a37.

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

[3] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (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 <0.5MB 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]//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]. 2016: arXiv:1611.01578[cs.LG]. https://arxiv.org/abs/1611.01578.

[9] LIU Hanxiao, KAREN S, YANG Yiming. DARTS: differentiable architecture search [C/OL]//7th International Conference on Learning Representations. London, UK: ICLR, 2019. 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]//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. Anchorage AK USA. New York, NY, 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]. https://arxiv.org/abs/2001.05936.

（编辑 陶晴）