

## 基于低开销可变形卷积的MobileNet再轻量化方法

孙长迪, 潘志松, 张艳艳

### 引用本文

孙长迪, 潘志松, 张艳艳. 基于低开销可变形卷积的MobileNet再轻量化方法[J]. 计算机科学, 2022, 49(12): 312-318.

SUN Chang-di, PAN Zhi-song, ZHANG Yan-yan. Re-lightweight Method of MobileNet Based on Low-cost Deformable Convolution [J]. Computer Science, 2022, 49(12): 312-318.

---

### 相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

#### 基于域自适应的肿瘤识别模型

Tumor Recognition Method Based on Domain Adaptive Algorithm

计算机科学, 2022, 49(12): 250-256. <https://doi.org/10.11896/jsjcx.220600008>

#### 改进型FCOS目标检测算法

Improved FCOS Target Detection Algorithm

计算机科学, 2022, 49(11A): 210900220-6. <https://doi.org/10.11896/jsjcx.210900220>

#### 基于改进EfficientNetV2的有害垃圾图像分类方法

Classification Method of Harmful Garbage Images Based on Improved EfficientNetV2

计算机科学, 2022, 49(11A): 211100100-5. <https://doi.org/10.11896/jsjcx.211100100>

#### 基于卷积神经网络的乳腺癌组织病理学图像分类研究综述

Review on Classification of Breast Cancer Histopathological Images Based on Convolutional Neural Network

计算机科学, 2022, 49(11A): 210800232-9. <https://doi.org/10.11896/jsjcx.210800232>

#### 复杂网络社团发现综述

Survey of Community Detection in Complex Network

计算机科学, 2022, 49(11A): 210800144-11. <https://doi.org/10.11896/jsjcx.210800144>

# 基于低开销可变形卷积的 MobileNet 再轻量化方法

孙长迪 潘志松 张艳艳

陆军工程大学指挥控制工程学院 南京 210007

(735120874@qq.com)

**摘要** 近年来,随着无人驾驶、无人机智能化、移动互联网的发展,低功耗、低算力的移动和嵌入式平台对轻量化的神经网络需求日益迫切。文中基于可变形卷积和深度可分离卷积思想的启发,提出了一种低开销可变形卷积,其兼具了可变形卷积的高效特征提取能力和深度可分离卷积的低运算量的优点。此外,在应用低开销可变形卷积的基础上,结合模型结构压缩的方法,设计了4种 MobileNet 网络再轻量化的方法。在 Caltech256, CIFAR100 和 CIFAR10 数据集上进行了实验,结果表明,低开销可变形卷积在运算量增加不明显的情况下,可以有效提高轻量级网络的分类准确度。并且,结合所提出的4种 MobileNet 再轻量化方法,可以将 MobileNet 网络的准确度提高 0.4%~1%,与此同时网络运算量可以下降 5%~15%,即显著提高了轻量化网络的各项性能,更加符合低功耗、低算力的现实需求,对于移动和嵌入式平台领域的智能化推进有着很重要的现实意义。

**关键词:** 低开销可变形卷积;可变形卷积;深度可分离卷积;MobileNet;轻量化;图像分类

**中图法分类号** TP391.4

## Re-lightweight Method of MobileNet Based on Low-cost Deformable Convolution

SUN Chang-di, PAN Zhi-song and ZHANG Yan-yan

College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China

**Abstract** In recent years, with the development of unmanned driving, intelligent UAV and mobile Internet, the demand for lightweight neural network from low-power, low-cost mobile and embedded platforms is increasingly urgent. Based on the idea of deformable convolution and depthwise separable convolution, this paper presents a low-cost deformable convolution, which has the advantages of high-efficiency feature extraction ability of deformable convolution and low computational complexity of depthwise separable convolution. In addition, on the basis of applying low-cost deformable convolution and combining with the method of model structure compression, 4 lightweight methods of MobileNet network are designed. Experiments on Caltech256, CIFAR100 and CIFAR10 datasets demonstrate that low-cost deformable convolution can effectively improve the classification accuracy of lightweight networks without significant increase in computational effort. Besides, the accuracy of the MobileNet network can be improved by 0.4%~1% by combining the 4 MobileNet re-lightening methods in this paper, while the network computing load can be reduced by 5%~15%, which significantly improves the performance of the lightweight network and better meets the practical needs of low power consumption and low computing power. It has very important practical significance for the advancement of intelligence in the field of mobile and embedded platforms.

**Keywords** Low-cost deformable convolution, Deformable convolution, Depthwise separable convolution, MobileNet, Lightweight, Image classification

## 1 引言

自 VggNet<sup>[1]</sup>, ResNet<sup>[2]</sup>等卷积神经网络<sup>[3]</sup> (Convolutional Neural Networks, CNN)问世以来, CNN 在机器视觉领域深耕细作,在不断刷新各大图像数据集的分类精度的同时,也带来了网络模型参数多和运算量大的问题。以 ResNet-152<sup>[2]</sup>网络为例,其在 ImageNet 数据集上 Top-1 准确率可以达到 78.3%, 伴随而来的是多达 20 MB 的海量参数、高达 4.3 GB

的浮点数运算量以及接近 230MB 的存储空间。这对存储空间有限、运算能力一般和功耗限制严格的移动和边缘设备极不友好,间接地限制了卷积神经网络在无人驾驶、无人机智能化、移动端等领域的应用。为了应对这些轻量级平台的智能算法实现需求,网络的轻量化日益受到关注。

目前主流的网络轻量化思路可以分为两类<sup>[3]</sup>:一类是从重构网络的基本单元入手<sup>[4]</sup>,通过设计更高效、更轻便的特征提取单元,在大幅度缩小网络尺寸的同时,仍能取得较好的

到稿日期:2021-12-02 返修日期:2022-05-12

基金项目:国家自然科学基金(62076251)

This work was supported by the National Natural Science Foundation of China(62076251).

通信作者:潘志松 (hotpzs@hotmail.com)

精度。如 2016 年前后提出的 SqueezeNet<sup>[5]</sup>, MobileNet<sup>[6]</sup>, ShuffleNet<sup>[7]</sup>等一系列网络,其中 MobileNet 系列依托深度可分离卷积(Depthwise Separable Convolution, DSC),在精度和运算量上达到了较好的平衡,应用最为广泛。另一类是从压缩网络结构的角度,在现有大型优秀网络的基础上,通过剪枝、蒸馏、二值化等手段<sup>[8]</sup>,对网络结构和参数进行压缩,以达到轻量化网络的目的。以上两类网络轻量化思路的重点都是如何权衡精度和运算量之间的矛盾,即当网络过于轻量化时,其学习能力较弱,难以达到令人满意的精度;当过于侧重提高精度时,网络轻量化效果又很难达到。如何在精度和运算量之间进行折中选择,是轻量化网络研究值得关注的问题。

本文受已有研究的启发,从重构网络的基本单元和压减网络结构两个方向同时入手,寻找出一条改良现有轻量化网络的新思路。首先,结合深度可分离卷积与可变形卷积<sup>[9]</sup>(Deformable Convolution, DC)的优秀思想,改良出一种对轻量化网络更加友好的低开销可变形卷积(Low-cost Deformable Convolution, LDC),并将其应用在现有的轻量化网络上。然后,利用低开销可变形卷积丰富高效的特征学习能力,逐步压减轻量化网络不同深度的结构,寻找出适应不同场景的轻量化网络再改良组合方法。

本文的主要贡献如下:

(1)提出了一种低开销可变形卷积,可以无缝替代现有轻量化网络中的卷积单元,在不改变网络其他结构的情况下,可以增强网络的特征学习能力,提高网络的分类精度。

(2)基于低开销可变形卷积在轻量化网络上的应用,提出了 4 种轻量化网络再调整的方法。实验结果表明,应用低开销可变形卷积与应用可变形卷积的效果相当甚至更好;适当缩减网络的层数,即在减少运算量的同时,仍能保持较好的分类精度。

本文第 2 节主要阐述了近年来轻量化的相关研究工作;第 3 节重点介绍了基于低开销可变形卷积的轻量化方法,并给出了具体的实例化网络;第 4 节通过实验验证了本文方法的有效性;最后总结全文。

## 2 相关工作

卷积神经网络起初是采用标准卷积单元进行特征值提取,即采用固定的卷积大小与形状。近年来提出的许多卷积神经网络改进方法不再拘泥于标准卷积单元,其中可变形卷积和深度可分离卷积应用得最为广泛。

### 2.1 可变形卷积

卷积神经网络在非常规尺寸和未知形态变换等应用场景下存在先天性的缺陷,其原因是卷积核固有的几何形状导致卷积核仅可以对特征图的固定位置进行采样,同层的卷积核的激活单元拥有相同的感受野,这就难以有效应对因位置、尺度、形状发生变化的相同物体<sup>[10]</sup>,如图 1(a)所示,三层  $3 \times 3$  标准卷积的一次采样并不能完全覆盖目标物体。

Dai 等为了解决以上问题,提出了可变形卷积,通过对标准卷积的采样点添加偏移地址,来实现卷积核自适应地定位激活单元,从而使卷积神经网络能自适应聚焦在目标物体

上<sup>[11]</sup>,如图 1(b)所示,三层  $3 \times 3$  可变形卷积的一次采样可以完全覆盖目标物体。

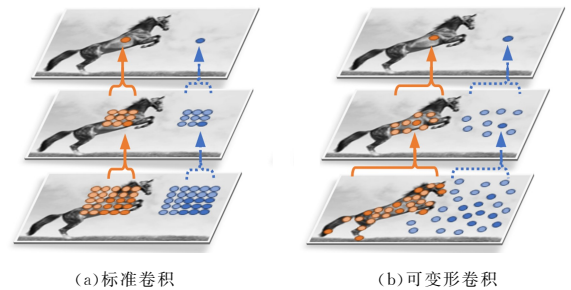


图 1 激活单元对比

Fig. 1 Comparison of activation unit

可变形卷积中的偏移地址通过标准卷积核来获得,仅需要额外增加很少的参数和运算量,通过梯度反向传播算法来融入网络训练,不需要增加额外的标签信息。

假设有一个大小为  $K \times K$  的卷积核,包含  $N(N=K^2)$  个采样点,  $W_n$  和  $P_n$  分别表示第  $n$  个采样点的卷积权重和卷积预先偏移值,当  $K=3$  且  $P_n \in \{(-1, -1), (-1, 0), \dots, (1, 1)\}$  时,表示大小为  $3 \times 3$  的标准卷积核。可变形卷积运算量的计算式如式(1)所示:

$$Y(P) = \sum W_n X(P + P_n + \Delta P_n) \Delta M_n \quad (1)$$

其中,  $X(P)$  表示输入特征图  $X$  中位置  $P$  的特征,  $Y(P)$  表示输出特征图  $Y$  中位置  $P$  的特征,  $\Delta P_n$  和  $\Delta M_n$  分别表示第  $n$  个位置学习到的偏移和调节参数,其中  $\Delta M_n \in [0, 1]$ ,  $P + P_n + \Delta P_n$  限定在  $X$  的尺寸内。

$3 \times 3$  的可变形卷积如图 2 所示,输入的特征图通过普通卷积层(Conv)生成,与输入特征图尺寸相同,但通道数为  $N=3^2$  的调节地址(Scalar Field)和通道数为  $2N$  的偏移地址(Offset Field),即输入特征图的每个特征点上都有  $2N$  个偏移量(Offsets)和  $N$  个调节量(Scalars),它们分别对应该特征点做  $3 \times 3$  卷积操作时(见图 2 中的蓝色部分)  $N$  个采样点水平和垂直维度的偏移量和调节量。

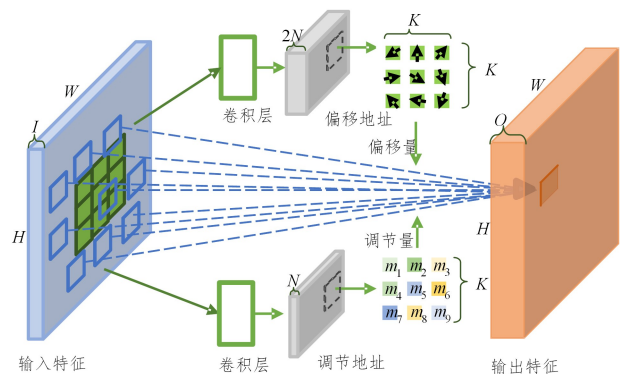


图 2 可变形卷积结构(电子版为彩图)

Fig. 2 Structure of deformable convolution

可变形卷积不仅可以学习卷积采样点的偏移量  $\Delta P_n$ ,还可以学习各卷积采样点的调节量  $\Delta M_n$ 。虽然可变形卷积学习到偏移量  $\Delta P_n$  后,会按照学习的目标形状进行特征采样,但依然会有采样点覆盖到非目标内容的情况,影响网络的判断。为了减小背景噪声等非目标内容对特征输出的影响,可变形

卷积还会学习各个采样点的影响调节量 $\Delta M_n$ ,使网络更加关注目标内容,减小非目标内容的影响。

2.2 深度可分离卷积

2017 年 Google 团队为应对智能手机和穿戴设备日益发展的需要,依托于深度可分离卷积思想,开发出了轻量级卷积神经网络 MobileNet,在算法层面有效地压缩了参数和运算量,已被广泛应用于移动设备和终端设备<sup>[12]</sup>。

深度可分离卷积由深度卷积(Depthwise Convolution)和逐点卷积(Pointwise Convolution)两部分构成。其中,深度卷积本质上属于分组卷积<sup>[13]</sup>(Group Convolution)的一种,即每个深度卷积核只对特征图的某一通道进行卷积(采样),每个特征图只被一个深度卷积核卷积(采样),得到的深度卷积特征图如图 3(a)所示;逐点卷积核在深度卷积特征图的基础上,逐点对深度卷积特征图进行 $1\times 1$ 的卷积(采样),得到的加权融合后的输出特征图如图 3(b)所示。

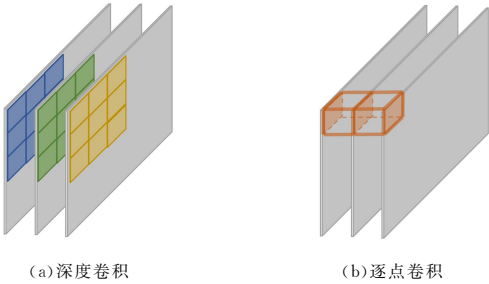


图 3 深度可分离卷积的结构  
Fig. 3 Structure of depthwise convolution

图 4(a)为标准卷积核的示意图,图 4(b)和图 4(c)分别是与其对应的深度卷积核和逐点卷积核的示意图。

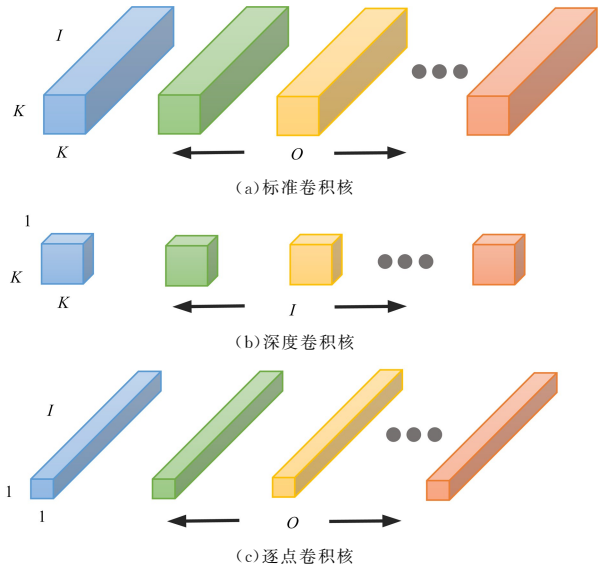


图 4 卷积核尺寸对比  
Fig. 4 Comparison of convolution sizes

图 4 中, $I$  表示输入特征图的通道数, $O$  表示输出特征图的通道数, $K\times K$  表示卷积核的大小, $H\times W$  表示输入特征图的大小。

得到普通卷积的运算量为:

$$K\times K\times I\times O\times H\times W$$

(2)

得到深度卷积的运算量为:

$$H\times W\times I\times K\times K$$

(3)

得到逐点卷积的运算量为:

$$H\times W\times I\times O$$

(4)

应用深度可分离卷积替代标准卷积后,运算量压缩比为:

$$\frac{H\times W\times I\times K\times K+H\times W\times I\times O}{K\times K\times I\times O\times H\times W}=\frac{1}{O}+\frac{1}{K^2}$$

(5)

相关实验表明<sup>[14]</sup>,卷积核大小为 $3\times 3$ 时,用深度可分离卷积替代标准卷积后,网络的运算量可以减少到原来的 $1/9\sim 1/8$ ,而网络的精度变化几乎很小。这就使得 MobileNet 系列网络在训练过程中,可以极大地缩减梯度反向传播所需的时间和运算量,而精度能得到最大程度的保证。

3 基于 LDC 的再轻量化方法

卷积神经网络在图像识别领域的成功,归功于卷积单元可以更加高效地学习到图像的分类特征。然而,卷积单元在参数和运算量上开销较大,不利于实现网络轻量化。为此,目前主流的轻量化网络虽大多仍保留卷积单元,但都将其应用在网络前端,即使用少量的卷积单元来提取简单特征。随后,使用诸如深度可分离卷积等低运算消耗的特征提取单元来整合深度特征,以实现网络分类。

本文方法主要考虑在改良标准卷积单元和网络结构的情况下,进行 LDC 方法的网络设计。

3.1 低开销可变形卷积模块(LDC)

可变形卷积的设计初衷是将其用于大型网络,通过少量增加参数( $1\sim 3\text{ kB}$ )和运算量( $10\sim 80\text{ MB}$ ),来自适应地学习特征采集的偏移与调节量,以大幅提高网络的分类精度。但当将其应用在轻量级网络上时,其增加的参数和运算量就略显庞大了。

本文受现有轻量级网络设计的启发,利用深度可分离卷积(DCS)运算量较少、精度适中的特性,将其运用在可变形卷积的偏移地址(Offset Field)和调节地址(Scalars Field)的学习上,提出了一种深度可分离的可变形卷积,即低开销的可变形卷积,如图 5 所示。其在大幅压缩可变形卷积操作所需运算量的同时,保持不低于可变形卷积的学习能力,且可以与标准卷积无缝替换,对轻量级网络更加友好。

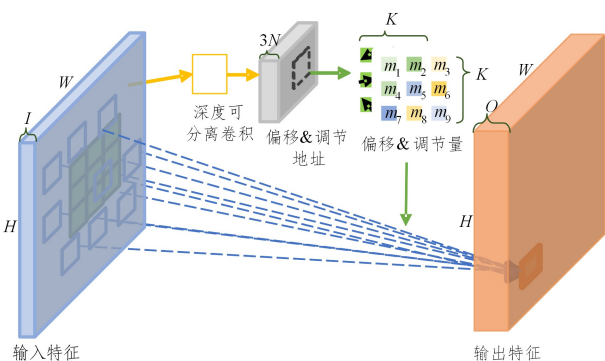


图 5 低开销可变形卷积  
Fig. 5 Low-cost deformable convolution

图 5 中, $I$  表示输入特征图的通道数, $O$  表示输出特征图的通道数, $H\times W$  表示特征图的大小,卷积核大小为 $K\times K$ ,调节地址通道为 $N=K\times K$ ,偏移地址通道为 $2N$ 。

当应用 1 层标准可变形卷积取代 1 层标准卷积时,其运算量为:

$$3N * H * W * I * K^2 + H * W * I * K^2 * O + \Delta$$

(6)

其中, $\Delta$  为固定值,表示除该层卷积外的其他层运算量之和,为常数项。

当应用 1 层低开销可变形卷积取代 1 层标准卷积时运算量为:

$$H * W * I * K^2 + 3N * H * W * I + H * W * I * K^2 * O + \Delta$$

(7)

应用低开销可变形卷积替代传统可变形卷积后,运算量压缩比为:

$$\frac{H * W * I * K^2 + 3N * H * W * I + H * W * I * K^2 * O + \Delta}{3N * H * W * I * K^2 + H * W * I * K^2 * O + \Delta}$$

$$= \frac{4 + O + \Delta'}{3N + O + \Delta'}$$

(8)

其中, $\Delta'$ 同式(6)中的 $\Delta$ ,为常数项。

当  $N=3 \times 3$  时,用 1 层低开销可变形卷积替代标准可变形卷积后,网络的整体运算量可减少 10%,而网络的分类精度有所提高,且在网络的训练过程中,可以极大地缩减梯度反向传播所需的时间和存储空间的占用<sup>[15]</sup>。

3.2 基于 LDC 的轻量级网络再轻量化方法

本文在结合 LDC 模块的基础上,提出了 4 种轻量级网络的再轻量化方法,以验证 LDC 模块在轻量级网络上应用的有效性,并探究其对轻量级网络结构调整的效果。

3.2.1 LDC0 再轻量化方法

在原有轻量级网络的基础上,在不改变原有网络结构的前提下,通过直接引入 LDC 模块无缝取代原有网络的卷积单元,得到再轻量化网络 LDC0。期望通过 LDC 模块强大的自适应特征学习能力,在仅增加网络少量的参数和运算量的情况下,使 LDC0 网络在分类精度上较原网络有更好的表现。

3.2.2 LDC1 再轻量化方法

在 LDC0 再轻量化方法的基础上,尝试改变原有网络结构,对原有轻量级网络的浅层结构进行网络压缩,以得到再轻量化网络 LDC1。同时,从改良特征提取单元和适当压缩网络结构的角入手,期望可以在不增加网络的参数和运算量的情况下,使 LDC1 在分类精度上较原网络有更好的表现。

3.2.3 LDC2 再轻量化方法

与 LDC1 再轻量化方法类似,在 LDC0 再轻量化方法的基础上,对原有轻量级网络的深层结构进行网络压缩,得到再轻量化网络 LDC2。与 LDC1 再轻量化方法做对比,以探究应用 LDC 模块后压缩轻量级网络不同深度网络结构的影响。

3.2.4 LDC3 再轻量化方法

结合 LDC1 和 LDC2 再轻量化方法,在 LDC0 再轻量化方法的基础上,同时对原有轻量级网络的浅层和深层结构进行网络压缩,得到再轻量化网络 LDC3,即同时从改良特征提取单元和大幅压缩网络结构的角入手,期望可以在大幅减少网络的参数和运算量的情况下,使 LDC3 网络在分类精度上仍然有不低于原网络的表现。

3.3 网络实例化

本文选取 MobileNet V3-small(以下简称为 Mobilenet)作为 Baseline 网络,通过本文的方法将 Mobilenet 网络再轻量化,

从而得到 Mobilenet-LDC0, Mobilenet-LDC1, Mobilenet-LDC2, Mobilenet-LDC3,并将其用于实验研究。

Mobilenet 的结构如表 1 所列,其中 Layers 代表层数,Input 代表输入通道,Operator 代表层结构,Exp 代表扩展通道数,Output 代表输出通道。

表 1 Mobilenet 结构表

Table 1 Structure of Mobilenet

| Layers | Input                  | Operator          | Exp | Output |
|--------|------------------------|-------------------|-----|--------|
| 1      | 224 <sup>2</sup> × 3   | Conv2d, 3 × 3     | —   | 16     |
| 2      | 112 <sup>2</sup> × 16  | Benck, 3 × 3      | 16  | 16     |
| 3      | 56 <sup>2</sup> × 16   | Benck, 3 × 3      | 72  | 24     |
| 4      | 28 <sup>2</sup> × 24   | Benck, 3 × 3      | 88  | 24     |
| 5      | 28 <sup>2</sup> × 24   | Benck, 5 × 5      | 96  | 40     |
| 6      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 240 | 40     |
| 7      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 240 | 40     |
| 8      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 120 | 48     |
| 9      | 14 <sup>2</sup> × 48   | Benck, 5 × 5      | 144 | 48     |
| 10     | 14 <sup>2</sup> × 48   | Benck, 5 × 5      | 288 | 96     |
| 11     | 7 <sup>2</sup> × 96    | Benck, 5 × 5      | 576 | 96     |
| 12     | 7 <sup>2</sup> × 96    | Benck, 5 × 5      | 576 | 96     |
| 13     | 7 <sup>2</sup> × 96    | Conv2d, 1 × 1     | —   | 576    |
| 14     | 7 <sup>2</sup> × 576   | Pool, 7 × 7       | —   | —      |
| 15     | 1 <sup>2</sup> × 576   | Conv2d 1 × 1, NBN | —   | 1 280  |
| 16     | 1 <sup>2</sup> × 1 280 | Conv2d 1 × 1, NBN | —   | 1 000  |

Mobilenet 共有 16 层,其中第 6,7 层和第 11,12 层的网络结构基本一致。鉴于本文讨论的是相同参数下不同方法对网络分类精度的影响,当剪除第 6,7,11,12 以外层的网络结构,如第 8 层时,因第 7 层的输出特征维度为 14<sup>2</sup> × 40,而第 9 层的输入特征维度为 14<sup>2</sup> × 48,所以需要引入新的变量来打通输入输出通道,而不同方法又难以统一,势必会影响实验结果。为了保证实验的公平性,同时又不失一般性,选择第 6 层作为浅层代表、第 11 层作为深层代表,进行网络结构压缩,以探究压缩不同深度网络结构所带来的影响。

3.3.1 应用可变形卷积的 Mobilenet-DC 网络

为了增加实验的准确性并增强对比性,本文用可变形卷积替代 Mobilenet 网络的第 1 层(Layers=1),即普通卷积层,所涉及的卷积核大小、步长、边缘扩充等参数保持不变,得到的网络 Mobilenet-DC0 与 Mobilenet 一同作为 Baseline 网络,以验证 LDC 模块的效果,其详细结构如表 2 所列。

表 2 Mobilenet-DC0 结构表

Table 2 Structure of Mobilenet-DC0

| Layers | Input                  | Operator          | Exp | Output |
|--------|------------------------|-------------------|-----|--------|
| 1      | 224 <sup>2</sup> × 3   | DC, 3 × 3         | —   | 16     |
| 2      | 112 <sup>2</sup> × 16  | Benck, 3 × 3      | 16  | 16     |
| 3      | 56 <sup>2</sup> × 16   | Benck, 3 × 3      | 72  | 24     |
| 4      | 28 <sup>2</sup> × 24   | Benck, 3 × 3      | 88  | 24     |
| 5      | 28 <sup>2</sup> × 24   | Benck, 5 × 5      | 96  | 40     |
| 6      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 240 | 40     |
| 7      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 240 | 40     |
| 8      | 14 <sup>2</sup> × 40   | Benck, 5 × 5      | 120 | 48     |
| 9      | 14 <sup>2</sup> × 48   | Benck, 5 × 5      | 144 | 48     |
| 10     | 14 <sup>2</sup> × 48   | Benck, 5 × 5      | 288 | 96     |
| 11     | 7 <sup>2</sup> × 96    | Benck, 5 × 5      | 576 | 96     |
| 12     | 7 <sup>2</sup> × 96    | Benck, 5 × 5      | 576 | 96     |
| 13     | 7 <sup>2</sup> × 96    | Conv2d, 1 × 1     | —   | 576    |
| 14     | 7 <sup>2</sup> × 576   | Pool, 7 × 7       | —   | —      |
| 15     | 1 <sup>2</sup> × 576   | Conv2d 1 × 1, NBN | —   | 1 280  |
| 16     | 1 <sup>2</sup> × 1 280 | Conv2d 1 × 1, NBN | —   | 1 000  |

3.3.2 轻量化 Mobilenet-LDC0 网络

本文将 3.2.1 节中的再轻量化方法应用于 Mobilenet 网络,用低开销可变形卷积替代 Mobilenet 网络的第 1 层(La-

yers=1),即普通卷积层,所涉及的卷积核大小、步长、边缘扩充等参数保持不变,得到 Mobilenet-LDC0 网络,详细结构如表 3 所列。

表 3 Mobilenet-LDC0 结构表  
Table 3 Structure of Mobilenet-LDC0

| Layers | Input                 | Operator       | Exp | Output |
|--------|-----------------------|----------------|-----|--------|
| 1      | 224 <sup>2</sup> ×3   | LDC,3×3        | —   | 16     |
| 2      | 112 <sup>2</sup> ×16  | Benck,3×3      | 16  | 16     |
| 3      | 56 <sup>2</sup> ×16   | Benck,3×3      | 72  | 24     |
| 4      | 28 <sup>2</sup> ×24   | Benck,3×3      | 88  | 24     |
| 5      | 28 <sup>2</sup> ×24   | Benck,5×5      | 96  | 40     |
| 6      | 14 <sup>2</sup> ×40   | Benck,5×5      | 240 | 40     |
| 7      | 14 <sup>2</sup> ×40   | Benck,5×5      | 240 | 40     |
| 8      | 14 <sup>2</sup> ×40   | Benck,5×5      | 120 | 48     |
| 9      | 14 <sup>2</sup> ×48   | Benck,5×5      | 144 | 48     |
| 10     | 14 <sup>2</sup> ×48   | Benck,5×5      | 288 | 96     |
| 11     | 7 <sup>2</sup> ×96    | Benck,5×5      | 576 | 96     |
| 12     | 7 <sup>2</sup> ×96    | Benck,5×5      | 576 | 96     |
| 13     | 7 <sup>2</sup> ×96    | Conv2d,1×1     | —   | 576    |
| 14     | 7 <sup>2</sup> ×576   | Pool,7×7       | —   | —      |
| 15     | 1 <sup>2</sup> ×576   | Conv2d 1×1,NBN | —   | 1 280  |
| 16     | 1 <sup>2</sup> ×1 280 | Conv2d 1×1,NBN | —   | 1 000  |

3.3.3 轻量化 Mobilenet-LDC1 网络

本文将 3.2.2 节中的再轻量化方法应用于 Mobilenet 网络,在 Mobilenet-LDC0 网络的基础上,将 Mobilenet-LDC0 网络的第 6 层(Layers=6)结构移除,得到 Mobilenet-LDC1 网络,探索结合可变形卷积强大的自学习能力来压缩网络浅层结构对网络分类精度的影响,其结构同删除第 6 层(Layers=6)的表 3 所列的结构。

3.3.4 轻量化 Mobilenet-LDC2 网络

本文将 3.2.3 节中的再轻量化方法应用于 Mobilenet 网络,在 Mobilenet-LDC0 网络的基础上将 Mobilenet-LDC0 网络的第 11 层(Layers=11)结构移除,得到 Mobilenet-LDC2 网络,与 Mobilenet-LDC1 网络做对比,探索结合可变形卷积强大的自学习能力压缩网络深层结构对网络分类精度的影响,其结构同删除第 11 层(Layers=11)的表 3 所列的结构。

3.3.5 轻量化 Mobilenet-LDC3 网络

本文将 3.2.4 节中的再轻量化方法应用于 Mobilenet 网络,在 Mobilenet-LDC0 网络的基础上,将 Mobilenet-LDC0 网络的第 6,11 层(Layers=6,Layers=11)结构同时移除,得到 Mobilenet-LDC3 网络,探索结合可变形卷积强大的自学习能力压缩网络多层结构对网络分类精度的影响,其结构同删除第 6,11 层(Layers=6,Layers=11)的表 3 所列的结构。

以上网络分别在 Pytorch 平台上通过 Thop 统计包进行参数和浮点数计算,以输出 1 000 类为例,各网络所需的参数量和浮点数运算量如图 6 所示。

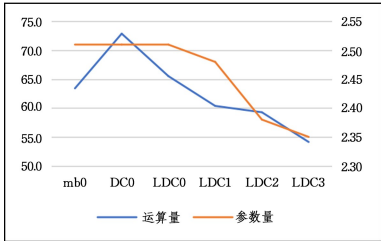


图 6 参数量与运算量统计图

Fig. 6 Statistical chart of parameters and FLOPs

图 6 中,mb0,DC0,LDC0,LDC1,LDC2,LDC3 分别为 Mobilenet 网络、Mobilenet-DC0 网络、Mobilenet-LDC0 网络、Mobilenet-LDC1 网络、Mobilenet-LDC2 网络、Mobilenet-LDC3 网络的缩写。

4 实验及结果分析

4.1 实验环境及数据集

本文的实验平台为 Pytorch 1.7.0 深度学习框架,通过 NVIDIA GTX2080TI 显卡(CUDA 11.2)进行训练。

实验数据集选取 Pytorch 平台下常用的图像数据集,包括 Caltech256<sup>[16]</sup>,CIFAR100,CIFAR10。其中,Caltech256 数据集共有 256 个类别,每个类别的图片超过 80 张,总共有 20 607 张图片。因其没有划分验证集和训练集,故本实验每个类别随机选取该类别总数 1/8 的图片作为验证集,其余作为训练集。

CIFAR100 数据集共有 100 个类别,每个类别有 600 张大小为 32×32 的彩色图像,其中 500 张作为训练集,100 张作为测试集。CIFAR100 属于层次数据集,对于每一张图片,它有 Fine\_labels 和 Coarse\_labels 两个标签,分别代表图像的细粒度和粗粒度标签,本实验仅使用 Fine\_labels 标签进行训练与测试。

CIFAR10 数据集与 CIFAR100 数据集的结构类似,有 10 种类别,由 60 000 张大小为 32×32 的彩色图片组成,每种类别包含 6 000 张图片,总计有 10 000 张测试图片和 50 000 张训练图片,分为 5 个训练批次和 1 个测试批次,其中测试批次是由在每种类别中的随机选择的 1 000 张图片组成,每种类别中的剩余图片共同组成 5 个训练批次。

4.2 实验设置

为了验证本文提出的轻量级网络再轻量化的方法的有效性,本文将应用本文方法所得的实例化网络 MobileNet-LDC0,MobileNet-LDC1,MobileNet-LDC2,MobileNet-LDC3 和 Baseline(即 MobileNet,MobileNet-DC0)网络,先后在 Caltech256,CIFAR100,CIFAR10 数据集上进行训练,作为横向对比实验,以验证不同数据集下各方法的有效性,同一数据集采用相同的超参数设置。同时,通过对比各网络之间的结果,作为纵向对比实验,以验证各方法之间的差异,以区分不同方法的适用场景。

在 Caltech256 数据集上训练时超参数统一设置为: EPOCH=100,Dropout=0,优化器为 SGD,LR=0.1,采用余弦退火调整策略, $T_i=5,\eta_{\min}=0$ 。其计算式为:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)) \tag{11}$$

其中, $\eta_{\max}$  为初始学习率, $\eta_{\min}$  为最小学习率, $T_{\text{cur}}$  为优化器上次重启后的时间间隔个数, $T_i$  为重启间隔数。

在 CIFAR100 数据集上训练时超参数统一设置为: EPOCH=100,Dropout=0.25,优化器为 SGD,LR=0.05,采用余弦退火调整策略, $T_i=5,\eta_{\min}=0$ ,其表达式同式(11)。

在 CIFAR10 数据集上训练时超参数统一设置为: EPOCH=100,Dropout=0.25,优化器为 SGD,LR=0.05,采用余弦退火调整策略, $T_i=5,\eta_{\min}=0$ ,其表达式同式(11)。

通过实验发现,在  $EPOCH=50$  时,各网络在以上数据集上都基本收敛,各数据集上损失值随迭代次数变化而变化的曲线如图 7 所示。

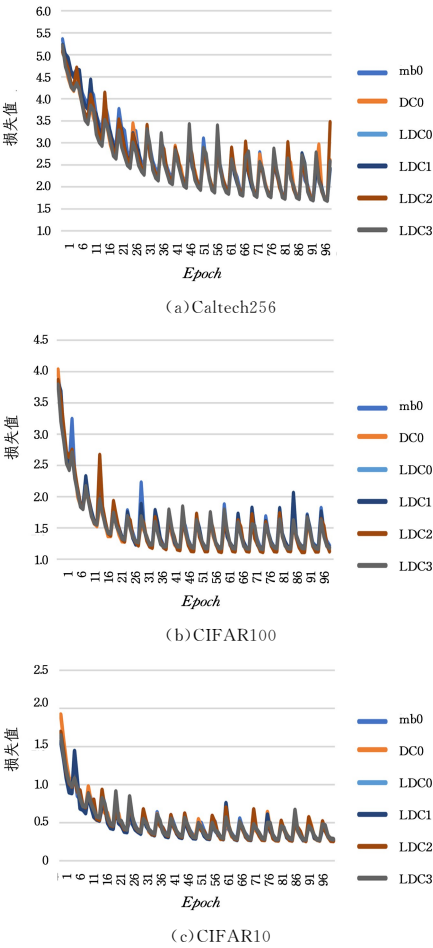


图 7 损失值随迭代次数变化而变化的曲线

Fig. 7 Curve of loss changing with epoch

4.3 评价标准

本文采用 Softmax 作为分类器来进行结果分类输出,其表达式为:

$$Y(P)=P(y=p|x,\theta_p)=\frac{e^{x^T \cdot \theta_p}}{\sum_{p=1}^C e^{x^T \cdot \theta_p}} \tag{9}$$

采用交叉熵作为损失函数来优化网络参数,其表达式为:

$$\min_{(W,b;\theta)} \epsilon(W,b;\theta)=\frac{1}{N} \sum_{n=1}^N [-\sum_{p=1}^C \gamma(y_n=p) \cdot \log P(y=p|x,\theta_p)] + \lambda R(\theta) \tag{10}$$

其中,  $C$  表示类别数,  $N$  表示样本数,  $\gamma$  表示 Dirichlet 函数,参数  $\theta=(\theta_1,\theta_2,\cdots,\theta_k)$ ,  $R(\cdot)$  表示正则化约束项,  $\lambda$  表示正则化系数。

将各网络的分类数都设置为 1 000,以输入  $1 \times 3 \times 224 \times 224$  维度的图像为例,采用 Thop 统计包来统计各网络所需的参数和运算量,并将其作为各网络轻量程度的衡量标准。

4.4 实验结果及分析

本实验对本文涉及的 6 种(含 Baseline)网络进行效果验证,所有实验在相同的实验环境下进行,各数据集下不同网络准确度的变化曲线如图 8 所示。

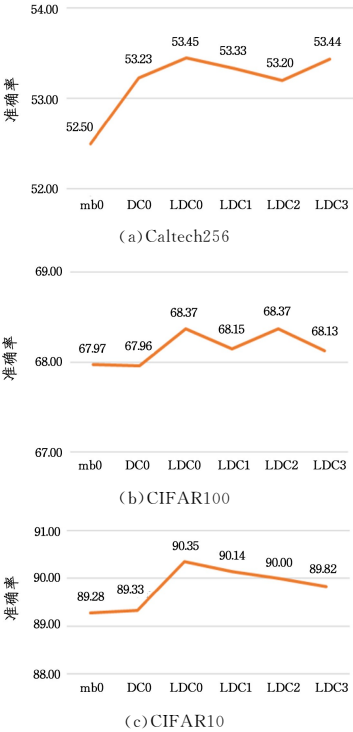


图 8 各数据集下不同网络准确度的变化曲线

Fig. 8 Variation curves of accuracy in different datasets

消融实验的结果如表 4 所列,表中深层代表对网络深层结构进行压缩,浅层代表对网络浅层结构进行压缩。

表 4 消融实验

Table 4 Ablation experiments

| 模型   | 运算量                | 参数量                | CIFAR100     | CIFAR10      | Caltech256   | DC | LDC | 浅层 | 深层 |
|------|--------------------|--------------------|--------------|--------------|--------------|----|-----|----|----|
| mb0  | $63.4 \times 10^6$ | $2.51 \times 10^6$ | 67.97        | 89.28        | 52.50        |    |     |    |    |
| DC0  | $72.9 \times 10^6$ | $2.51 \times 10^6$ | 67.96        | 89.33        | 53.23        | ✓  |     |    |    |
| LDC0 | $65.5 \times 10^6$ | $2.51 \times 10^6$ | <b>68.37</b> | <b>90.35</b> | <b>53.45</b> |    | ✓   |    |    |
| LDC1 | $60.4 \times 10^6$ | $2.48 \times 10^6$ | 68.15        | 90.14        | 53.33        |    | ✓   | ✓  |    |
| LDC2 | $59.3 \times 10^6$ | $2.38 \times 10^6$ | <b>68.37</b> | 90.00        | 53.20        |    | ✓   |    | ✓  |
| LDC3 | $54.1 \times 10^6$ | $2.35 \times 10^6$ | 68.13        | 89.82        | 53.44        |    | ✓   | ✓  | ✓  |

对比 Mobilenet-LDC0 和 Mobilenet 网络的结果, Mobilenet-LDC0 网络的分类准确度提高了 0.4%~1.07%,浮点数运算量仅增加了 2.1MB,证明引入 LDC 模块的网络模型具有更强的特征提取能力,进而在分类精度上表现更好,代价是少量增加了网络的浮点数运算量。

对比 Mobilenet-LDC0 和 Mobilenet-DC0 网络的结果, Mobilenet-LDC0 网络的分类准确度提高了 0.2%~1.02%,浮点数运算量却减少了 7.4 MB,这证明引入 LDC 模块的网络模型比引入 DC 模块的网络模型更加高效,具有更强的特征提取能力和更低的浮点数运算量,即 LDC

模块更加契合轻量级网络。

对比 Mobilenet-LDC2, Mobilenet-LDC1 和 Mobilenet 网络的结果,前两个网络的分类准确度分别提高了 0.4%~0.72%和 0.18%~0.86%,参数量分别减少了 0.13 MB 和 0.03 MB,浮点数运算量分别减少了 4.1 MB 和 3 MB,这证明轻量级网络在引入 LDC 模块并适当缩减网络结构后,网络仍能学习与融合到足够的特征以支撑分类需要,且 LDC2 方法在浮点数运算量和参数量上的表现都优于 LDC1 方法,在不同数据集上的表现波动更小,即泛化能力更好。

对比 Mobilenet-LDC3 和 Mobilenet 网络的结果, Mobilenet-LDC3 网络的分类准确度分别提高了 0.16%~0.94%,参数量减少了 0.16 MB,浮点数运算量减少了 9.3 MB,这证明轻量级网络在引入 LDC 模块,并缩减多层网络结构后,网络仍可以学习到支撑分类需要的特征,但在不同数据集上的表现波动较大,在鲁棒性上 LDC3 方法不如 LDC0 方法。

综上所述,本文提出的 LDC0 方法以少量增加浮点数运算量为代价,较大幅度地提高了网络的分类精度,适用于对参数量和浮点数运算不太敏感,而对精度有进一步要求的应用场景;LDC1 和 LDC2 方法在少量减少网络参数量和浮点数运算量的情况下,仍可以小幅度地提高网络的分类精度,适用于对浮点数运算较敏感,而对精度仍有进一步要求的应用场景,且 LDC2 方法相对更佳;LDC3 方法在大幅度减少网络参数量和浮点数运算量的情况下,仍可以小幅度地提高网络的分类精度,适用于对参数量和浮点数运算量特别敏感,而对精度仍有要求的应用场景。

**结束语** 本文从重构网络的基本单元和压减网络结构两个方向出发,针对轻量化网络再轻量这一现实任务开展研究。受可变形卷积和深度可分离卷积的启发,提出了低开销可变形卷积模块,通过重构卷积单元,强化神经网络的特征提取效能,使之具有高效且自适应的特征提取能力。同时,基于低开销可变形卷积模块,再辅以网络的结构压减,设计出 4 种场景下的轻量级网络再轻量化方法,在不降低轻量级网络分类精度的情况下,能大幅压缩网络的运算量和参数量。最后,依托轻量级 Mobilenet 3-small 网络,对所提的 4 种再轻量化方法开展实例化检验,实验进一步证明了低开销可变形卷积高效的特征提取能力与本文提出的再轻量化方法的有效性。

## 参 考 文 献

- [1] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv:1409.1556, 2014.
- [2] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2016.
- [3] PENGCHENG B I, LUO J, CHEN W, et al. Research on Lightweight Convolutional Neural Network Technology [J/OL]. [http://en.cnki.com.cn/Article\\_en/CJFDTotal-JSGG201916005.htm](http://en.cnki.com.cn/Article_en/CJFDTotal-JSGG201916005.htm).
- [4] CHENG Y, WANG D, ZHOU P, et al. A Survey of Model Compression and Acceleration for Deep Neural Networks[J]. arXiv: 1710.09282, 2017.

- [5] 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]. arXiv:1602.07360, 2016.
- [6] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. arXiv:1704.04861, 2017.
- [7] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[J]. arXiv:1707.01083, 2017.
- [8] DENG L, LI G, HAN S, et al. Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey[J]. Proceedings of the IEEE, 2020, 108(4): 485-532.
- [9] DAI J, QI H, XIONG Y, et al. Deformable Convolutional Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017: 764-773.
- [10] CHEN F, WU F, XU J, et al. Adaptive deformable convolutional network[J/OL]. [https://www.researchgate.net/publication/346289132\\_Adaptive\\_deformable\\_convolutional\\_network](https://www.researchgate.net/publication/346289132_Adaptive_deformable_convolutional_network).
- [11] ZUO Z, ZHANG W, ZHANG D, et al. A remote sensing image semantic segmentation method by combining deformable convolution with conditional random fields[J]. Acta Geodaetica et Cartographica Sinica, 2020, 3(3): 39-49.
- [12] ZHANG Y F. Lightweight of deep learning and its application in image recognition[D]. Jinhua: Zhejiang Teachers' University, 2019.
- [13] HE Y, ZHANG X, SUN J. Channel Pruning for Accelerating Very Deep Neural Networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 1398-1406.
- [14] GUO J, LI Y, LIN W, et al. Network Decoupling: From Regular to Depthwise Separable Convolutions[J]. arXiv: 1808.05517, 2018.
- [15] SUN W, ZHOU X, ZHANG X, et al. A Lightweight Neural Network Combining Dilated Convolution and Depthwise Separable Convolution[C/OL]// 2020. [https://link.springer.com/chapter/10.1007/978-3-030-48513-9\\_17](https://link.springer.com/chapter/10.1007/978-3-030-48513-9_17).
- [16] GRIFFIN G, HOLUB A, PERONA P. Caltech-256 Object Category Dataset[R]. Pasadena (Los Angeles, USA): Caltech Technical Report, 2007.



**SUN Chang-di**, born in 1991, postgraduate. His main research interests include artificial intelligence and pattern recognition.



**PAN Zhi-song**, born in 1973, Ph.D, professor, Ph.D supervisor. His main research interests include artificial intelligence and pattern recognition.