基于特征膨胀卷积模块的轻量化技术研究

许新征[1,2],李 杉[1]
(1. 中国矿业大学计算机科学与技术学院,江苏徐州 221116; 2. 教育部矿山数字化工程中心,江苏徐州 221116)

摘 要: 本文从卷积神经网络模型的网络结构入手,利用特征复用的思想,设计了高效的特征膨胀卷积模块.该模块减少了标准卷积模块的输出通道数,引入了多分支结构.通过各个分支上的廉价操作对标准卷积操作的输出特征图进行变换和融合,产生新的特征图.模块的最终输出由各个分支上生成的特征图进行合并连接得到.特征膨胀卷积模块利用特征复用思想复用模型中的特征,在降低模型计算量的同时,丰富了特征图隐含的信息,提高了模型的性能.最后,将特征膨胀卷积模块代替标准卷积模块,设计了轻量化的VGG16(Visual Geometry Group 16-Layer)模型和残差结构,并且在CIFAR数据集和ILSVRC2012(ImageNet Large Scale Visual Recognition Challenge 2012)数据集上取得了较好的分类效果.

关键词: 卷积神经网络;轻量化;特征复用;特征膨胀卷积;深度学习;图像分类
基金项目: 国家自然科学基金(No.61976217)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2023)02-0355-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210559

Research of Lightweight Convolution Neural Network Based on Feature Expansion Convolution

XU Xin-zheng[1,2], LI Shan[1]
(1. *School of Computer Science and Technology*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*;
2. *Engineering Research Center of Mining Digital Ministry of Education*, *Xuzhou*, *Jiangsu* 221116, *China*)

Abstract: This paper starts with the network structure of convolution neural network model, and uses the idea of feature reuse to design an efficient feature expansion convolution module. The module reduces the number of output channels of standard convolution module and introduces multi branch structure. Through the cheap operation on each branch, the output feature map of standard convolution operation is transformed and fused to generate a new feature map. The final output of the module is obtained by merging the feature graphs generated on each branch. The feature expansion convolution module uses the idea of feature reuse to reuse the features in the model, which not only reduces the calculation of the model, but also enriches the hidden information of the feature graph and improves the performance of the model. Finally, the feature expansion convolution module is used to replace the standard convolution module, and the lightweight VGG16 (Visual Geometry Group 16-Layer) model and residual structure are designed, and good classification results are achieved on CIFAR and ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge 2012) datasets.

Key words: convolutional neural network; lightweight; feature reuse; feature expansion convolution; deep learning; image classification
Foundation Item(s): National Natural Science Foundation of China (No.61976217)

1 引言

自从人工智能诞生以来,得益于硬件技术和软件技术的不断积累和创新,人工智能技术取得了一次又一次的突破性发展[1].硬件技术的发展推动互联网与各个行业深度融合,这个过程催生出了大量的数据,为人工智能的发展提供了必要的土壤[2].自2012年Hinton提出AlexNet网络模型以来[3],深度学习已经经过了十多年的发展.在这十多年里,以反向传播理论为核心,出现了大量的非常有价值的卷积神经网络模型.这些模型在图像处理领域得到了广泛的应用和发展,并

收稿日期:2021-05-05;修回日期:2021-07-21;责任编辑:梅志强

且正在与各个行业的发展紧密结合[4~7]. 然而,尽管硬件技术不断发展,出现了大量的应用于深度学习模型的高性能计算平台,这些硬件设备的计算能力依旧不能满足所有卷积神经网络模型的需求[8].

在处理实际问题的过程中,深度学习模型往往需要处理规模庞大,内容复杂多变的数据. 为了解决这样的问题,学术界也推出了大型的数据集,比如VOC(Visual Object Class)[9],COCO(Microsoft Common Objects in Context)[10],ImageNet[11]等,来辅助训练卷积神经网络模型. 通过前人的不懈努力,许多优秀的卷积神经网络模型被设计出来,如AlexNet,VGG(Visual Geometry Group)[12],GoogLeNet[13],ResNet(Residual Network)[14]等. 这些模型在ImageNet大规模视觉识别挑战赛[15](ImageNet Large Scale Visual Recognition Challenge)中都取得了非常优异的成绩,这些模型所代表的设计理念和设计原则也为后面的研究指明了方向. 然而,在完成比如实时人脸支付、实时ETC(Electronic Toll Collection)收费等实际任务时,都需要考虑到新的应用场景的复杂因素,这对模型提出了更高的要求. 为了可以解决实际问题,能够在终端满足检测识别任务的需求,其中非常重要的一条就是要求在不降低模型性能的基础上减少模型的计算复杂度,这深刻地反映了卷积神经网络的轻量化技术的必要性[16, 17].

卷积神经网络的轻量化技术是通过研究模型的参数和结构来降低模型开销的基础研究[18,19]. 特征复用的主要思想是对卷积计算过程中产生的中间特征进行多次直接或是间接复用,以获取更加丰富的特征信息. 为了缓解有限的硬件资源和复杂模型之间的矛盾,本文通过从特征复用的角度对卷积神经网络的轻量化技术展开了研究. 本文的主要工作包含以下三个方面:

(1) 利用特征复用思想设计了特征膨胀卷积模块. 该模块可以用很小的代价生成多角度的特征,为模型提供更丰富的图像特征信息.

(2) 使用特征膨胀卷积模块代替了标准卷积计算结构,设计出了高效的轻量化卷积神经网络模型,使之在实际计算的过程中使用更少的计算资源获得较好的性能.

(3) 在CIFAR和ImageNet数据集上对特征膨胀卷积模块的实际效果进行了详细的讨论,验证了本文提出的模块的有效性.

2 相关工作

卷积神经网络模型是一个前馈神经网络模型,卷积层是其中的重要组成部分,承担着提取图像特征的重要作用[20]. 因此针对卷积神经网络的卷积计算结构进行轻量化优化是一种有效的减少模型计算量的方法.

标准卷积计算通过将卷积核与输入图像按通道进行卷积操作,得到一组计算结果. 然后对计算结果按照输出通道进行按元素相加,这样就得到输入图像的一组输出特征图. 将标准卷积的滤波器尺寸设置为1×1时,就构成了点卷积结构. 由于参数很少,点卷积在进行卷积计算的时候可以节省大量的计算量,以很小的代价改变输入特征的通道数[21]. 在卷积计算过程中将特征分组,然后对每个分组分别进行卷积计算,最后将计算结果进行拼接得到最终的输出特征图,这样就变成了分组卷积[22],分组卷积有利于计算的并行化[23]. 当输入特征数、输出特征数和分组数相等时,就变成了深度卷积. 深度卷积不能自由变换通道数,也不能进行特征融合. 但是可以通过极小的代价,扩大模型感受野. 通过对点卷积和深度卷积的组合,可以设计出在结构上代替标准卷积的特征提取结构,比如深度可分离卷积结构[22~26]等.

单纯由深度卷积和点卷积组成的特征提取模块,在结构上可以代替标准卷积模块. 但是,由于其中包含的参数过少,在实际应用的过程中,也会出现训练时间过长,模型精度不够的情况. 这是由于模型的参数减少导致的表达能力下降. 特征复用思想可以有效地增强模型的表达能力[27].

特征复用技术是通过高密度地复用网络中的特征以降低网络计算量和提高网络性能的方法[28~30]. 2015年,He[31]等在卷积神经网络模型中加入跳跃连接结构设计了残差模块,通过残差模块实现特征复用[32],首次在实践中验证了特征复用思想的优越性. 密集连接方式充分利用跳跃连接可以将浅层网络特征直接传递到深层网络的特点,在卷积模型的所有层中尽可能地增加了跳跃连接结构,设计出了Dense block模块[32]. Ghost[33]模块从特征图的角度出发,采用廉价操作生成Ghost特征代替模型中的冗余特征,在减少模型参数的同时也可以有效提高模型的特征表达能力.

本文沿用Ghost模块的设计思路,利用特征复用思想对模块中的特征进行复用,采用点卷积和深度卷积获取不同角度的特征. 在减少了卷积计算过程中的计算量的同时增强了模型的特征表达能力.

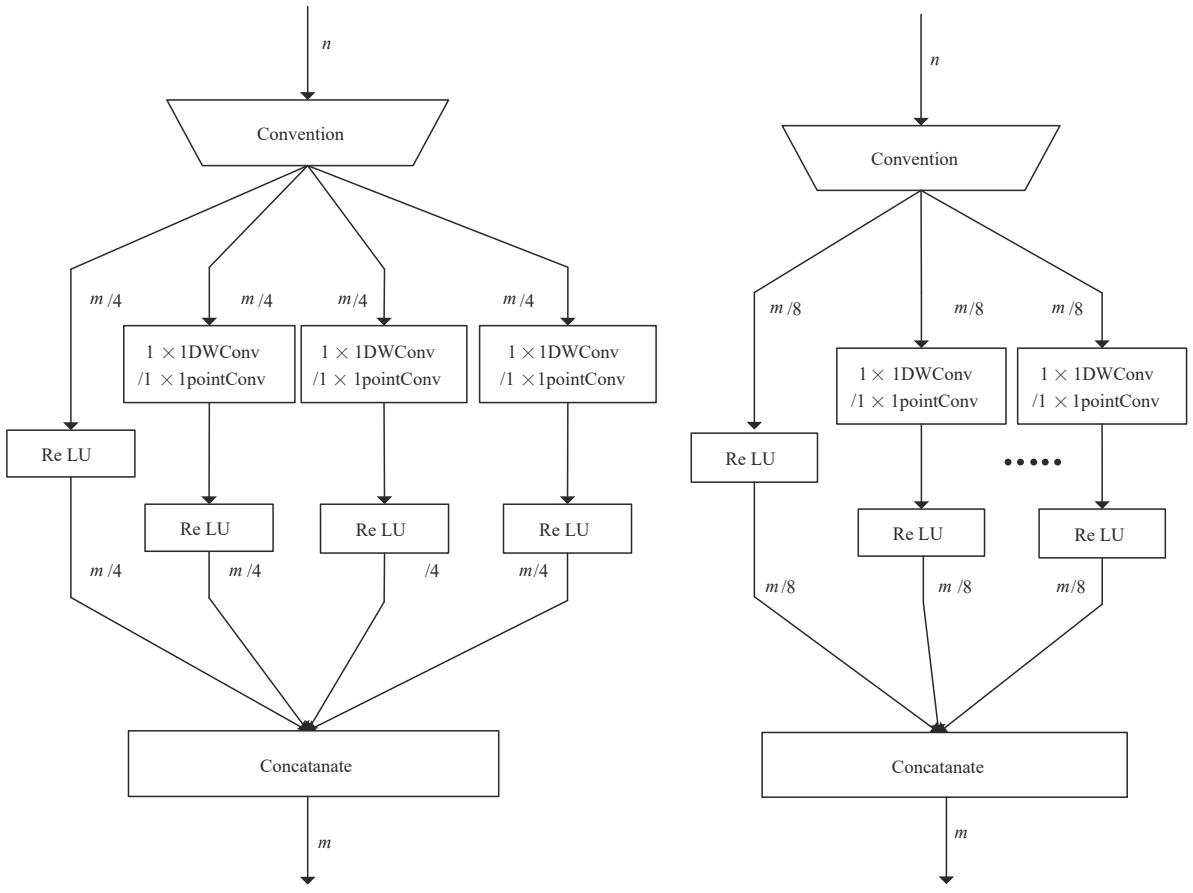
3 特征膨胀卷积模块设计

在卷积神经网络中,特征提取是整个模型的重要功能. 在卷积计算的过程中,特征在同一个输出通道中进行融合并生成新的特征. 但是模型中生成的特征具有一定的冗余性,因此完全可以利用特征复用的思

想,采用更加廉价的手段生成新的特征. 在Ghost模块中,Han等[33]人通过引入一系列线性变换的方式,对特征进行了直接复用. 但是,Ghost模块在进行特征复用的过程中,没有进行特征融合,只是根据计算出的特征进行简单变换,将一个特征变换为多个相似的特征来增大特征数量. 此外,分析文献[33]的数据可知,在CIFAR10数据集上进行测试时,当分支数s=2时获得的精度最高,并且会随着s的增大而减小. 而1−1/s代表的是Ghost模块所生成的Ghost特征比例,s的值越小,Ghost模块生成的特征的比例也就越小. 因此本文尝试探究,当利用Ghost方法生成的特征比例进一步减小的时候,模型的性能或许会有更进一步的提高. 所以,在延续了Ghost模块思想的基础上,引入廉价的多分支结构,设计出了基于特征复用思想的特征膨胀卷积模块(Feature Expansion convolution module, FExpand).

3.1　特征膨胀卷积模块的结构设计

特征膨胀卷积模块如图1所示,模块的输入为n个通道,输出为m个通道. 首先将根据分支数s将标准卷积的输出通道变窄为原来的1/4或1/8,生成原始特征图. 然后基于原始特征图在多个分支上对特征通道数进行恢复. 这些分支分为三类,第一类将上层网络的特征直接传递到下一层中,称为复制(Copy)分支. 第二类是采用类似Ghost模块的处理方法,通过引入简单的线性变换,利用极小的代价模拟出网络中的冗余特征图,称为影子(Ghost)分支. 第三类分支采用廉价的特征融合结构对上层输出的特征图进行融合,生成新的特征图,称为融合(Fuse)分支. 最后对三类分支结构的输出进行合并连接,作为模型的整体输出.



(*a*) 4分支结构　　(*b*) 8分支结构

图1　特征膨胀卷积模块

Copy分支的操作比较简单,直接传递即可. 因此对后两类分支进行详细描述. Ghost分支与文献[33]中的设计方法相同,本文中使用的线性变换单元大小都为1×1. 使用1×1和3×3的线性变换单元进行测试时,发现两者对模型的精度影响差异非常有限. 而使用1×1大小的线性变换单元可以更多地减少模型中的参数. Fuse分支的主要目的在于通过特征融合生成新的特征图. 因此设计特征融合结构时,没有采

用普通卷积,而是采用点卷积结构. 普通卷积采用大小为$k_w \times k_h$(k_w,k_h分别为卷积核的宽度和高度)的滤波器进行特征提取的过程中,模型在原始特征图上的感受野会从$k_w \times k_h$增大到$(2\times k_w-1)\times(2\times k_h-1)$,如图2所示. 这会增加特征图中的深度特征信息. 点卷积计算能够使用最少的计算量对特征通道进行线性组合生成新的特征,同时不会扩大当前层的感受野,有利于维持当前层在感受野上的一致性. Ghost分支和Fuse分支在计算过程中都保持特征图的通道数不变. 由于Copy分支输出的特征信息是经过激活函数处理之前的特征信息,因此在最后进行特征合并之前,需要对各分支提取的特征分别进行激活,以保证信息的一致性.

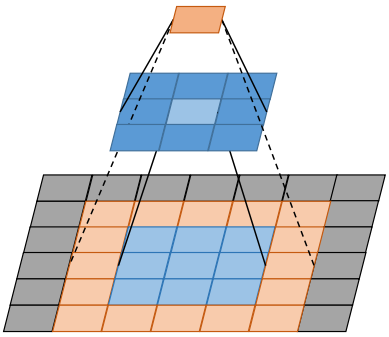


图2 卷积计算感受野变化图

关于分支数量,虽然在理论上可以设计得更多,以便于在两类特征图的组合上获得更加灵活的方案. 但是在实际操作中,分支数进一步扩大的时候,原始卷积获取的原始特征图数量会进一步减少,同时卷积核中的参数也会进一步减少,从而影响网络提取特征的性能. 特别是浅层网络在进行特征提取的时候,这种情况会非常严重. 而保留一定的参数空间来进行特征提取是必要的. 综合考虑,本文认为4分支和8分支是两种比较合适的方案.

特征膨胀卷积利用三种类型的分支提取了三种不同类型的特征,同时减少了网络参数的数量. 多分支的结构能够使用廉价操作获得模型中冗余的特征,同时能够更加灵活地改变冗余特征在模型中的占比. 在提高了整体网络的特征提取能力的同时,也提供了更灵活的探索网络模型特征结构的方法.

3.2 特征膨胀卷积模块的参数量分析

本文为特征膨胀卷积模块设计了两种结构,即4分支结构和8分支结构. 两种结构中第一分支均为Copy分支,剩下的由Ghost分支和Fuse分支组成. 对于一个大小为$n\times m\times k_w\times k_h$的卷积核,当采用4分支结构时,其参数量如式(1)~(4)所示.

$$Q=Q_c+Q_g+Q_f \tag{1}$$

$$Q_c=((n\times m\times k_w\times k_h))/4 \tag{2}$$

$$Q_g=N_g\left(1\times 1\times \frac{m}{4}\right)=N_g\times\frac{m}{4} \tag{3}$$

$$Q_f=N_f\left(1\times 1\times \frac{m}{4}\times\frac{m}{4}\right)=N_f\times\frac{m^2}{16} \tag{4}$$

在式(3)与式(4)中,N_g表示特征膨胀卷积结构Ghost分支数量,N_f表示Fuse分支数量,其中$N_g+N_f=3$. 与原始的卷积核中的参数相比,参数减少的比例如式(5)所示.

$$\begin{aligned} r&=\frac{n\times m\times k_w\times k_h}{\frac{(n\times m\times k_w\times k_h)}{4}+N_g\times\frac{m}{4}+N_f\times\frac{m^2}{16}}\\ &=\frac{16\times n\times k_w\times k_h}{4\times n\times k_w\times k_h+4\times N_g+N_f\times m}\\ &=\frac{4}{1+\frac{N_g}{4\times n\times k_w\times k_h}+\frac{N_f\times m}{4\times n\times k_w\times k_h}} \end{aligned} \tag{5}$$

在卷积核中$n\gg N_g$并且$n\gg N_f$,所以参数减少的比例主要取决于特征膨胀卷积模块的种类以及输入输出通道数. 当选择4分支结构时,最多可以将参数减少到原本的1/4. 当选择8分支结构的时候$N_g+N_f=8$,计算方法与式(5)相同,参数最多会减少到原本的1/8. 根据N_g和N_f的不同,生成的两种特征图在模型中的比例也不相同.

4 基于FExpand的卷积神经网络

在构建卷积神经网络模型的过程中,特征膨胀卷积模块可以从结构上直接简单代替标准的卷积模块,在减少计算量的同时提高网络提取特征的能力. 本节以特征膨胀卷积模块为核心,设计了基于VGG16模型结构的轻量化卷积神经网络模型,设计了基于残差结构的特征提取模块.

4.1 基于FExpand的VGG16模型设计

在VGG16网络模型中,通过对卷积层、归一化层以及激活层的反复堆叠构成的模块进行特征提取. 其中,卷积层提取了特征图的空间信息,并完成通道之间的特征融合;归一化层调整了特征图中数据的分布,缓解了梯度弥散和梯度爆炸的问题;激活层对数据进行非线性变换,增加了模型的表示能力. 反复堆叠会不断增加模型的复杂度,增加模型的表示能力,增加相对于原始图像的感受野. 特征膨胀卷积模块在功能上与标准卷积结构是相同的,可以完成对图像的特征信息的提取和融合,并且能够进行灵活的通道数变换. 因此,可以通过保持原始的VGG16模型结构不变,直接使用特征膨胀卷积模块替换标准卷积层,设计新的轻量化卷积神经网络模型. 基于特征膨胀卷积模块设计的轻量

化卷积神经网络模型如图3所示.

根据分支数目,特征膨胀卷积模块分为4分支结构的模块和8分支结构的模块. 在构建新的轻量化模型结构的时候,不同层采用不同的特征膨胀卷积模块,模型中的三种特征图的数量会呈现出不同的比例. 当直接使用4分支模块代替VGG16网络模型中的全部卷积层时,可以尽可能多地获取模型中生成的原始特征图,使得网络能够直接从原始图像数据中获取更多的图像特征信息. 而当直接使用8分支模块代替VGG16网络模型中的全部卷积层时,可以尽可能多地减少网络模型中的参数,减少模型在提取特征的过程中的计算量.

在设计模型时也可以将两种特征膨胀卷积模块混合,使用4分支模块代替VGG16网络模型中浅层的卷积层,然后使用8分支模块代替模型中深层的卷积层. 这样模型浅层可以更多地从原始图像数据中获取特征信息,同时网络深层可以进一步减少模型的计算量,对前面单独使用一种特征膨胀卷积模块的方案的优势进行了一定的综合. 但是不建议使用8分支模块代替VGG16网络模型处于浅层的卷积层,用4分支模块代替网络深层的卷积层. 这是因为VGG16的第一层输出通道数只有64,因此使用8分支模块代替浅层卷积层时,特征膨胀卷积模块内部生成的原始特征图就只有8个,从原始图像中提取出来的图像信息过少,不利于模型深层进一步提取图像特征.

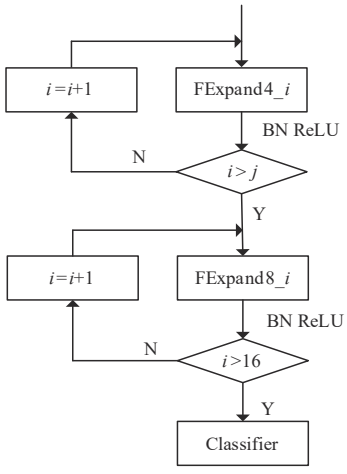


图3　基于特征膨胀卷积模块的VGG16设计

在特征膨胀卷积模块内部,由于Ghost分支和Fuse分支的数量不同,两种特征的数量比例也会呈现出差异,这同样会对模型的性能产生影响. 在4分支结构和8分支结构中,N_g和N_f的值,以及生成的两种特征图在模块中所占的比例如表1所示. 由于文献[33]对Ghost分支生成的特征图的比例高于50%的情况已经进行了比较充分的讨论,本文主要针对比例不大于50%的情况进行相关实验.

表1　Ghost特征与Fuse特征的组合比例

分支类型	N_g数	N_f数	N_g(%)	N_f(%)
4	1	2	25	50
4	2	1	50	25
8	1	6	12.5	75
8	2	5	25	62.5
8	3	4	37.5	50
8	4	3	50	37.5
8	5	2	62.5	25
8	6	1	75	12.5

本文针对上述两种情况分别设计了实验,讨论了不同情况下,基于特征膨胀卷积模块的轻量化卷积神经网络模型的性能,分析了不同类型的特征对模型的性能的影响,性能最优的轻量化模型结构如表2所示.

表2　基于特征膨胀卷积模块的最优轻量化VGG16模型

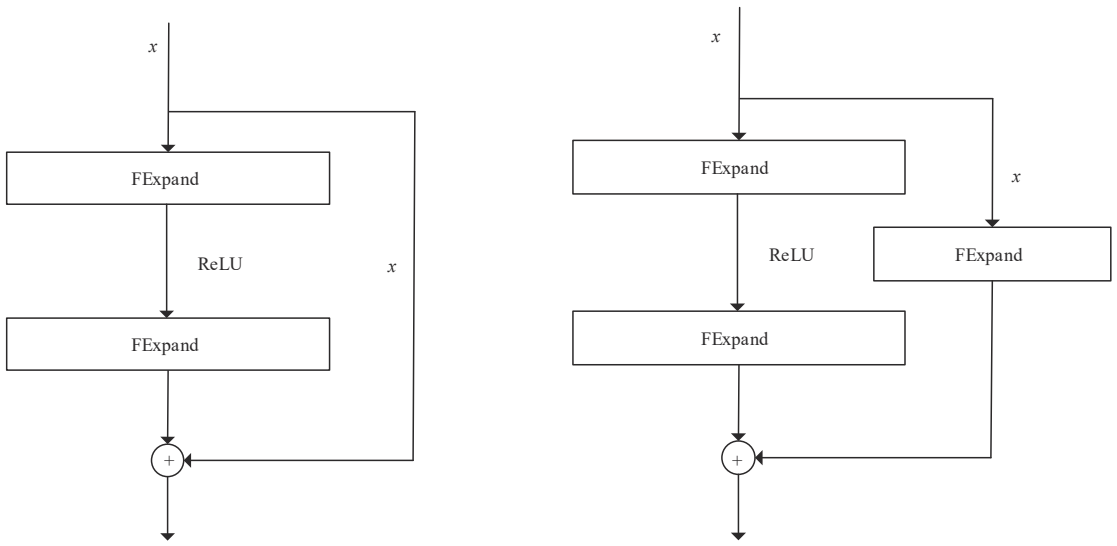
Layer	Module	m, $[C, N_g, N_f]$
Block1	FExpand4	16×4, [1,1,2]
Block2	FExpand4	16×4, [1,1,2]
Pooling	MaxPooling	
Block3	FExpand4	32×4, [1,1,2]
Block4	FExpand4	32×4, [1,1,2]
Pooling	MaxPooling	
Block5	FExpand4	64×4, [1,1,2]
Block6	FExpand4	64×4, [1,1,2]
Block7	FExpand4	64×4, [1,1,2]
Pooling	MaxPooling	
Block8	FExpand4	64×4, [1,1,2]
Block9	FExpand4	64×8, [1,1,6]
Block10	FExpand8	64×8, [1,1,6]
Pooling	MaxPooling	
Block11	FExpand8	64×8, [1,1,6]
Block12	FExpand8	64×8, [1,1,6]
Block13	FExpand8	64×8, [1,1,6]
Pooling	AvePooling	
Classfier	Fully Connected+Softmax	

4.2　基于FExpand的残差模块设计

残差结构是在标准卷积操作上加上一个跳跃连接构成的,本文所设计的模块,能够对不同层次的图像特征进行融合,因此本节尝试将特征膨胀卷积模块与跳跃连接相融合,设计基于特征膨胀卷积的残差模块,进一步提高特征复用的频率. 其结构如图4所示. 本文设计的特征膨胀卷积模块可以很好地替代标准卷积模

块,因此在残差模块中可以直接替换. 残差结构通过对应元素相加的方式对不同层的特征图进行融合,以保证特征通道数不会因为跳跃连接而出现加倍的情况. 特征图在通道数不变时可以直接相加,但是当通道数出现变化时就需要进行调整. 本文在通道数不变时,采用如图4(*a*)所示的结构,当出现特征通道数不匹配时,则加入一个点卷积结构对通道数进行调整,如图4(*b*)所示.

特征膨胀卷积模块从卷积操作的角度通过廉价操作,以极小的代价生成新的特征,本质上还是对同一层的特征图上从不同的角度进行的特征提取. 而跳跃连接,则从更大的尺度上,对不同层的特征进行融合,将浅层模型的特征直接传递到模型深层的同时,也让模型在自动优化参数时,能够将深层模型的变化直接反馈到浅层. 本节通过将两者结合,能够更加全面的验证特征复用思想的优越性.



(*a*) 特征通道数相同 (*b*) 特征通道数不匹配

图4 基于特征膨胀卷积模块的残差模块

5 实验结果与分析

本文基于特征复用的思想,在卷积操作中引入多分支结构,设计了能够提取更丰富的图像信息的特征膨胀卷积模块. 在VGG16的网络结构基础上,基于特征膨胀卷积模块设计了轻量化的网络模型. 与原始的网络相比,新的轻量化模型具有更少的参数数量.

5.1 基于FExpand的轻量化VGG16实验

本节基于VGG16模型的结构框架,使用4分支和8分支的特征膨胀卷积模块代替标准的卷积模块,构建了新的轻量化卷积模型. 首先单独使用4分支模块和8分支模块对VGG16模型进行优化,然后同时使用两种模块进行优化,最后综合比较各个方法设计出的轻量化模型的性能,选择最好的优化方案. 本节设计了不同的方案,并在CIFAR10数据集上对模型的性能进行了测试,其结果分别如表3和表4所示.

如表3中所示,在只使用4分支模块代替卷积模块构建轻量化模型时,网络中的参数数量减少了大约73%. 当只使用8分支模块代替卷积操作时,网络中的参数数量下降到了原本的大约1/8,计算量则下降到原本的1/12. 改变两种模块中的Ghost分支和Fuse分支数的比例时,对模型的参数数量和计算量没有影响,但是会对模型的性能产生少量的影响. 当模块中的Fuse分支增多的时候,模型在CIFAR10上的性能会有少量的

表3 基于特征膨胀卷积模块的轻量化VGG16模型在CIFAR10数据集上的实验结果

	(C,N_g,N_f)	N_f比例	Weights(M)	Reduce(%)	Acc(%)
FExpand4	1,1,2	50%	3.91	74.18	92.7
FExpand4	1,2,1	25%	3.80	73.44	92.7
FExpand8	1,1,6	75%	2.02	86.28	92.5
FExpand8	1,2,5	62.5%	1.99	86.48	92.5
FExpand8	1,3,4	50%	1.96	86.68	92.4
FExpand8	1,4,3	37.5	1.94	86.82	91.9
VGG16	-	-	14.72	-	92.3

提升,相比于原始的VGG16模型,最多可以增加0.4%.这表明模型的轻量化主要依赖分支的数量,而增加Fuse特征有利于提高模型的性能.

观察表3可知,当使用8分支结构,且Ghost分支数为1,Fuse分支数为6时获得的效果最好,而使用4分支结构时,Ghost分支数为1,Fuse分支数为2时模型的性能最好.因此,在同时使用两种特征膨胀卷积模块改进VGG16模型时,4分支模块中各分支数为C=1,N_g=1,N_f=2,8分支模块中,各分支数为C=1,N_g=1,N_f=6.在浅层使用4分支模块,深层使用8分支模块.

在同时使用两种特征膨胀卷积模块对VGG16模型进行优化时,首先使用8分支模块替换全部标准卷积模块,然后从浅层逐步将8分支模块替换为4分支模块,直至全部替换完成.全部的结果如表4所示,第一列表示替换为4分支模块的前j层的层数.分析可知,随着逐步替换的推进,模型的参数量会不断增加,同时模型的性能随着参数增加也会有少量提升.当使用4分支模块替换前9层时,模型的性能到达最优.这表明同时使用两种特征膨胀卷积模块,使用4分支结构替换模型浅层,使用8分支结构替换模型深层的方案是合理的.也说明模型浅层应该适当保留一定的参数,以便于能够直接从原始图像中获取足够多的特征信息.

表4　基于两种特征膨胀卷积模块的改进VGG16模型在CIFAR10数据集上的实验结果

Layer	Weights(M)	Acc(%)
j=1	2.023	92.5
j=2	2.023	92.5
j=3	2.033	92.5
j=4	2.053	92.5
j=5	2.093	92.5
j=6	2.173	92.6
j=7	2.243	92.6
j=8	2.403	92.6
j=9	2.703	92.7
j=10	3.003	92.7
j=11	3.313	92.7
j=12	3.613	92.7
j=13	3.913	92.7

5.2　基于FExpand的轻量化模型验证实验

本文基于特征复用思想,设计了基于特征膨胀卷积模块的轻量化VGG16模型,并且在CIFAR10数据集上进行了测试,检测出了性能最好的结构.本节则尝试在更多数据集上对本文所设计的两种轻量化模型进行实验验证,并且与其他的方法进行比较.

在训练的过程中采用随机梯度下降方法对模型的参数进行优化,训练批次为160次.初始学习率为0.01,在训练过程中会在训练批次超过50次之后,每10个批次学习率会按照指数下降.采用更细粒度的精度测试方法,在每一个批次训练的过程中,会实时测试当前的损失值,并在损失值最小的时候测试当前参数的性能.采用更细粒度的测试方法可以在一定程度上提高模型最终的检测效果.模型在CIFAR数据集(包含CIFAR10和CIFAR100)和ImageNet子集ILSVRC-2012上测试的结果与其它模型的比较分别如表5和表6所示(表5中GMAC表示10亿次乘加运算).

由表5可知,在VGG16模型结构下,基于特征膨胀卷积模块设计的卷积神经网络在参数数量和计算量上与原始网络相比,参数数量减少了81.64%,模型在计算过程中的计算量则减少了4.019%.与使用Ghost模块优化的VGG16相比,本方法在减少参数量上有优势,但是在计算量上要稍微多一点.在ResNet50模型结构下,与原始的VGG16模型和Ghost优化之后的模型相比,本文设计的模型在参数数量和计算量上都是具有优势的.从性能上看,本文所设计的模型在CIFAR数据集上的分类精度与相同结构的其它模型相比有少量的提升,与其他卷积神经网络模型相比也具有较好的结果.综合上述分析,说明本文所设计的模型具有一定的价值.

由表6中在10-crop ImageNet数据集上进行测试的结果可知,本文设计的基于特征膨胀卷积模块的轻量

表5　轻量化VGG16模型在CIFAR10和CIFAR100数据集上的测试结果

Models	Weights(M)	GMAC	CIFAR10 (%)	CIFAR100 (%)
VGG-16[12]	14.72	0.316	92.3	72.3
FitNet[36]	2.5	0.382	91.61	64.96
Highway Network[29]	2.3	0.372	92.24	67.76
DenseNet121[32]	15.3	0.143	94.81	80.36
ResNet-50[14]	25.557	0.086	93.03	77.78
SE-Net[34]	47.7	0.259	94.79	80.02
CondenseNet[35]	0.52	0.006	94.48	76.36
Ghost-VGG-16[33]	7.387	0.156	93.7	68.80
Ghost-ResNet50[33]	12.36	0.02	92.7	72.6
VGG16-our	2.703	0.189	92.7	73.4
ResNet50-our	5.48	0.01	93.3	78.1

表6 轻量化VGG16模型在ImageNet数据集上的测试结果 单位:%

Models	Top-1 error	Top-5 error
VGG-16 [12]	71.93	90.67
DenseNet121[32]	76.39	93.34
ResNet-50 [14]	77.15	93.29
Ghost-ResNet [33]	75	92.3
VGG16-our	75.68	92.6
ResNet50-our	77.16	93.34

化卷积神经网络,与原始的结构和经过Ghost模块改进的结构相比都具有一定的优势. 在Top-1错误率上,本文设计的模型比原始的VGG16模型高3.75%,与原始的ResNet50模型相比有微弱的优势. 在Top-5错误率上,本文设计的模型比原始的VGG16高大约2%,与原始的DenseNet101模型的性能相同. 表6中的数据表明,本文所设计的特征膨胀卷积模块可以补足Ghost模块在特征提取能力上的不足,在减少模型的参数和计算量的同时,少量提高模型的分类性能. 这表明在模型中提取充足的复杂的图像特征是可以提高模型的分类性能的.

上述实验结果表明,本文所提出的特征膨胀卷积模块是具有更强的特征提取能力,能够通过多分支结构,提取出更丰富的图像特征,增强卷积神经网络模型的性能.

6 结论

本文从卷积神经网络的结构出发,遵循特征复用的思想,设计了特征膨胀卷积模块. 特征膨胀卷积模块利用缩减输出通道的标准卷积模块获取原始特征,通过Ghost模块中的廉价操作获取Ghost特征,通过Fuse廉价操作获取Fuse特征,并将三种特征结合作为模块的输出. 之后,对本文提出的特征膨胀卷积模块进行了实验验证,对比实验表明了其有效性. 尽管如此,本文的工作依旧有一定的拓展空间:(1)本模块在进行特征提取的过程中没有加入深度信息,在后续设计的过程中,可以考虑在模块中加入深度信息,进一步提取更复杂的图像特征;(2)对网络模型的层数与性能之间的关系还需进一步的深入研究.

参考文献

[1] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251. (in Chinese)

[2] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1): 1-17.
LU H T, ZHANG Q C. Applications of deep convolutional neural network in computer vision[J]. Journal of Data Acquisition and Processing, 2016, 31(1): 1-17. (in Chinese)

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[4] 权宇, 李志欣, 张灿龙, 等. 融合深度扩张网络和轻量化网络的目标检测模型[J]. 电子学报, 2020, 48(2): 390-397.
QUAN Y, LI Z X, ZHANG C L, et al. Fusing deep dilated convolutions network and light-weight network for object detection[J]. Acta Electronica Sinica, 2020, 48(2): 390-397. (in Chinese)

[5] 赖小波, 许茂盛, 徐小媚. 多分类CNN的胶质母细胞瘤多模态MR图像分割[J]. 电子学报, 2019, 47(8): 1738-1747.
LAI X B, XU M S, XU X M. Glioblastoma multiforme multi-modal MR images segmentation using multi-class CNN[J]. Acta Electronica Sinica, 2019, 47(8): 1738-1747. (in Chinese)

[6] 王琦, 谭娟. 基于人工智能技术的光学超精密检测技术[J]. 激光杂志, 2021, 42(2): 156-160.
WANG Q, TAN J. Optical ultra-precision detection technology based on artificial intelligence technology[J]. Laser Journal, 2021, 42(2): 156-160. (in Chinese)

[7] 赵耀霞, 吴桐, 韩焱. 基于卷积神经网络的复杂构件内部零件装配正确性识别[J]. 电子学报, 2018, 46(8): 1983-1988.
ZHAO Y X, WU T, HAN Y. Identifying the correctness of fit of internal components based on a convolutional neural network[J]. Acta Electronica Sinica, 2018, 46(8): 1983-1988. (in Chinese)

[8] 郭棉, 张锦友. 移动边缘计算环境中面向机器学习的计算迁移策略[J]. 计算机应用, 2021, 41(9): 2639-2645.
GUO M, ZHANG J Y. Computation offloading policy for machine learning in mobile edge computing environments [J]. Journal of Computer Applications, 2021, 41(9): 2639-2645. (in Chinese)

[9] VICENTE S, CARREIRA J, AGAPITO L, et al. Reconstructing pascal voc[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 41-48.

[10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft CO-

CO: Common objects in context[C]//European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[11] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 1026-1034.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-05-05]. https://arxiv.org/abs/1409.1556.

[13] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1-9.

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[15] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[16] CHENG Y, WANG D, ZHOU P, et al. Model compression and acceleration for deep neural networks: The principles, progress, and challenges[J]. IEEE Signal Processing Magazine, 2018, 35(1): 126-136.

[17] KWON D, KIM H, KIM J, et al. A survey of deep learning-based network anomaly detection[J]. Cluster Computing, 2019, 22(1): 949-961.

[18] SINDHWANI V, SAINATH T, KUMAR S. Structured transforms for small-footprint deep learning[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Cambridge: MIT Press, 2015: 3088-3096.

[19] CHEN W L, WILSON J T, TYREE S, et al. Compressing neural networks with the hashing trick[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. Cambridge: JMLR, 2015: 2285-2294.

[20] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//CVPR' 14: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1725-1732.

[21] LI Y Y, BU R, SUN M C, et al. PointCNN: Convolution on X-transformed points[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2018: 828-838.

[22] IOANNOU Y, ROBERTSON D, CIPOLLA R, et al. Deep roots: Improving CNN efficiency with hierarchical filter groups[EB/OL]. (2016-05-20) [2021-05-05]. https://arxiv.org/abs/1605.06489.

[23] 朱虎明, 李佩, 焦李成, 等. 深度神经网络并行化研究综述[J]. 计算机学报, 2018, 41(8): 1861-1881.
ZHU H M, LI P, JIAO L C, et al. Review of parallel deep neural network[J]. Chinese Journal of Computers, 2018, 41(8): 1861-1881. (in Chinese)

[24] HOWARD ANDREW G, ZHU MENGLONG, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-27) [2021-05-05]. http://arxiv.org/abs/1704.04861.

[25] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.

[26] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1314-1324.

[27] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[EB/OL]. (2016-05-23)[2021-05-05]. https://arxiv.org/abs/1605. 07146.

[28] HUANG G, SUN Y, LIU Z, et al. Deep networks with stochastic depth[C]//European Conference on Computer Vision. Cham: Springer, 2016: 646-661.

[29] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Highway networks[EB/OL]. (2015-05-03) [2021-05-05]. https://arxiv.org/abs/1505.00387.

[30] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 936-944.

[31] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2755-2763.

[32] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2261-2269.

[33] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]//2020 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1577-1586.

[34] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE, 2019: 2011-2023.

[35] HUANG G, LIU S C, MAATEN L V D, et al. CondenseNet: An efficient DenseNet using learned group convolutions[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2752-2761.

[36] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: Hints for thin deep nets[EB/OL]. (2014-12-19)[2021-05-05]. https://arxiv.org/abs/1412.6550.

作者简介

许新征 男，1980年8月生，安徽宿州人，博士，教授．主要从事机器学习与数据挖掘、人工智能与模式识别、医学图像处理等方面的研究．

E-mail: xxzheng@cumt.edu.cn

李 杉 男，1995年8月生，湖北松滋人，硕士研究生，主要从事深度学习和计算机视觉等方面的研究．