http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0287

# 基于结构重参数化的深度可分离卷积神经网络

陈红[1]，闫建国[2]，杨华[1]，张靖[1]，李伟[3]，杨靖[1,4,*]

(1. 贵州大学 电气工程学院，贵阳 550025；2. 中国电建集团贵州工程有限公司，贵阳 550025；3. 贵州省普通高等学校工程研究中心，贵阳 550025；4. 贵州大学 贵州省“互联网+”协同智能制造重点实验室，贵阳 550025)

**摘 要**：针对目前卷积神经网络(CNN)模型大部分采用的单分支深度卷积方法影响模型的表达能力、占用大量参数和Flops的问题，提出一种新的轻量级CNN模型——基于结构重参数化的深度可分离卷积神经网络(DSCNN)。该模型的特征提取模块结构重参数化混合深度卷积(RepMiX)能融合不同通道和空间位置之间的信息，实现多尺度特征融合；基于结构重参数化的DSCNN中的重参数化非对称空间算子(RepASO)通过具有不同功能的多分支结构来学习不同信道特征信息，提高模型特征学习能力；RepMiX和RepASO均结合结构重参数化技术和深度可分离卷积(DS-Conv)思想，实现了训练和推理阶段的结构解耦，在减少模型参数和Flops的同时加速模型推理；在Tiny-ImageNet-200、CIFAR-10、CIFAR-100数据集及自建铝锭表面缺陷分类数据集上进行了对比实验，结果表明：基于结构重参数化的DSCNN实现了更高的浮点运算速度，并保持了具有竞争力的精度。

**关 键 词**：卷积神经网络；结构重参数化；深度可分离卷积；非对称卷积；结构解耦
**中图分类号**：V557+.1; TP751
**文献标志码**：A **文章编号**：1001-5965(2026)06-2145-11

卷积神经网络(convolutional neural network, CNN)在表示高维复杂函数方面表现出强大的能力，已成功应用于各种视觉场景[1]。与视觉变换器(vision Transformer, ViT)相比，CNN计算复杂度更低，效率更高，因此，在计算机视觉应用中占据主导地位。目前，为了达到更高的精度，主流的CNN体系结构复杂而庞大，对设备的计算和存储能力有很高的要求。为了在移动设备上实现实际应用，需要在保持高精度的同时，建立参数量少、Flops更少的轻量级CNN模型。

深度可分离卷积(depthwise separable convolution, DS-Conv)是在MobileNetV1[2]中提出的，与普通卷积不同，其由2个连续的部分组成：① 深度卷积(depthwise convolution, DW-Conv)，用于提取空间特征；② 点卷积(pointwise convolution, PW-Conv)，用于提取信道特征。DS-Conv减少了计算负担和模型参数，成为后续轻量级CNN[3]及现代大规模CNN[4-5]的基本设计组件。在MobileNetV1的基础上，MobileNetV2[6]引入了倒残差模块(inverted residual block, IRB)，以减轻特征损失并增强表示能力。利用IRB设计范式作为基本模块，学者提出了具有不同整体架构的各种轻量级模型。除了IRB，研究人员还设计了具有不同结构的高效基本块[7]，以增强轻量级CNN的表达能力。然而，架构的不

**收稿日期**：2024-05-07；**录用日期**：2024-06-25；**网络出版时间**：2024-07-03 15:00
**网络出版地址**：link.cnki.net/urlid/11.2625.V.20240703.0858.001
**基金项目**：国家自然科学基金(61640014,61963009)；贵州省教育厅创新群体(黔教合KY字[2021]012)；贵州省科技支撑计划(黔科合支撑[2022]一般017,黔科合支撑[2023]一般411,黔科合支撑[2024]一般051)；贵州省教育厅工程研究中心(黔教技[2022]040,黔教技[2022]043)；中国电建集团科技项目(DJ-ZDXM-2020-19,DJ-ZDXM-2022-44)；贵州省双碳研究院开放课题(DCRE-2023-13)；中铝国际工程股份有限公司科技重大专项(CJ2022JS-ZD02)
* **通信作者**. E-mail：jyang7@gzu.edu.cn

*引用格式：陈红，闫建国，杨华，等. 基于结构重参数化的深度可分离卷积神经网络[J]. 北京航空航天大学学报，2026，52(6)：2145-2155.*
*CHEN H，YAN J G，YANG H，et al. Deep separable convolutional neural networks based on structural reparameterization[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(6)：2145-2155 (in Chinese).*

一致性导致结构冗余，使用不同的基本模块和架构导致不公平的容量比较，模糊了架构选择与性能提升之间的因果关系。

此外，虽然一些工作[8]将DW-Conv分解为并行的低秩分支以节省计算成本，但轻量级模型追逐更低Flops和参数的主要主干在于PW-Conv。ShuffleNet[9]进行了信道分组和洗牌，但不同分组的信道之间信息交换不足。IGCV3[10]提出了交错组卷积，其将常规卷积分解为多层组卷积，结构更为复杂。ChannelNets[11]将稀疏连接引入PW-Conv，对于每个输入通道，输出通道只关注其中的一小部分。

为解决上述问题，本文提出了一种参数和Flops更少的轻量级CNN模型，该模型在不影响精度的前提下尽可能轻量化，实现了精度和速度的平衡。本文重点讨论构建轻量级CNN模型的4个关键部分：整体架构、特征提取模块、空间算子和基本模块。主要贡献如下：

1）设计了特征提取模块结构重参数化混合深度卷积（reparameterization MiXer, RepMiX），一种由RepMixer和混合模块组成的类似Transformer的架构。RepMixer结合了多个并行DW-Conv，并结合重参数化方法来加快推理，混合模块混合信道和位置信息。

2）设计了重参数化非对称空间算子（reparameterized asymmetric spatial operator, RepASO）作为基于结构重参数化的深度可分离卷积神经网络（deep separable convolutional neural network, DSCNN）模型的空间算子。模型训练时，RepASO为并行多分支架构，以弥补可学习参数减少带来的影响，同时增强模型的表征能力。模型推理时，RepASO利用重参数化将多分支架构转化为单一分支，提高模型的推理速度。

3）基于RepMiX和RepASO模块，提出了名为基于结构重参数化的DSCNN的轻量级CNN模型。实验表明，相比现有轻量级CNN模型，基于结构重参数化的DSCNN具有更快的浮点运算速度，并保持了相当甚至更高的精度。

# 1 相关理论

## 1.1 结构重参数化

结构重参数化[2,12]是重参数化的一种代表性方法，使用某一结构转移过来的参数对新结构进行参数化。ACNet[13]在训练中使用常规卷积分支，并构建垂直和水平2个卷积分支，推理时，这些分支会转化为单分支的常规卷积。RepVGG[12]在训练期间采用多分支拓扑结构，通过结构重参数化实现训练和推理结构的解耦。MobileOne[14]利用结构重参数化，在训练期间为DW-Conv和PW-Conv多个并行分支，在推理过程中没有分支，而是采用直通式架构。结构重参数化的具体方法与RepVGG是一致的，即把批归一化（batch normalization, BN）融合到前面的卷积中，合并并行卷积参数。

总之，尽管轻量级CNN模型在减少模型参数和计算复杂度方面取得了一定的成功，但由于设计目标是轻量级，不得不在模型准确性和学习能力等方面做出妥协，面临如何在速度和准确性之间取得平衡的挑战。本文提出的基于结构重参数化的DSCNN模型，是将结构重参数化技术、DS-Conv与非对称卷积方法相结合的新方法，该方法为解决上述问题提供了新的解决方案。

## 1.2 深度卷积

DW-Conv是常规卷积的一种高性能变体，已被广泛采用为许多神经网络的关键构建模块。对于输入$F \in R^{m\times n\times w}$，DW-Conv应用$m$个滤波器来计算输出。每个滤波器在输入通道上的DW-Conv可以写成如下形式：

$$\hat{G}_{k,l,m}=\sum_{i,j}\hat{k}_{i,j,m}*F_{k+i-1,l+j-1,m} \tag{1}$$

式中：$\hat{k}$为大小为$D_k\times D_k\times M$的深度卷积核，核$\hat{k}$中的第$m$个滤波器应用于$F$的第$m$个通道，在对第$m$个通道进行滤波后生成输出特征图$\hat{G}$。

与普通卷积相比，DW-Conv将Flops从$m\times n\times k^2\times c^2$减少到$m\times n\times k^2\times c$。虽然DW-Conv（随后通常是PW-Conv）能有效减少Flops，但不能简单地取代常规卷积，因为这会导致精度显著下降。因此，在实际应用中，通常会增加DW-Conv的通道数或采用多分支架构来弥补精度的下降。

# 2 本文算法

## 2.1 整体架构设计

许多轻量级CNN模型具有相似的基本模块，但由于其整体架构存在差异，可能会出现不必要的冗余，这可以通过统一整合来解决。在包含不同基本模块的轻量级CNN模型中，架构设计缺乏一致性和统一标准，导致模型的整体结构各不相同，模糊了架构选择与性能提升之间的因果关系。因此，建立合理的整体架构尤为重要。参考CNN[5]和ViTs[15]的架构，本文提出了一种新的架构，称为基于结构重参数的DSCNN，图1展示了模型的整体架构。

Shalk：其为旨在使用具有多分支架构的过渡层，而不是相对较大的过渡层，在开始阶段捕捉更多细节。首先，创建并行特征提取网络，该网络

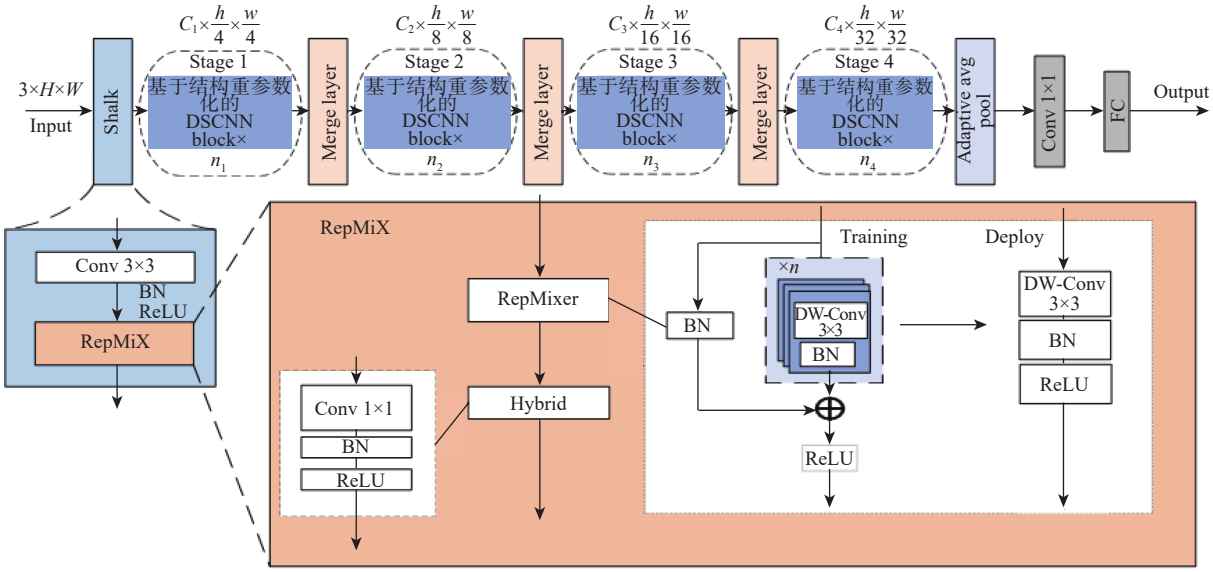



图 1　基于结构重参数化的 DSCNN 的整体架构

Fig. 1　Overall architecture of DSCNN based on structural reparameterization

由多个3×3 DW-Conv 和 BN 层组成，还采用了直连分支 (identity 分支) 来复用输入特征。其次，使用1×1卷积进行通道混合。

Stages: 基于结构重参数化的 DSCNN 共包括 4 个 Stages，每个阶段由重复的基于结构重参数化的 DSCNN 块 (包含 RepASO 算子的块) 堆叠组成，类似于 IRB。第 1～4 阶段的计算比例为 1∶2∶8∶2。本文每个阶段的基于结构重参数化的 DSCNN 块数指定为 [1, 2, 8, 2]。

Downsampling layer: 在第 1～3 阶段后，会增加一个单独的下采样层，该层使用步长为 2 的卷积核对前一阶段的输出特性进行卷积运算。该下采样层将输入分辨率降低一半，通道数增加一倍，之后是一个 BN 层以稳定训练。

Head: 其为模型的最后部分，位于第 4 阶段后。首先，使用自适应平均池化技术获得特征向量，然后，应用一组1×1卷积层进一步混合特征信息。最后，将输出的特征信息输入最后一个全连接分类器，产生最终输出。

### 2.2　特征提取模块 RepMiX

近年来，Transformer 架构在各种任务中取得了巨大成功。本设计从 ConvMixer[16] 和 RepViT[17] 中汲取灵感，设计了 RepMiX，其由空间混合算子 RepMixer 和通道混合算子 Hybrid 模块组成。RepMiX 能融合不同通道和空间位置间的特征信息，旨在从图像或特征数据中提取深层信息，实现多尺度特征融合[18]。

在模型训练模式下，使用具有多分支结构的转换层来捕捉更多细节，并在不同尺度上获得相应的输出结果。每个分支都会独立学习混合空间信息，并在一定范围内建立跨维交互。每个分支都包含一个3×3 DW-Conv 层和 BN 层。BN 函数表达式为

$$y_i=\gamma\frac{x_i-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{2}$$

式中：$\mu$为均值；$\sigma^2$为方差；$\gamma$为缩放因子；$\beta$为偏移；$\varepsilon$为一个小数，防止分母为零。

这种设计可以确保每一层激活函数的输入在所有维度上都有相似的分布，有助于网络更全面地捕捉输入数据的结构和特征。并且，为了适应不同批大小和训练阶段的需要，在本文模型中，采用指数移动平均(exponential moving average, EMA)动态调整 BN 训练过程中均值和方差。在 BN 中，EMA 通过结合当前批次的统计量和之前累积的统计量来计算，用于更新累积的均值和方差，公式如下：

$$\begin{cases}\mu_{\text{running}}=\xi\mu_{\text{batch}}+(1-\xi)\mu_{\text{running}}\\ \sigma^2_{\text{running}}=\xi\sigma^2_{\text{batch}}+(1-\xi)\sigma^2_{\text{running}}\end{cases} \tag{3}$$

式中：$\mu_{\text{batch}}$和$\sigma^2_{\text{batch}}$为当前批次数据的均值和方差；$\mu_{\text{running}}$和$\sigma^2_{\text{running}}$为累积的均值和方差；$\xi$为平滑常数。

此外，RepMixer 运算器还包括可复用输入特征的直连分支。直连分支引入了跨层连接，允许前一个特征直接传递到后续层。这样，网络就能更好地捕捉低层与高层特征之间的信息流。最后，所有3×3 DW-Conv 分支的输出与直连分支的输出相加，并向前传播。

在部署模式下，为加快推理速度，合并卷积层和 BN 层。卷积计算公式如下：

$$\boldsymbol{y}=\boldsymbol{\omega}*\boldsymbol{X}+\boldsymbol{b} \tag{4}$$

式中：$\boldsymbol{\omega}$为卷积核的权重；$\boldsymbol{b}$ 为卷积层的偏置，通常在建立模型时设置为 0；$\boldsymbol{X}$为输入特征矩阵；$\boldsymbol{y}$ 为卷积层的输出，即 BN 层的输入。

将式(4)代入式(2),可得

$$\boldsymbol{y}_i=\boldsymbol{\gamma}\frac{\boldsymbol{\omega}*\boldsymbol{X}+\boldsymbol{b}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}}+\boldsymbol{\beta}=\left(\boldsymbol{\gamma}\frac{\boldsymbol{\omega}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}}\right)*\boldsymbol{X}+\left(\boldsymbol{\gamma}\frac{\boldsymbol{b}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}}+\boldsymbol{\beta}\right) \tag{5}$$

融合后,原来的卷积层和BN层可以用一个新的卷积操作来表示,其中,$\boldsymbol{\omega}_{\text{new}}=\boldsymbol{\gamma}\frac{\boldsymbol{\omega}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}}$,$\boldsymbol{b}_{\text{new}}=\boldsymbol{\gamma}\frac{\boldsymbol{b}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}}+\boldsymbol{\beta}$。在训练阶段,BN层仍然需要计算和存储每个批次的均值、方差、尺度和偏移,训练时的内存需求和计算资源并没有减少;但在推理时,由于BN层被融合到之前的卷积中,减少了对均值、方差、尺度和偏移的计算和存储,不需要存储BN层的中间统计量,从而减少了推理时的计算负担和内存使用量。

利用结构重参数化技术,对图1中的RepMixer结构进行剪枝。在推理过程的多分支训练过程中获得的模型参数将用于单分支结构的参数化。该操作可以将多个分支融合为一个3×3卷积,直接处理输入特征图,加速模型推理。将多分支结构转换为单分支结构的数学表达式为

$$\text{bn}(\boldsymbol{M},\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma},\boldsymbol{\beta})_{:,i,:,:}=(W_{:,i,:,:}-\mu_i)\frac{\gamma_i}{\sigma_i}+\beta_i \tag{6}$$

$$\begin{cases}W'_{:,i,:,:}=\dfrac{\gamma_i}{\sigma_i}W_{:,i,:,:}\\ b'_i=\beta_i-\dfrac{\mu_i\gamma_i}{\sigma_i}\end{cases} \tag{7}$$

$$\begin{cases}W'_{\text{new}}=\displaystyle\sum_{j=1}^{n}W'_j\\ b'_{\text{new}}=\displaystyle\sum_{j=1}^{n}b'_j\end{cases} \tag{8}$$

式中:$M$为BN层的输入特征图;$\mu_i$和$\sigma_i$分别为训练样本中单次迭代的均值和标准差;$\gamma_i$和$\beta_i$为网络学习参数;$W_{:,i,:,:}$为训练后的卷积核权重;$W'_{:,i,:,:}$和$b'_i$分别为转换后的权重和偏置;$W'_{\text{new}}$和$b'_{\text{new}}$分别为将多分支结构合并为单分支结构后的卷积核权重和偏置。

RepMiX利用混合算子进行信道混合,促进信道信息交换,并对上一步生成的特征图进行特征提取。整合从不同分支获得的特征图,并通过BN层和ReLU激活函数引入非线性映射,更有效地结合通道间的线性信息。

### 2.3 空间算子设计

构建学习特征信息的多分支空间算子,称为RepASO,如图2所示。RepASO的每个分支都有不同的作用,实现学习多尺度特征的目的[19]。

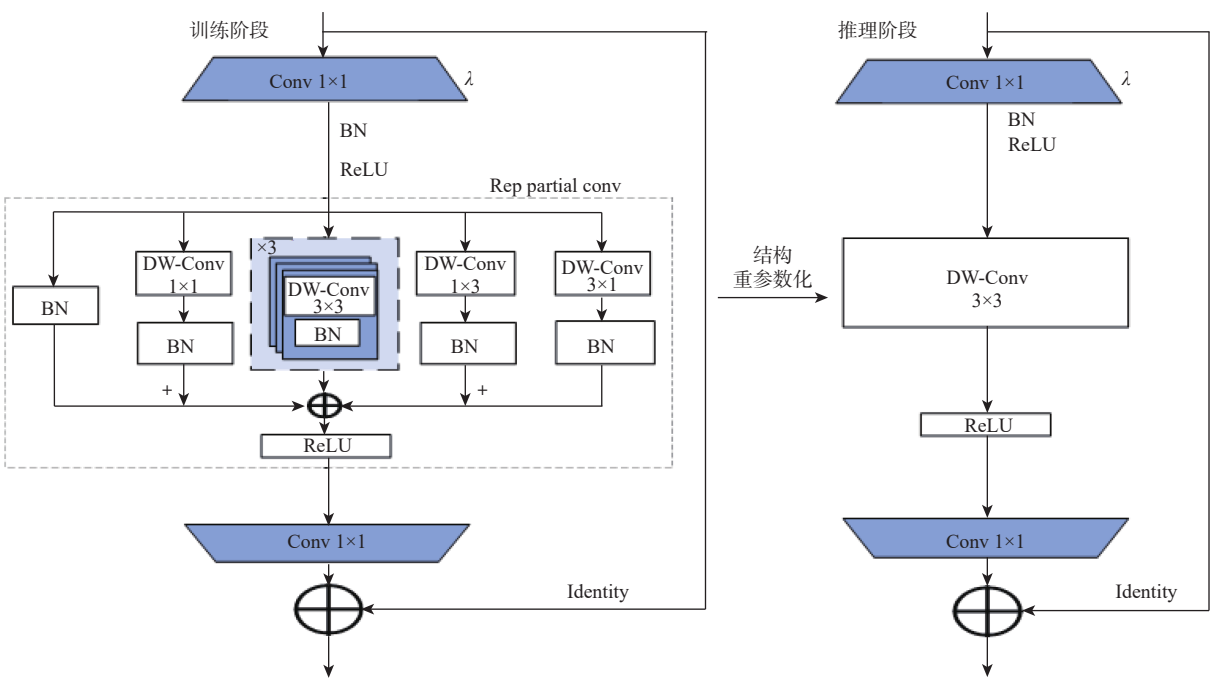



图2 重参数化非对称空间算子

Fig. 2 Reparameterized asymmetric spatial operator

卷积核的大小决定了其能观察到的输入数据的局部范围,也称为感受野。与5×5或7×7等较大的卷积核相比,3×3卷积核的感受野较小,但使用3×3卷积核可以显著减少参数。通过堆叠多层3×3卷积,网络可以学习局部和全局范围的特征。因此,本文选择参数和Flops较少的3×3卷积核。为补偿减少的特征通道,构建$N$个(默认情况下$N$=3)3×3 DW-Conv并行分支。此外,构建一个1×1 DW-Conv分支,组合1×1卷积和DW-Conv,提取中心位置[20]多个通道之间的非线性关系。为缓

解梯度消失，加速网络收敛，增加身份映射分支。

在处理具有明显水平或垂直结构的图像时，需要捕捉不同方向的特征，因此，构建水平1×3和垂直3×1 DW-Conv 并行分支，与3×3 DW-Conv 分支一起构成了非对称卷积结构。为提供更详细的解释，$\boldsymbol{F}'$为融合卷积核，$\boldsymbol{b}$ 为偏置，$\boldsymbol{F}$、$\overline{\boldsymbol{F}}$和$\hat{\boldsymbol{F}}$分别为3×3、1×3和3×1卷积核变换的输出，有

$$\boldsymbol{F}'=\frac{\gamma}{\sigma}\boldsymbol{F}\oplus\frac{\overline{\gamma}}{\overline{\sigma}}\overline{\boldsymbol{F}}\oplus\frac{\hat{\gamma}}{\hat{\sigma}}\hat{\boldsymbol{F}} \tag{9}$$

$$\boldsymbol{b}=-\frac{\gamma\mu}{\sigma}-\frac{\overline{\gamma}\overline{\mu}}{\overline{\sigma}}-\frac{\hat{\gamma}\hat{\mu}}{\hat{\sigma}}+\beta+\overline{\beta}+\hat{\beta} \tag{10}$$

式中：$\gamma$、$\overline{\gamma}$、$\hat{\gamma}$、$\beta$、$\overline{\beta}$和$\hat{\beta}$分别为 3 条卷积分支的网络学习参数；$\mu$、$\overline{\mu}$、$\hat{\mu}$、$\sigma$、$\overline{\sigma}$和$\hat{\sigma}$分别为 3 条卷积分支训练后得到的均值和标准差。

在融合非对称卷积分支后，相当于一个新的卷积操作，其中，卷积核和偏置是多分支训练过程中获得的参数，用$\boldsymbol{y}=\boldsymbol{F}'*\boldsymbol{X}+\boldsymbol{b}$表示。图 3 直观展示了上述过程。结合上述结构，得到了一种名为 RepASO 的多分支算子，其具有七分支结构。7 个分支中的每个分支都配有一个 BN 层，每个分支的输出相加后进行前向传播。

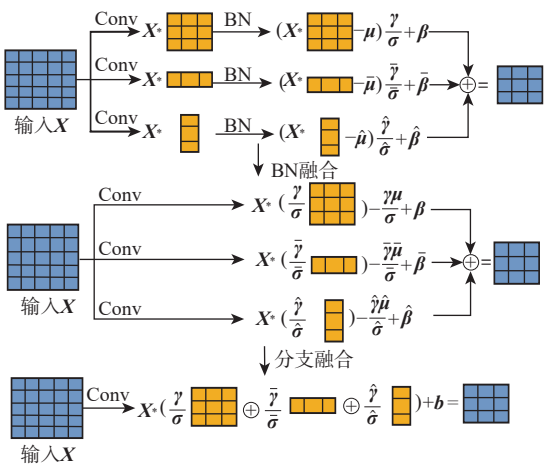



图 3　非对称卷积融合原理

Fig. 3　Principle of asymmetric convolution fusion

RepASO 是对 Inception[21] 结构的简单借鉴，其用参数和计算成本更低的 DW-Conv 取代了普通卷积，并引入非对称卷积来学习水平和垂直方向特征，但 RepASO 并不简单。根据卷积的可加性和结构重参数化原理，在推理过程中可以将这 7 个分支重参数化为一个 DW-Conv 分支，如图 2 所示，既加快了模型推理速度，又不损失模型学习能力，即 RepASO 结合了功能丰富的多分支拓扑结构和结构重参数化技术，前者在模型训练过程中提高模型学习能力，后者在模型推理过程中将多功能分支重参数化为单个分支，加快推理速度，而不仅仅是简单的 DW Inception 结构。

在 RepASO 中，每个分支都利用 DW-Conv 来有效减少 Flops，但其不能简单地取代传统的卷积，因为可能会导致模型精度下降。因此，在多分支结构之后，增加了一个 1×1 PW-Conv。通过这种 PW-Conv 操作，前一层输出特征图的通道之间进行信息交换，以弥补 DW-Conv 带来的精度损失。此外，PW-Conv 还能将前一层输出特性图的通道数量减半，从而降低模型参数和计算复杂度，提高模型效率。

### 2.4　合并基本块

在基于结构重参数化的 DSCNN 模型中，每个 stage 都由基于结构重参数化的 DSCNN 块堆叠而成，且每个 stage 后面都有一个合并层，该合并层的作用是进行图像块下采样操作，主要是使用大小为 2×2、步长为 2 的卷积核对上一层输出进行通道扩充和特征图压缩，作为模型的通道算子。合并基本块由空间算子和通道算子组成，即每个 stage 及其后面的合并层构成了基于结构重参数化的 DSCNN 模型的基本块。基本块是轻量级 CNN 的关键组成部分，负责提取和整合特征，通过串联和堆叠的方式形成整个模型的架构。

### 2.5　基于结构重参数化的 DSCNN

基于基于结构重参数化的 DSCNN 的整体架构（包括 RepMiX、RepASO 空间算子和合并层），本文提出一种名为基于结构重参数化的 DSCNN 的新型轻量级 CNN 模型，表 1 列出了基于结构重参数化的 DSCNN 的架构。实验表明，基于结构重参数化

**表 1　基于结构重参数化的 DSCNN 网络架构**

**Table 1　Architecture of DSCNN based on structural reparameterization**

| 名称 | 输出大小 | 模型 | 数量 |
|---|---|---|---|
| Stalk | $\frac{h}{4}\times\frac{w}{4}$ | 3×3　40<br>RepMixer　40<br>FFN　40 | 1 |
| Stage 1 | $\frac{h}{4}\times\frac{w}{4}$ | 1×1　80<br>RepASO　80<br>1×1　40 | 1 |
| Merge layer | $\frac{h}{8}\times\frac{w}{8}$ | 2×2 Conv　步长 2<br>BN　ReLU | 1 |
| Stage 2 | $\frac{h}{8}\times\frac{w}{8}$ | 1×1　160<br>RepASO　160<br>1×1　80 | 2 |
| Merge layer | $\frac{h}{16}\times\frac{w}{16}$ | 2×2 Conv　步长 2<br>BN　ReLU | 1 |
| Stage 3 | $\frac{h}{16}\times\frac{w}{16}$ | 1×1　320<br>RepASO　320<br>1×1　160 | 8 |
| Merge layer | $\frac{h}{32}\times\frac{w}{32}$ | 2×2 Conv　步长 2<br>BN　ReLU | 1 |
| Stage 4 | 7×7 | 1×1　640<br>RepASO　640<br>1×1　320 | 2 |
| Head | 1×1 | 7×7 AvgPool2d　320<br>1×1　1 280<br>FC　1 000 | 1 |

的DSCNN在参数量、浮点运算数和浮点运算速度方面优于现有的轻量级CNN模型，同时，保持了有竞争力的精度。

## 3 实 验

为验证模型的有效性，在Tiny-ImageNet-200、CIFAR-10和CIFAR-100数据集上进行了性能测试。为评估模型的泛化性能，通过实地拍摄创建了铝锭表面缺陷分类数据集，并与多个轻量级CNN模型进行比较。实验环境如表2所示。

表2 实验环境

Table 2 Experimental environment

| 序号 | 环境 | 版本 |
|---|---|---|
| 1 | Use the system | Ubuntu 22.04 LTS |
| 2 | PyTorch | 1.13.1 |
| 3 | CUDA | 11.6 |
| 4 | GPU | NVIDIA RTX3090Ti |
| 5 | CPU | i9-12900KF |
| 6 | PyCharm | 2022 Community |
| 7 | Python | 3.9 |
| 8 | RAM | 32 GB |
| 9 | SSD | 500 GB |

### 3.1 数据集

使用Tiny-ImageNet-200、CIFAR-10和CIFAR-100数据集验证了本文模型的有效性。Tiny ImageNet-200数据集包含200个不同的类别，图像大小为224像素×224像素，共包含100 000幅图像，每类分别包含500幅训练图像、50幅验证图像和50幅测试图像。CIFAR-10数据集由10类彩色图像组成，每类包含6 000张图像，图像大小为32像素×32像素，包含60 000幅彩色图像，50 000张图像用于训练，其余用于验证。CIFAR-100数据集包含100个类别的60 000张彩色图像，图像大小为32像素×32像素，每个类别500幅图像用于训练，100幅图像测试。通过实地拍摄创建了铝锭表面缺陷分类数据集，用于对铝锭表面缺陷进行分类。铝锭表面缺陷分类数据集由约2 500张不同尺寸的图像组成，数据集按1∶9的比例分为测试集和训练集。

### 3.2 实验参数设置

对于Tiny-ImageNet-200、CIFAR-10和CIFAR-100图像分类任务，对数据进行随机裁剪、随机水平翻转和归一化增强[22]。使用AdamW(Adam with Weight Decay)优化器进行优化，权重衰减系数为0.005，动量为0.9。批量大小设置为240，模型训练时间为300个epoch，初始学习率设置为0.004，并采用余弦退火学习率调整策略，使用20个epoch预热，初始值为$1\times10^{-6}$，最小学习率设为$1\times10^{-5}$，调整公式为

$$l = l_{\min}+0.5(l_{\max}-l_{\min})\cdot(1+\cos(t_{\text{cur}}/T_{\text{total}}\pi)) \tag{11}$$

式中：$l$为当前训练步骤的学习率；$l_{\min}$为最小学习率；$l_{\max}$为初始学习率；$t_{\text{cur}}$为当前训练步骤的索引(从1开始)；$T_{\text{total}}$为训练步骤的总数。

使用软目标交叉熵作为损失函数，在传统的交叉熵损失函数中，目标标签以单击编码的形式表示。而使用软目标交叉熵时，目标标签是一种概率分布，具体计算公式为

$$\boldsymbol{L} = -\frac{1}{B_{\text{size}}}\sum_{j=1}^{B_{\text{size}}}\sum_{i=1}^{n}\boldsymbol{y}_{ij}\ln\hat{\boldsymbol{y}}_{ij} \tag{12}$$

式中：$B_{\text{size}}$为批量大小；$\boldsymbol{y}_{ij}$为目标标签的概率分布；$\hat{\boldsymbol{y}}_{ij}$为模型的预测输出。

对于每个类别，计算真实概率与预测概率之间的交叉熵，将所有类别的交叉熵相加，得出最终损失值。

在铝锭表面缺陷图像分类任务中，使用与上述相同的方法对捕获的照片数据进行缺陷增强操作。使用AdamW优化器进行优化，权重衰减系数为0.005，动量为0.9。批量大小设置为16，进行100次训练，初始学习率为0.004。采用余弦退火学习率调整策略来调整学习率，选择软目标交叉熵作为损失函数。

### 3.3 实验结果与分析

比较本文提出的基于结构重参数化的DSCNN模型与现有轻量级CNN模型的性能，包括MobileNeXt、FasterNet[23]、EdgeNeXt_S/XS/XXS、MobileVit_S/XS/XXS、GhostNetV1/V2、ConvNeXt_Tiny/Small、ShuffleNetV1/V2，及重量级CNN模型ResNet-50的性能。

表3列出了基于结构重参数化的DSCNN与其他轻量级CNN模型在Tiny-ImageNet-200数据集上的实验结果。本文目的是要达到模型精度与速度之间的均衡，因此，根据模型的Top-1精度，将实验数据分为4组，以比较近似精度的模型之间的速度和模型大小。通过设置不同的模型输入特征维度和基于结构重参数化的DSCNN块的堆叠数量，得到基于结构重参数化的DSCNN模型的4个不同版本DSCNN_T0/1/2/S；其中，DSCNN_T0/1/2每个stage的基于结构重参数化的DSCNN块数指定为[1, 2, 8, 2]，输入特征维度分别为40、64和96。DSCNN_S每个stage的基于结构重参数化的DSCNN块数指定为[1, 2, 13, 2]，输入特征维度为128。从表3可以得出，基于结构重参数化的DSCNN模型在各组GPU上实现了更高的吞吐量和浮点运算速度。

表 3　基于结构重参数化的 DSCNN 与其他轻量级 CNN 模型在 Tiny-ImageNet-200 数据集上的性能比较

Table 3　The performance comparison of DSCNN based on structural reparameterization and other light-weight CNN models on the Tiny-ImageNet-200 dataset

| 模型 | 参数量 | 浮点运算数 | GPU吞吐量/(帧·s$^{-1}$) | 浮点运算速度/$10^9$ s$^{-1}$ | Top-1精度/% |
|---|---|---|---|---|---|
| EdgeNeXt_XXS | $1.3\times10^6$ | $0.20\times10^9$ | 5 268 | 29.62 | 63.26 |
| MobileVit_XXS | $1.3\times10^6$ | $0.37\times10^9$ | 3 472 | 27.21 | 63.81 |
| FasterNet_T0 | $3.9\times10^6$ | $0.34\times10^9$ | 8 609 | 50.21 | 58.30 |
| DSCNN_T0 | $3.7\times10^6$ | $0.32\times10^9$ | 6 912 | 44.02 | 60.52 |
| EdgeNeXt_XS | $2.34\times10^6$ | $0.41\times10^9$ | 3 894 | 32.16 | 66.56 |
| MobileVit_XS | $2.39\times10^6$ | $0.92\times10^9$ | 1 964 | 33.37 | 67.55 |
| GhostNetV2 | $6.13\times10^6$ | $0.17\times10^9$ | 4 228 | 14.84 | 66.89 |
| FasterNet_T1 | $7.61\times10^6$ | $0.85\times10^9$ | 5 364 | 56.17 | 64.35 |
| DSCNN_T1 | $7.38\times10^6$ | $0.89\times10^9$ | 4 782 | 50.57 | 65.48 |
| ConvNeXt_Tiny | $28.59\times10^6$ | $3.47\times10^9$ | 1 143 | 43.15 | 71.96 |
| ResNet-50 | $25.55\times10^6$ | $3.33\times10^9$ | 1 466 | 42.56 | 72.03 |
| MobileNeXt | $3.51\times10^6$ | $0.31\times10^9$ | 4 128 | 32.04 | 68.90 |
| EdgeNeXt_S | $5.59\times10^6$ | $0.97\times10^9$ | 2 598 | 41.93 | 71.61 |
| MobileVit_S | $5.64\times10^6$ | $1.79\times10^9$ | 1 535 | 41.07 | 70.30 |
| FasterNet_T2 | $14.98\times10^6$ | $1.91\times10^9$ | 3 541 | 61.19 | 68.52 |
| DSCNN_T2 | $14.19\times10^6$ | $1.90\times10^9$ | 2 981 | 55.15 | 69.60 |
| Swin_T | $28.3\times10^6$ | $4.51\times10^9$ | 600 | 36.49 | 78.11 |
| DSCNN_S | $29.1\times10^6$ | $4.38\times10^9$ | 1 435 | 63.25 | 78.90 |

从表 3 可以看出，DSCNN_T0/1/2 的浮点运算速度与 MobileViT_S/XS/XXS 模型大小相当，但在 GPU 上基于结构重参数化的 DSCNN 模型的吞吐量分别比 MobileViT_S/XS/XXS 高出 2.0 倍、2.4 倍和 1.9 倍。在基于 Top-1 精度水平的 4 组中，DSCNN_T0/1/2 在 GPU 上的吞吐量和浮点运算速度均高于其他轻量级网络，仅低于 FasterNet_T0/1/2。但值得注意的是，DSCNN_T0/1/2 的 Top-1 精度比 FasterNet_T0/1/2 高 1%～2%。此外，DSCNN_S 的 Top-1 精度达到 78.90%，与新兴的 swin_T 模型不相上下。此外，在 GPU 上，DSCNN_S 的推理吞吐量提高了 2.39 倍，浮点运算速度提高了 1.73 倍。表 3 的结果分析表明，基于结构重参数化的 DSCNN 在精度和速度之间取得了理想的平衡。图 4 显示，基于结构重参数化的 DSCNN 在精度和吞吐量之间的平衡方面明显优于其他方法。

在 CIFAR-10 和 CIFAR-100 数据集上进行对比实验，结果如表 4 所示。精度主要检验预测结果是否正确，计算公式为

$$P_{\text{recision}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}} \tag{13}$$

式中：$T_{\text{P}}$表示真阳性情况；$F_{\text{P}}$表示假阳性情况。

表 4 中，基于结构重参数化的 DSCNN 达到了 90.42% 的最佳精度，这验证了包括 RepASO 在内的所提组件的有效性和泛化能力。除了在 CIFAR-10 数据集上的精度比 MobileViT_XS 低 0.37% 之外，DSCNN_T0 在保持较低参数和浮点运算数（分别比 MobileViT_XS 低 $1.67\times10^6$ 和 $0.6\times10^9$）的同时，表现优于其他方法，0.37% 的精度损失在可接受范围内。MobileVit_XXS 的浮点运算次数与 DSCNN_T0 相近，但在 CIFAR-10 和 CIFAR-100 数据集上的精度分别比 DSCNN_T0 低 6.11% 和 9.46%。此外，与 FasterNet_T0 相比，DSCNN_T0 的参数量和浮点运算次数相当或更低，但在 CIFAR-10 和 CIFAR-100 数据集上的精度分别提高了 1.01% 和 1.14%。结果表明，本文模型在降低复杂度的同时，保持了良好的表达能力和有竞争力的性能。

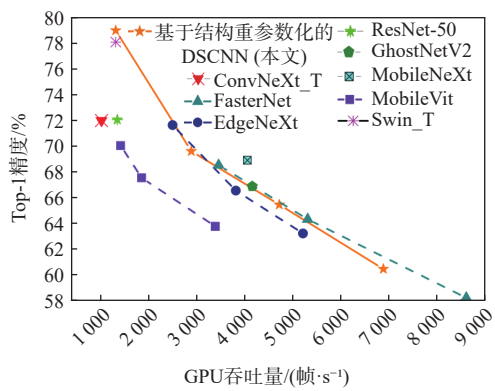



图 4　DSCNN 与其他轻量模型的性能对比

Fig. 4　Performance comparison between DSCNN and other lightweight models

铝锭表面缺陷分类数据集的样本图像如图 5

表 4 在 CIFAR-10 和 CIFAR-100 数据集上比较基于结构重参数化的 DSCNN 与其他轻量级 CNN 模型

Table 4 Comparing DSCNN based on structural reparameterization with other light-weight CNN models on CIFAR-10 and CIFAR-100 datasets

| 模型 | 参数量 | 浮点运算数 | CIFAR-10 精度/% | CIFAR-100 精度/% |
|---|---|---|---|---|
| EdgeNeXt_XS | $1.83\times10^6$ | $0.41\times10^9$ | 86.97 | 62.83 |
| ShuffleNetV1 | $2.28\times10^6$ | $0.16\times10^9$ | 85.98 | 61.35 |
| MobileNeXt | $2.51\times10^6$ | $0.31\times10^9$ | 86.77 | 54.30 |
| GhostNetV2 | $6.13\times10^6$ | $0.17\times10^9$ | 86.17 | 62.43 |
| MobileVit_XXS | $1.30\times10^6$ | $0.37\times10^9$ | 84.31 | 56.53 |
| DSCNN_T0 | $3.45\times10^6$ | $0.32\times10^9$ | 90.42 | 65.99 |
| FasterNet_T0 | $3.66\times10^6$ | $0.34\times10^9$ | 89.41 | 64.85 |
| ConvNeXt_Tiny | $27.97\times10^6$ | $3.47\times10^9$ | 89.07 | 64.76 |
| ResNet-50 | $25.55\times10^6$ | $3.33\times10^9$ | 90.57 | 65.76 |
| MobileVit_XS | $5.12\times10^6$ | $0.92\times10^9$ | 90.79 | 65.17 |

和图 6 所示。在研究该数据集时，对基于结构重参数化的 DSCNN 与多个轻量级 CNN 模型的准确性、网络参数和浮点运算数进行了比较。图 7 直观展示了基于结构重参数化的 DSCNN 模型的性能。图中：圆圈大小表示浮点运算数。可以看出，与重量级 CNN 模型 ResNet-50 相比，基于结构重参数化的 DSCNN 的参数量和浮点运算数明显较低，而检测精度却高出约 6%。与 MobileNeXt、MobileVit-s、EdgeNeXt-s 和 GhostNetV2 等轻量级 CNN 模型相比，基于结构重参数化的 DSCNN 的检测精度分别提高 1.5%、3%、5.5% 和 6.6%。

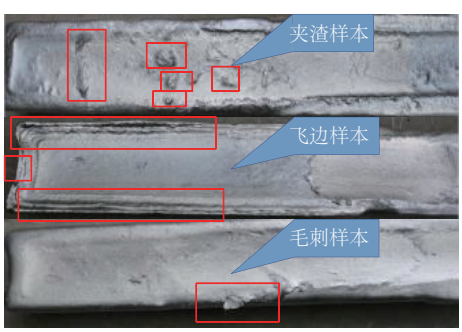



图 5 缺陷样本

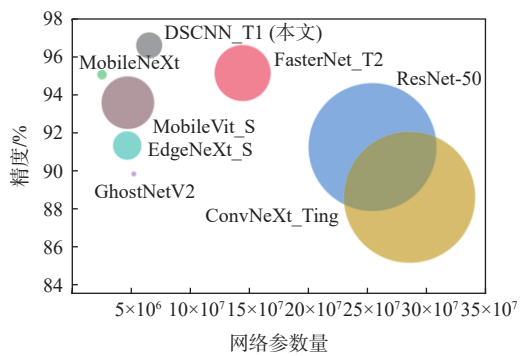
Fig. 5 Defective sample

为直观地显示模型提取了哪些特征进行预测，使用铝锭表面缺陷的图像进行特征图可视化，使用 3 张不同类型缺陷的图像来生成热力图，如图 8 所示。图中：热力图中的颜色越鲜艳，说明网络对该区域越感兴趣，该区域越容易被观察到。与其他方法相比，基于结构重参数化的 DSCNN 模型能更好地覆盖样本图像中的缺陷区域，且更关注更突出、更容易观察到的位置（热力图中最亮的颜色）。即基于结构重参数化的 DSCNN 模型学会了利用目标对象



图 6 合格样本

Fig. 6 Qualified sample




圆圈大小代表 Flops

图 7 基于结构重参数化的 DSCNN 和多个轻量级 CNN 模型在铝锭表面缺陷分类数据集上的性能比较

Fig. 7 Performance comparison of DSCNN based on structural reparameterization and multiple light-weight CNN models on aluminum ingot defect detection dataset

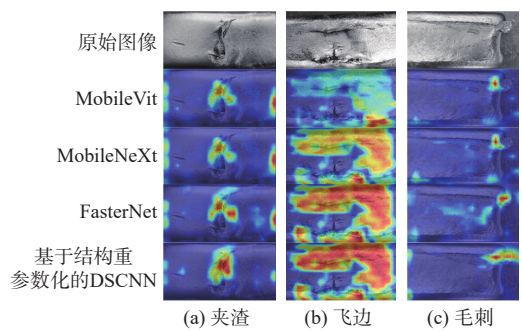



图 8 不同模型热力图的比较

Fig. 8 Comparison of heat maps of different models

区域的信息，能够关注具有不同特征和高辨别度的特征，并从中提取特征进行分类和回归。

### 3.4 消融实验

为探索 RepASO 空间算子的最佳配置，进行消融研究。表 5 列出了实验结果，表中的每一行代表在 RepASO 中包含不同核大小的卷积、身份映射、非对称卷积和 DW-Conv 分支的情况。实验 A、B 和 C 表明，使用大的卷积核并不能提高模型的性能。相反，其会对模型的性能产生负面影响。此外，实验 C、D、E、F、G、H、I 和 J 验证了添加非对

表 5　不同配置的 RepASO 在各个数据集上的实验结果比较

Table 5　Comparison of experimental results on various datasets with different configurations of RepASO

| 序号 | 卷积个数 | | | | | | 身份映射 | CIFAR-10 精度/% | CIFAR-100 精度/% | Tiny-ImageNet Top-1精度/% | 铝锭表面缺陷分类数据集精度/% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 7×7 | 5×5 | 3×3 | 3×1 | 1×3 | 1×1 | | | | | |
| A | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 88.18 | 61.75 | 58.92 | 92.89 |
| B | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 88.24 | 62.80 | 59.07 | 92.95 |
| C | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 88.44 | 63.24 | 59.20 | 93.15 |
| D | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 88.85 | 63.24 | 59.32 | 93.44 |
| E | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 88.87 | 63.34 | 59.37 | 93.54 |
| F | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 88.57 | 63.13 | 59.42 | 93.47 |
| G | 0 | 0 | 3 | 1 | 1 | 0 | 0 | 88.88 | 63.60 | 59.29 | 93.34 |
| H | 0 | 0 | 3 | 1 | 1 | 1 | 0 | 88.89 | 63.68 | 60.31 | 94.14 |
| I | 0 | 0 | 3 | 1 | 1 | 0 | 1 | 88.78 | 63.51 | 59.32 | 94.38 |
| J | 0 | 0 | 3 | 1 | 1 | 1 | 1 | 90.42 | 65.99 | 60.52 | 96.50 |

称卷积分支和多个 3×3 DW-Conv 分支可以提高表示能力。

对 RepMiX 进行消融分析，获得了 RepMiX 的最佳结构形式，主要确定 RepMixer 中 DW-Conv 分支的最佳数量（用 $r$ 表示），表 6 列出了实验结果。可以看出，每增加一条分支，训练阶段的模型参数量仅增大 $0.02\times10^6$，当 $r$=4 时，模型开始出现过拟合。因此，$r$=3 为最佳，此时模型在 CIFAR-10 和 CIFAR-100 数据集上的准确率最高；3 条 DW-Conv 分支以增加较少的训练阶段模型参数为代价，实现了更高的识别准确率。这种设计允许用户为不同的推理设备任意选择合适的计算块，如果对识别准确率要求不高，也可以减少分支数以降低内存开销。

表 6　RepMixer 中不同 $r$ 值的 RepMiX 在 CIFAR-10 和 CIFAR-100 数据集上的定量实验结果

Table 6　Quantitative experimental results of the RepMiX with different $r$ values in RepMixer on CIFAR-10 and CIFAR-100 datasets

| $r$ | 原始模型参数量 | CIFAR-10精度/% | CIFAR-100精度/% |
|---|---|---|---|
| 1 | $3.46\times10^6$ | 89.03 | 63.52 |
| 2 | $3.52\times10^6$ | 89.33 | 63.83 |
| 3 | $3.58\times10^6$ | 90.42 | 65.99 |
| 4 | $3.64\times10^6$ | 88.17 | 63.55 |

## 4　结　论

针对当前 CNN 模型较大、对设备计算能力要求较高和实时性较差的问题，本文提出了一种轻量化网络模型——基于结构重参数化的 DSCNN。

1）RepMiX 和 RepASO 在训练时的多分支并结构可融合不同信道和多尺度特征信息，并行架构避免了网络深度带来的过拟合问题。

2）在推理过程中利用训练保存的参数将 RepMiX 和 RepASO 重参数化为单分支，在保证原有性能的同时，简化模型结构，加速了模型推理。

3）模块设计时，结合 DS-Conv 替代普通卷积的思想，使用 DW-Conv 设计多分支结构，之后增加一个 PW-Conv，弥补了 DW-Conv 使用造成的精度损失，以本文使用的3×3卷积核来计算，模型参数量和计算量下降到使用普通卷积的 1/9 左右。

4）在 Tiny-ImageNet-200、CIFAR-10、CIFAR-100 数据集及自建铝锭表面缺陷分类数据集上，基于结构重参数化的 DSCNN 在保持相当精度的同时，参数和 Flops 更少，浮点运算速度更快，在速度和精度之间实现了很好的权衡。

## 参考文献（References）


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

[2] DING X H, ZHANG X Y, HAN J G, et al. Scaling up your kernels to 31×31: revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11953-11965.

[3] TAN M X, LE Q V. EfficientNetV2: smaller models and faster training[EB/OL]. (2021-06-23)[2024-01-05]. https://doi.org/10.48550/arXiv.2104.00298.

[4] LIU Z, MAO H, WU C Y, et al. A ConvNet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11966-11976.

[5] YU W H, ZHOU P, YAN S C, et al. InceptionNeXt: when inception meets ConvNext[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 5672-5683.

[6] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4510-4520.

[7] ZHANG J N, LI X T, LI J, et al. Rethinking mobile block for efficient attention-based models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2024: 1389-1400.

[8] LIU S W, CHEN T L, CHEN X H, et al. More ConvNets in the 2020s: Scaling up kernels beyond 51×51 using sparsity[EB/OL]. (2023-03-03)[2024-01-07].https://doi.org/10.48550/arXiv.2207.03620.

[9] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.

[10] SUN K, LI M J, LIU D, et al. IGCV3: interleaved low-rank group convolutions for efficient deep neural networks[EB/OL]. (2018-07-20)[2024-01-08]. https://doi.org/10.48550/arXiv.1806.00178.

[11] GAO H Y, WANG Y, CAI L, et al. ChannelNets: compact and efficient convolutional neural networks via channel-wise convolutions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2570-2581.

[12] DING X H, ZHANG X Y, MA N N, et al. RepVGG: making VGG-style ConvNets great again[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13728-13737.

[13] DING X H, GUO Y C, DING G G, et al. ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 1911-1920.

[14] VASU P K A, GABRIEL J, ZHU J, et al. MobileOne: an improved one millisecond mobile backbone[EB/OL]. (2023-03-28)[2024-01-10]. https://doi.org/10.48550/arXiv.2206.04040.

[15] LIU Z, HU H, LIN Y T, et al. Swin transformer V2: scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11999-12009.

[16] TROCKMAN A, KOLTER J Z. Patches are all you need?[EB/OL]. (2022-01-24)[2024-01-12]. https://doi.org/10.48550/arXiv.2201.09792.

[17] WANG A, CHEN H, LIN Z J, et al. RepViT: Revisiting mobile CNN from ViT perspective[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 15909-15920.

[18] 邱云飞, 张家欣, 兰海, 等. 融合张量合成注意力的改进 ResNet 图像分类模型[J]. 激光与光电子学进展, 2023, 60(6): 0610008.
QIU Y F, ZHANG J X, LAN H, et al. Improved ResNet image classification model based on tensor synthesis attention[J]. Laser & Optoelectronics Progress, 2023, 60(6): 0610008(in Chinese).

[19] 朱逢乐, 刘益, 乔欣, 等. 基于多尺度级联卷积神经网络的高光谱图像分析[J]. 吉林大学学报 (工学版), 2023, 53(12): 3547-3557.
ZHU F L, LIU Y, QIAO X, et al. Analysis of hyperspectral image analysis based on multi-scale cascaded convolutional neural network[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(12): 3547-3557 (in Chinese).

[20] 程小辉, 李钰, 康燕萍. 基于中间图特征提取的卷积网络双标准剪枝[J]. 计算机工程, 2023, 49(3): 105-112.
CHENG X H, LI Y, KANG Y P. Double standard pruning of convolution network based on feature extraction of intermediate graph[J]. Computer Engineering, 2023, 49(3): 105-112(in Chinese).

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

[22] 赵利, 王雷全, 张俊三, 等. 基于双通道特征增强的高光谱图像分类[J]. 激光与光电子学进展, 2023, 60(12): 1210012.
ZHAO L, WANG L Q, ZHANG J S, et al. Hyperspectral image classification based on dual-channel feature enhancement[J]. Laser & Optoelectronics Progress, 2023, 60(12): 1210012(in Chinese).

[23] CHEN J R, KAO S H, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 12021-12031.

# Deep separable convolutional neural networks based on structural reparameterization

CHEN Hong[1]，YAN Jianguo[2]，YANG Hua[1]，ZHANG Jing[1]，LI Wei[3]，YANG Jing[1, 4, *]

(1. Electrical Engineering College，Guizhou University，Guiyang 550025，China；
2. China Power Construction Group Guizhou Engineering Co.，Ltd，Guiyang 550025，China；
3. Engineering Research Center of Guizhou Higher Education Institutions，Guiyang 550025，China；
4. Guizhou University，China Guizhou Provincial Key Laboratory of Internet + Intelligent Manufacturing，Guiyang 550025，China)

**Abstract**：A new lightweight convolutional neural network (CNN) model called the deep separable convolutional neural network (DSCNN) based on structural reparameterization is proposed, aiming at the single-branch deep convolutional approach used in the majority of the current CNN models, which not only affects the expressive ability of the model but also occupies a large number of parameters and Flops. Firstly, the feature extraction module (reparameterization MiXer, RepMiX) of the model can fuse information between different channels and spatial locations, realizing multi-scale feature fusion. Secondly, the reparameterized asymmetric spatial operator (RepASO) in the DSCNN learns different channel feature information through a multi-branch structure with different functions, which improves the model feature learning ability; meanwhile, both RepMiX and RepASO combine the structural reparameterization technique and the idea of depth wise separable convolution (DS-Conv) to realize structural decoupling in the training and inference phases, which accelerates the model inference while reducing the model parameters and Flops. Finally, comparative experiments are carried out on the Tiny-imagenet-200, CIFAR-10, and CIFAR-100 datasets in addition to a self-constructed dataset for the classification of aluminum ingot surface defects. The experimental results demonstrate that the DSCNN maintains competitive accuracy while achieving higher floating-point speeds.

**Keywords**：convolutional neural networks；structural reparameterization；depthwise seperable convolution；asymmetric convolution；structural decoupling

**Received**：2024-05-07；**Accepted**：2024-06-25；**Published Online**：2024-07-03 15：00

**URL**：link.cnki.net/urlid/11.2625.V.20240703.0858.001

**Foundation items**：National Natural Science Foundation of China (61640014, 61963009); Innovation Group of Guizhou Education Department (Qianjiaohe KY[2021]012); Guizhou Provincial Science and Technology Projects (Qiankehe Zhicheng [2022]Yiban017, Qiankehe Zhicheng [2023]Yiban411, Qiankehe Zhicheng [2024]Yiban051); Engineering Research Center of Guizhou Education Department (Qianjiaoji[2022]040, Qianjiaoji[2022]043); Science and Technology Project of Power Construction Corporation of China, Ltd (DJ-ZDXM-2020-19, DJ-ZDXM-2022-44); Guizhou Dual Carbon Research Institute Open Subjects (DCRE-2023-13); China Aluminum International Engineering Corporation Limited Science and Technology Major Project (CJ2022JS-ZD02)

*** Corresponding author**. E-mail：jyang7@gzu.edu.cn