中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

# 基于特征分布调整的深度神经网络二值量化方法

刘畅, 陈莹

引用本文:
刘畅,陈莹. 基于特征分布调整的深度神经网络二值量化方法[J]. 控制与决策, 2024, 39(6): 1840−1848.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1945

---

# 基于特征分布调整的深度神经网络二值量化方法

刘 畅, 陈 莹†

(江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122)

**摘 要:** 二值卷积神经网络(BNNs)由于其占用空间小、计算效率高而受到关注. 但由于量化激活特征的正负部分分布不均等问题,二值网络和浮点深度神经网络(DNNs)之间存在着明显的性能差距,影响了其在资源受限平台上的部署. 二值网络性能受限的主要原因是特征离散性造成的信息损失以及分布优化不当造成的语义信息消失. 针对此问题,应用特征分布调整引导二值化,通过调整特征的均值方差均衡特征分布,减小离散性造成的信息损失. 同时,通过分组激励与特征精调模块设计,调整优化量化零点位置,均衡二值化激活分布,最大程度保留语义信息. 实验表明,所提出方法在不同骨干网络、使用不同数据集时均能取得较好效果,其中在CIFAR-10上使用ResNet-18网络量化后网络准确率仅损失0.4 %,高于当前主流先进二值量化算法.

**关键词:** 特征分布；均值方差调整；语义信息保留；模型压缩；二值神经网络；模型量化

**中图分类号:** TP183 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2022.1945

**引用格式:** 刘畅,陈莹. 基于特征分布调整的深度神经网络二值量化方法[J]. 控制与决策, 2024, 39(6): 1840-1848.

# Feature distribution guided binary neural networks

*LIU Chang, CHEN Ying*†

(Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education，Jiangnan University，Wuxi 214122，China)

**Abstract:** In recent years, binary neural networks (BNNs) have received attention due to their small memory consumption and high computational efficiency. However, there exists a significant performance gap between BNNs and floating-point deep neural networks (DNNs) due to problems, such as imbalanced distributions of positive and negative parts of quantized activation features, which affects their deployment on resource-constrained platforms. The main reason for the limited accuracy of binary networks is the information loss caused by feature discretization and the disappearance of semantic information caused by improper distribution optimization. To address this problem, this paper applies feature distribution adjustment to guide binarization, which adjusts the mean-variance of features to balance the feature distribution and reduce the information loss caused by discretization. At the same time, through the design of group excitation and feature fine-tuning module, the quantization zero points are optimized to balance the binarization activation distributions and retain the semantic information to the maximum extent. Experiments show that the proposed method achieves better results on different backbone networks using different datasets, in which only 0.4 % of accuracy is lost after binarizing ResNet-18 on CIFAR-10, which surpasses the current mainstream BNNs.

**Keywords:** feature distribution；mean and variance adjustment；semantic information speicherung；model compression；binary neural networks；neural network quantization

## 0 引 言

近年来,深度卷积神经网络(DNNs)在计算机视觉、语音识别等领域的大多数任务中发挥了巨大作用,如图像分类、物体检测、语义分割和跟踪. 通常情况下,为了提高网络性能,模型的复杂性会随之提升,最先进的深度神经网络通常涉及数百万的参数,在计算过程中需要数十亿的FLOPs,这便给边缘设备的部署带来了相当大的问题. 模型压缩技术应运而生,其中包括模型剪枝[1]、模型量化[2]、模型蒸馏[3]、低秩分解[4]和轻量级模块设计[5-6].

自 Courbariaux 等[7] 在 Binary-connect 中提出二进制神经网络(BNNs)以来,网络二进制化一跃成为

**收稿日期:** 2022-11-10；**录用日期:** 2023-03-06.
**基金项目:** 国家自然科学基金项目(62173160).
**责任编委:** 张国山.
†通讯作者. E-mail: chenying@jiangnan.edu.cn.

最有潜力的网络压缩和加速推理方法之一. 与传统的32位浮点网络相比, BNNs只用1位来表示网络权重及激活, 大大降低了存储负担. 此外, BNNs将浮点乘加运算替换为XNOR和bitcount, 极大地提高了内存访问和卷积操作的效率. 然而, 与32位网络相比, 二值量化网络仍然存在性能大幅下降的问题, 这促使大量研究人员致力于减少全精度网络和二值网络之间的性能差距, 并研究出许多行之有效的解决方案. 为提高二值网络性能, XNOR-Net[8]在权重和激活中引入实值缩放因子, 有效提高了BNN在大规模数据集上的性能. IR-Net[9]在前向传播中对权重进行平衡与标准化, 缓解了二进制化的表现能力和离散性问题, 最大限度地保留了信息. SD-BNN[10]通过自适应调整激活与权重符号分布, 改善输出的符号分布. 众多研究表明, 通过调整权重及激活分布可以在一定程度上保留全精度网络信息、降低特征离散性造成的信息损失.

由于未经处理的图像要学习的特征范围较宽, 特征分布峰值与二值分割阈值距离较大, 对语义信息的识别能力堪忧. 鉴于此, 本文提出基于特征分布调整的深度神经网络二值量化方式(feature distribution guided binary neural networks, FDG-BNN), 通过分组激励模块(grouped excitation module, GEM)优化阈值分割, 配合特征精调模块(feature refinement module, FRM)均衡激活分布, 进而使得BNNs可对特征分布进行动态调整, 更好地利用特征信息进行二值化, 最大程度保留语义信息.

## 1　相关工作

二值神经网络实现了非常高的压缩率, 并提高了内存访问效率, 使其前向推理速度大大提升. 然而, 当网络权重和激活值都被直接量化为1位时, 网络的识别性能会急剧下降. 同时, 由于符号函数不能保证二值化激活仍然保留全精度激活的可辨别特征, 部分二值神经网络算法通过结构调整、降低梯度误差及训练策略的改进方法调整量化方式以影响激活分布. 同时, 大多主流方法通过引入实值缩放因子解决分布问题, 尽管对性能提升有一定帮助, 但将计算量扩大了一倍以上, 很大程度上抵消了二值量化的优势. 因此如何在不提升计算量的情况下, 选择更合理的量化零点以调整激活的分布十分重要.

SD-BNN[10]通过调整通道均值改变量化阈值, 找到最优量化效果. 然而, 只改变均值很难有效地处理各具特异性的激活, 特征的离散性使得分布优化不当时对准确率的影响较大, 在较大的均值调整范围中, 若学习的均值调整参数不精确或不适合输入, 则二值化的输出将失去大量的信息. 因此, 本文提出基于特征分布调整的二值量化方法, 均衡调整特征分布优化阈值分割, 进而更好地保留图片语义信息, 解决二值神经网络离散性带来的优化挑战.

## 2　本文方法

本节首先分析FDG-BNN提出的必要性, 随后详细介绍FDG-BNN的整体架构及其中分组激励模块和特征重组模块的设计.

### 2.1　问题分析

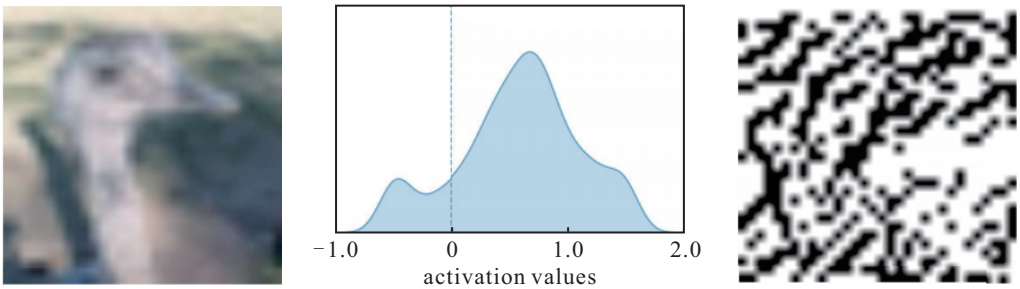
二值神经网络将卷积操作中全精度表示的权重及特征通过sign函数转化为+1、−1, 将原本连续全精度权重及特征转化为离散的1-bit数值, 而其中包含的信息也随之压缩. 在训练实值网络时, 权重和激活是连续的实值, 可以轻易重塑或移动分布. 然而, 对于二值神经网络而言, 对分布的学习关键而困难, 二元卷积中的激活只能从+1、−1中选取数值. 输入实值特征图中, 在符号函数之前做一个小的分布变化可能会导致完全不同的输出二元激活, 这将直接影响到特征的信息量, 并显著影响最终的准确性.

一般情况下, 若输出二元激活特征中具有判别性的有用信息和不具有判别性的无用信息的比例较大, 即信噪比较高, 则网络实值激活特征分布的正负性相对均衡, 对于网络的优化及其性能有益. 相较于实值特征, 二值化后的特征所包含信息量明显减小, 其所包含的有用信息与无用信息必然都随之减少, 而输入实值特征图分布的微小变化能对二值网络中有用信息占比造成的影响增大, 对网络性能的影响更为突出, 故对输入实值特征图进行分布优化显得格外重要. 好的分布优化方式可以很大程度上减小离散性造成的信息损失, 更好地保留其中包含的语义信息.

ReActNet[11]发现了网络性能与特征分布间的联系性, 进而设计了更具泛化性的sign及ReLU函数, 提升了基本网络结构的二值化性能, 也开启了特征分布调整的优化方向. 但仅以函数泛化很难了解实际特征分布, 无法自适应地对其分布进行调整. 此后, IR-Net[9]从信息流角度出发, 通过最大化量化前后特征熵减小信息损失, 进而保证语义信息的保留. 但信息熵配合损失函数的优化方式在直观性上仍稍显不足, 无法均衡特征分布. SD-BNN[10]提出激活自分布及权重自分布两个模块, 直接调整特征及权重的符号分布, 改善卷积输出的符号差异, 进而调整量化零点. 对特征符号的直接调整解决了原本量化零点不是最优分割点的问题, 也实现了一定的性能提升, 但其对语

义信息的保留能力仍较为不足.

图 1 分别展示了 CIFAR-10 中随机选取的一张图像、训练结束后第 1 层卷积的激活分布及其量化零点值(图中竖线)、图片对应二值的化激活图. 如图 1 所示, SD-BNN 方法的量化零点很难对正、负激活做出较好的平衡, 因此影响了激活分布的调整, 进而导致二值化激活图语义信息完全消失. 一般而言, 由于输入图片的均值分布较宽, 离散点分布不集中, 其中若存在部分错误信息则对全局语义信息的影响较大. 虽然通过添加缩放因子可以使二值化后的权重激活更加接近其实际值, 但相同幅度的平均调整在高方差分布中会比低方差分布造成更少的激活符号变化, 难以起到较好的信息提取作用, 即使缩放因子能起到一定的修正效果, 依旧无法解决分布优化不当造成的影响. 基于此, 本文提出基于特征分布的调整方法, 利用原特征的均值方差等信息协助均衡特征分布, 对原有语义信息起到了更好的保留作用.




(a) CIFAR-10 中随机图像 (b) 第 1 层卷积激活分布 (c) 二值化激活图

**图 1 SD-BNN 可视化效果**

### 2.2 基于特征分布调整的二值量化

为进一步减小特征离散性造成的信息损失, 防止分布优化不当造成的语义信息消失, 本文使用分组激活配合特征精调来调整指导量化过程.

#### 2.2.1 总体架构

所提出的 FDG-BNN 对特征进行针对性调整, 可以显著均衡特征分布, 优化阈值分割, 同时调整分布宽度, 提升激活值跳变概率, 优化训练效果.

FDG-BNN 由两个模块组成如图 2 所示, 包括进行全局特征提取、分组精调与零点精调的分组激励模块 GEM (grouped excitation module) 和缩放精调与缩放整合的特征精调模块 FRM (feature refinement module), $A_r$ 和 $A_b$ 分别为该模块的输入和输出特征矩阵.

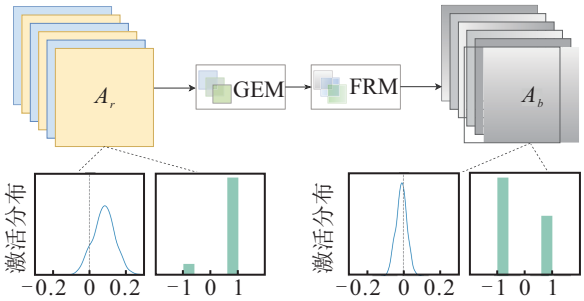



**图 2 FDG-BNN 卷积块结构及特征分布**

GEM 模块使用全局特征提取配合全局平局池化进行分组特征提取, 再进行分组精调, 最后使用零点精调配合全连接、shortcut 将信息整合, 达到调整整体特征的同时以通道方式改变零点的效果.

FRM 模块使用特征精调方式计算并融合 GEM 调整的特征矩阵在空间上的均值和方差, 并使用缩放整合使融合后的特征成为像素级的缩放矩阵, 调整后的激活值缩放到适当分布后, 应用信号函数对其进行二进制化.

#### 2.2.2 GEM

GEM 构造细节如图 3 所示, 其中 FC 表示全连接层. 该结构类似 squeeze-excitation 模块[12], GEM 调整输入张量中每个通道分布的均值, 而非通道权重. 维度为 $C \times H \times W$ 的输入张量被送入 GEM, 通过全局平均池化, 提取出大小为 $C \times 1 \times 1$ 的全局特征, 将该全局特征输入到分组连接中进行特征提取. 分组连接的作用包含两部分, 首先对特征进行分组精调, 其次可通过设置分组数控制复杂度.

分组连接中, 每一个滤波器都只接受其感受野内的特征, 分组间无信息交互, 因此通过全连接层 FC 对各个独立的特征进行聚合并进行后续处理.

首先对输入进行特征提取, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_C] \in \mathbb{R}^{C\times H\times W}$, 针对任意 $\boldsymbol{x}_k \in \boldsymbol{X}$, 有

$$\boldsymbol{y}_k = F_g(\boldsymbol{x}_k) = \frac{1}{H \times W}\sum_{i}^{H}\sum_{j}^{W}\boldsymbol{x}_k(i,j). \quad (1)$$

如式 (1) 所示, 计算得出与 $\boldsymbol{X}$ 内参数一一对应的 $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_C] \in \mathbb{R}^{C\times 1\times 1}$ 再进行分组. 假设全局特征被分成 $g$ 组, 每个全连接层中有 $C/g$ 个输入标

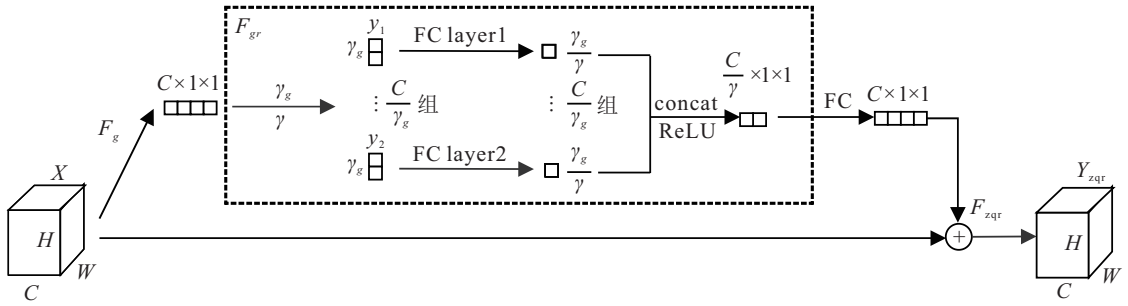



**图 3**　GEM结构

量. 需要注意的是,$g$应该能被$C$整除. 使用超参数$\gamma_g$表示分组后每组标量数$C/g$,即每层拥有多少个输入标量. 每层输出$\gamma_g/\gamma$个标量,其中$\gamma$为通道缩减超参数[12],分组数为$C/\gamma_g$,因此分组连接模块产生的总输出数$N_{\text{gcm}}$可以表示为

$$N_{\text{gcm}}=\frac{\gamma_g}{\gamma}\times g=\frac{\gamma_g}{\gamma}\times\frac{C}{\gamma_g}=\frac{C}{\gamma}. \tag{2}$$

过程中$\boldsymbol{Y}$被拆分为$C/\gamma_g$组,其拆分方式为

$$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\dots,\boldsymbol{y}_g,\dots,\boldsymbol{y}_{\frac{C}{\gamma_g}}]. \tag{3}$$

每组为

$$\boldsymbol{Y}_g=[\boldsymbol{y}_{\gamma_g\times(g+1)+1},\boldsymbol{y}_{\gamma_g\times(g+1)+2},\dots,\boldsymbol{y}_{\gamma_g\times(g+1)+\gamma_g}]\in\mathbb{R}^{\gamma_g\times1\times1}. \tag{4}$$

分组精调在减小计算量的同时,对不重要通道信息起到了过滤作用. 通过使用全局平均池化产生通道的统计数据来实现将全局空间信息分组并重组到一个通道中.

同时,为避免连续全连接层带来的运算资源浪费,针对通道数较多的卷积层激活,用分组精调($F_{gr}$)方式代替连续多个全连接层,不再利用所有通道的均值池化信息,而是通过其中几个相邻通道计算总体表现及所需的偏移情况,再结合标准的全连接方式达到更好的效果. 计算方式如下:

$$\boldsymbol{Y}'_C=F_{gr}(\boldsymbol{Y})=\boldsymbol{W}_{\text{FC}}[\boldsymbol{W}_1\boldsymbol{Y}_1,\boldsymbol{W}_2\boldsymbol{Y}_2,\dots,\boldsymbol{W}_g\boldsymbol{Y}_g,\dots,\boldsymbol{W}_{\frac{C}{\gamma_g}}\boldsymbol{Y}_{\frac{C}{\gamma_g}}]. \tag{5}$$

其中:$[\boldsymbol{W}_1,\boldsymbol{W}_2,\dots,\boldsymbol{W}_g,\dots,\boldsymbol{W}_{\frac{C}{\gamma_g}}]$;$\boldsymbol{W}_g$和$\boldsymbol{Y}_g$分别为划分到第$g$组的权重和特征;$\boldsymbol{W}_g\in\mathbb{R}^{1\times\gamma_g}$,$\boldsymbol{Y}_g\in\mathbb{R}^{\gamma_g\times1\times1}$,$\boldsymbol{W}_{\text{FC}}\in\mathbb{R}^{C\times\frac{C}{\gamma_g}}$为普通全连接层权重矩阵,通过全连接层将$\boldsymbol{Y}'_C$转化为$C\times1\times1$向量,便于与原始输入融合以实现特征分布调整.

随后,为了捕捉通道间信息进行汇总,对分组连接模块的输出使用零点精调($F_{\text{zqr}}$)方式,有

$$\boldsymbol{Y}_{\text{zqr}}=F_{\text{zqr}}(\boldsymbol{Y}'_C)=\boldsymbol{X}+\text{sigmoid}(\boldsymbol{Y}'_C). \tag{6}$$

其中:$\boldsymbol{Y}'_C\in\mathbb{R}^{C\times1\times1}$,结合原图特征得到$\boldsymbol{Y}_{\text{zqr}}\in\mathbb{R}^{C\times H\times W}$,与输入张量尺度一致. 配合ReLU计算激活可保证其灵活性与非相互排斥性,确保网络学习到通道之间的非线性相互作用,并允许多个通道被强调,而非单一激活.

### 2.2.3　FRM

虽然GEM可以对特征图进行通道维度的调整,然而仅通过该调整能带来的性能提升有限. 为了实现特征图在空间维度上的精细化调整,本文提出FRM模块,将输入进行注意力分布调整后,通过空间注意力机制生成灰度变换矩阵,与输入进行按位点乘,对特征图的分布进行精细调整,使其更有判别性.

FRM的构造细节如图4所示. 其中输入张量和输出张量均具有$C$个通道,形状为$H\times W$. 在输入空间中,配合零点精调使用缩放精调($F_s$)与缩放整合($F_{\text{sr}}$)方式针对均值和方差进行聚合,计算出此通道的统计信息. 将均值和方差的统计信息张量通过连

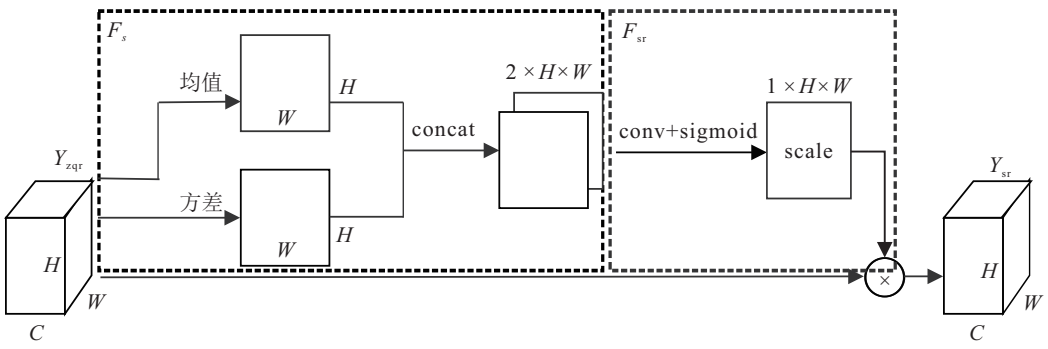



**图 4**　FRM结构

接操作融合得到双通道中间特征图,在卷积和信号函数的作用下得到如下缩放因子矩阵:

$$\boldsymbol{Y}_s = F_s(\boldsymbol{Y}_{\rm zqr}) = [\boldsymbol{y}_{\rm mean}, \boldsymbol{y}_{\rm var}] \in \mathbb{R}^{2\times H\times W}. \quad (7)$$

其中: $\boldsymbol{Y}_{\rm zqr} \in \mathbb{R}^{C\times H\times W}$; $\boldsymbol{y}_{\rm mean} = {\rm Mean}(\boldsymbol{Y}_{\rm zqr}) \in \mathbb{R}^{1\times H\times W}$, ${\rm Mean}(\cdot)$ 表示取均值; $\boldsymbol{y}_{\rm var} = {\rm Var}(\boldsymbol{Y}_{\rm zqr}) \in \mathbb{R}^{1\times H\times W}$, ${\rm Var}(\cdot)$ 表示求方差. 随后,在广播机制下,缩放因子矩阵与输入张量进行逐元相乘,整合得到输出矩阵如下:

$$\boldsymbol{Y}_{sr} = F_{\rm sr}(\boldsymbol{Y}_s) \times \boldsymbol{Y}_{\rm zqr} = {\rm sigmoid}({\rm conv}(\boldsymbol{Y}_s)) \times \boldsymbol{Y}_{\rm zqr}. \quad (8)$$

其中: $\boldsymbol{Y}_s \in \mathbb{R}^{2\times H\times W}$, $F_{\rm sr}(\boldsymbol{Y}_s) \in \mathbb{R}^{1\times H\times W}$, $\boldsymbol{Y}_{\rm zqr} \in \mathbb{R}^{C\times H\times W}$,最终得到输出 $\boldsymbol{Y}_{\rm sr} \in \mathbb{R}^{C\times H\times W}$ 与输入张量尺度一致.

## 3 实验分析

为验证所提出基于特征分布调整的二值量化方法的有效性以及相比于其他主流先进算法的优势,进行比较实验;此外,为验证引入的超参数的稳健性进行消融分析;最后,对所提出方法的复杂度和速度进行进一步分析.

### 3.1 实施细节

使用两个基准分类数据集以评估所提出FDG-BNN的有效性: 1) CIFAR-10包含10种物体的60 000张图像,其中包括50 000张训练图像和10 000张测试图像. 2) TinyImageNet包含200个类别的100 000张训练图像和10 000张验证图像. 文中涉及的验证准确率的其他主流先进方法延用其原文数值,如VGG-small[13]、ResNet-20[14]和ResNet-18[14].

所有实验都基于Pytorch和NVIDIA GTX 2080Ti实现. 在数据增强方面,应用随机水平翻转、 随机裁剪及归一化,CIFAR-10的输入图像尺寸为32 × 32, TinyImageNet为64×64. 在反向传播过程中,采用IR-Net[9]提出的误差衰减估计器(EDE)方式解决符号函数的不可分割性问题. 在训练过程中,批次大小设定为128,初始化学习率为0.08,使用SGD优化器,动量设定为0.9,使用余弦退火调整学习率. 参考SD-BNN[10]实验设置,CIFAR-10训练600个epoch, TinyImageNet训练100个epoch. 除分类器的输出维度不同外,两个数据集上的模型复杂度相同. 每个分组连接层中的分组规模超参数与通道缩减超参数$\gamma$一致,默认情况下参考SD-BNN[10]设置为16.

### 3.2 准确率对比

为验证所提出FDG-BNN的有效性,在CIFAR-10和TinyImageNet数据集上进行实验,并与其他BNN进行比较,包括XNORNet[8]、BNN[15]、BNN-DL[16]、IR-Net[9]、BinaryDuo[17]、ReActNet[11]、ReCU[18]、SD-BNN[10]等,其中SD-BNN[10]的准确率结果来自作者公开的源码在本文相同设置下的运行结果,其余方法的准确率结果来自原文.

表1显示了各种方法在CIFAR-10上的性能比较,* 表示方法使用Bi-RealNet结构. 由表1可见,所提出FDG-BNN在所有情况下均实现了最佳准确率,其中采用ResNet-18的FDG-BNN比SD-BNN提高了0.3 %,并将与全精度对应的二值化性能差距缩小到0.4 %. 对于VGG-Small,所提出FDG-BNN获得了与SD-BNN相同的性能,这是由于使用FDG-BNN的VGG-Small比使用SD-BNN的VGG-Small更深,而且梯度可能在没有short-cut结构的帮助下变得不稳定.

**表1 不同新型的二值卷积神经网络在CIFAR-10分类数据集上的验证**

| 模型 | 方法 | 权重/激活位数 | 准确率/% |
|---|---|---|---|
| VGG-Small | 全精度 | 32/32 | 91.7 |
| | XNOR-Net[8] | 1/1 | 89.8 |
| | BNN[15] | 1/1 | 89.9 |
| | BNN-DL[16] | 1/1 | 90.0 |
| | IR-Net[9] | 1/1 | 90.4 |
| | BinaryDuo[17] | 1/1 | 90.4 |
| | SD-BNN[10] | 1/1 | 90.6 |
| | FDG-BNN | 1/1 | 90.6 |
| ResNet-20 | 全精度 | 32/32 | 90.8 |
| | DSQ[19] | 1/1 | 84.1 |
| | IR-Net[9] | 1/1 | 85.4 |
| | IR-Net* | 1/1 | 86.5 |
| | FDA-BNN[20] | 1/1 | 86.2 |
| | BD-BNN[21] | 1/1 | 86.5 |
| | SD-BNN[10] | 1/1 | 86.5 |
| | FDG-BNN | 1/1 | 86.6 |
| ResNet-18 | 全精度 | 32/32 | 93.0 |
| | BNN-DL[16] | 1/1 | 90.5 |
| | IR-Net[9] | 1/1 | 91.5 |
| | RBNN[22] | 1/1 | 92.2 |
| | BD-BNN[21] | 1/1 | 92.4 |
| | SD-BNN[10] | 1/1 | 92.3 |
| | FDG-BNN | 1/1 | 92.6 |

**表2 不同新型的二值卷积神经网络在TinyImageNet分类数据集上使用ResNet-18的验证**

| 方法 | 权重/激活位数 | 准确率/% |
|---|---|---|
| 全精度 | 32/32 | 62.66 |
| ReActNet[11] | 1/1 | 40.00 |
| BiRealNet[23] | 1/1 | 41.37 |
| BNN-BN[24] | 1/1 | 42.93 |
| SD-BNN[10] | 1/1 | 48.23 |
| IR-Net[9] | 1/1 | 48.64 |
| ReCU[18] | 1/1 | 49.29 |
| FDG-BNN | 1/1 | 49.62 |

表2显示了在TinyImageNet上使用ResNet-18的各种二值化SOTA的性能. 所有实验均在本文设置

下重新运行,并展示top-1准确率结果. 由于任务更具挑战性,性能结果表现出很大的差异性. 可以看出,所提出FDG-BNN较SD-BNN提高了1.39%,并超过ReActNet的9.62%. 实验结果表明,所提出FDG-BNN优于其他SOTA方法.

### 3.3　消融实验

本节通过消融实验进一步表明所提出FDG-BNN的有效性,并解释其中部分结构的设计依据. 消融研究包括5部分:超参数$\gamma_g$大小的选择、GEM与FRM的先后顺序、在空间方面不同处理的效果、算法复杂度以及可视化分析.

1) 超参数$\gamma_g$设置的消融实验.

本次消融实验使用ResNet-18结构在CIFAR-10上进行比较,在$\gamma = 4$的情况下确定每个分组连接块中使用不同$\gamma_g$的相应效果,其中$g \in \left\{1, 2, 4, \dfrac{C}{\gamma}, \dfrac{C}{2\gamma}, \dfrac{C}{4\gamma}\right\}$. 分组的数量对优化效果产生的影响(组的数量为$C/\gamma_g$)如图5所示. 图5中:$g \in \{1, 2, 4\}$表示特征图分组数量固定,$g \in \left\{\dfrac{C}{\gamma}, \dfrac{C}{2\gamma}, \dfrac{C}{4\gamma}\right\}$表示特征图分组数量随通道数自适应调整.

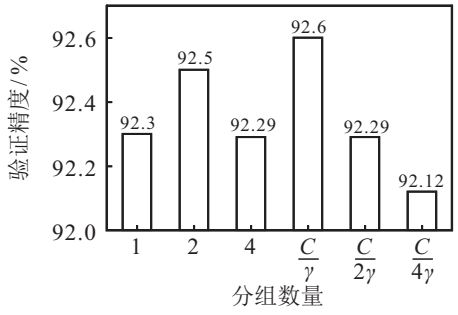



**图5　$\gamma_g$使用ResNet-18在CIFAR-10上的结果**

观察可知,分组操作可对准确率造成一定影响. 当分组数量固定时,适中的分组数量可以达到较好的验证准确率,但依旧不及随通道数自适应调整的分组状态可达到的最高准确率. 当分组数量固定,即$g \in \{1, 2, 4\}$时,性能趋势变化与分组数量无显著相关性,其原因是网络每层的特征图通道数是不相同的,通常而言网络浅层通道数较小,深层通道数较大,固定的分组数量导致卷积无法接收来自上一层的最优输入特征.

在自适应调整的分组中,输入通道数与$\gamma_g$共同决定分组数. 当$\gamma_g$固定时,输入通道数越多则分组数越多,进而在一定程度上降低了网络的复杂度,减小了过拟合. 然而复杂度的降低将不可避免地导致准确率下降,因此在分组规模选取时需考虑两者之间的平衡. 由图5可见:当$\gamma_g = \gamma$时,准确率表现最佳,为92.60;当$\gamma_g = 4\gamma$时,准确率表现最差,为92.12. 值得注意的是,若$\gamma_g = 1$,则分组连接模块退化为全连接模块,这有助于在通道中获得全局注意力信息,性能也将随之提高,但代价是计算量的极大增加. 另一方面,若$\gamma_g = 4\gamma$,则构成全分离的连接模块,对于中间特征的每一个标量而言,这相当于线性变换的操作,因此这样简单的BNN很难去表征一个复杂的高级语义信息. 对于$\gamma_g = \gamma$,每个输出标量均由组内的几个特定激活决定,以此更好地保留语义信息并减少过拟合.

2) 有关GEM与FRM先后顺序的消融实验.

在CIFAR-10和TinyImageNet上的实验表明,GEM与FRM的顺序会影响算法效果. 实验结果如表3所示,FDG-BNN中,GEM在前相比于FRM在前可以获得更好的性能. 在CIFAR-10上使用ResNet-18,GEM在前比FRM在前效果提高了0.3%,使用ResNet-20提高了0.1%. 在TinyImageNet上使用ResNet-18,GEM在前比FRM在前效果提高了1.2%. 上述结果可以在一定程度上说明,即使FRM因为均值聚类具有了调整分布均值的作用,但依旧不能像GEM那样明确地完成均值调整.

**表3　不同新型的二值卷积神经网络在CIFAR-10和TinyImageNet数据集上的验证**

| 数据集 | 模型 | 设定 | 准确率/% |
|---|---|---|---|
| CIFAR-10 | ResNet-18 | FRM在前 | 92.3 |
| | | GEM在前 | 92.6 |
| | ResNet-20 | FRM在前 | 86.5 |
| | | GEM在前 | 86.6 |
| TinyImageNet | ResNet-18 | FRM在前 | 48.4 |
| | | GEM在前 | 49.6 |

3) 对FRM中不同空间处理的消融实验.

为进一步确认改进效果,本文应用平均聚合和方差聚合两种聚合方式. 通过对比实验讨论两种操作的效果,包括在FRM中只使用平均聚合和只使用方差聚合,并进行了一个2核卷积聚合的对照实验.

如图6所示:“仅均值”和“仅方差”分别表示FRM中仅应用了相应的聚合,C2 conv表示有2个核的卷积聚合,“均值 & 方差”表示所提出方法的性能. 方差聚合的性能比均值聚合的性能好0.01,表明FRM中的方差运算是有效的,但均值与方差的组合可以使性能达到最佳,这有助于稳定GEM分布的调整. 卷积聚合获得了最差的性能,并使所提出的FDG-BNN准确率下降了1.33%,这意味着卷积引入的线性变换很难明确地产生有效的分布比例因子. 此外,卷积聚合引入了更多的参数和计算成本,而平均和方差操作需要少量计算开销.

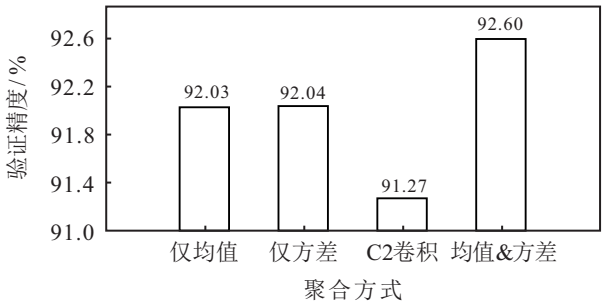



**图 6** ResNet-18在CIFAR-10上使用不同空间处理结果

4) 复杂度和速度.

每个组的组数不同、分组连接块的数量不同导致了不同的复杂性. 在标准卷积层中,更多的组意味着更少的参数和更少的计算[5],然而,由于设备和内存缓存的不同,推理速度可能无法从小的理论复杂性改进中获益. 本文对GPU推理速度进行了测试.

表4和表5呈现了使用ResNet-18和VGG-Small的不同组的参数量、FLOPs及其对应速度. 当组的数量与分组规模$\gamma_g$无关时,组的数量越大复杂性越高,推理速度越慢;当组数与分组规模$\gamma_g$有关时,$\gamma_g$越大复杂性越高,推理速度越快.

**表 4** CIFAR-10上使用ResNet-18的复杂度和推理速度

| 超参数 | 参数量/M | FLOPs/M | 速度/FPS |
|---|---|---|---|
| $g=1$ | 11.32 | 11.33 | 43.10 |
| $g=2$ | 11.28 | 11.30 | 40.05 |
| $g=4$ | 11.26 | 11.28 | 40.60 |
| $g=\dfrac{C}{\gamma}$ | 11.25 | 11.26 | 40.19 |
| $g=\dfrac{C}{2\gamma}$ | 11.25 | 11.27 | 40.84 |
| $g=\dfrac{C}{4\gamma}$ | 11.26 | 11.28 | 41.50 |

**表 5** CIFAR-10上使用VGG-Small的复杂度和推理速度

| 超参数 | 参数量/M | FLOPs/M | 速度/FPS |
|---|---|---|---|
| $g=1$ | 4.71 | 5.13 | 134.09 |
| $g=2$ | 4.70 | 5.12 | 133.01 |
| $g=4$ | 4.69 | 5.11 | 131.43 |
| $g=\dfrac{C}{\gamma}$ | 4.68 | 5.10 | 127.33 |
| $g=\dfrac{C}{2\gamma}$ | 4.69 | 5.11 | 129.66 |
| $g=\dfrac{C}{4\gamma}$ | 4.69 | 5.11 | 135.27 |

对不同方法的ResNet18网络在两种数据集上进行了复杂度对比,表6展示了对比方法的参数量和推理速度结果. 由表6可见,SD-BNN方法和FDG-BNN方法为了对激活特征图进行分布调整,引入额外的参数量,其中FDG-BNN方法在CIFAR-10数据集和TinyImageNet数据集相比,IR-Net分别增加了0.08 M和0.06 M的参数量,对于当下绝大多数嵌入式和移动设备而言是可以忽略的,推理速度上比BNN-BN方法快大约5 FPS. 另一方面,FDG-BNN相比SD-BNN而言,在两个数据集上都减小了0.07 M的参数量,在CIFAR-10数据集的推理速度上慢0.23 FPS, TinyImageNet数据集上的推理速度基本持平. 因此,与SD-BNN相比,在牺牲较小推理速度的情况下,FDG-BNN实现了更小的参数量和准确率性能.

**表 6** 不同方法的ResNet18网络在CIFAR-10和TinyImageNet数据集上的复杂度对比

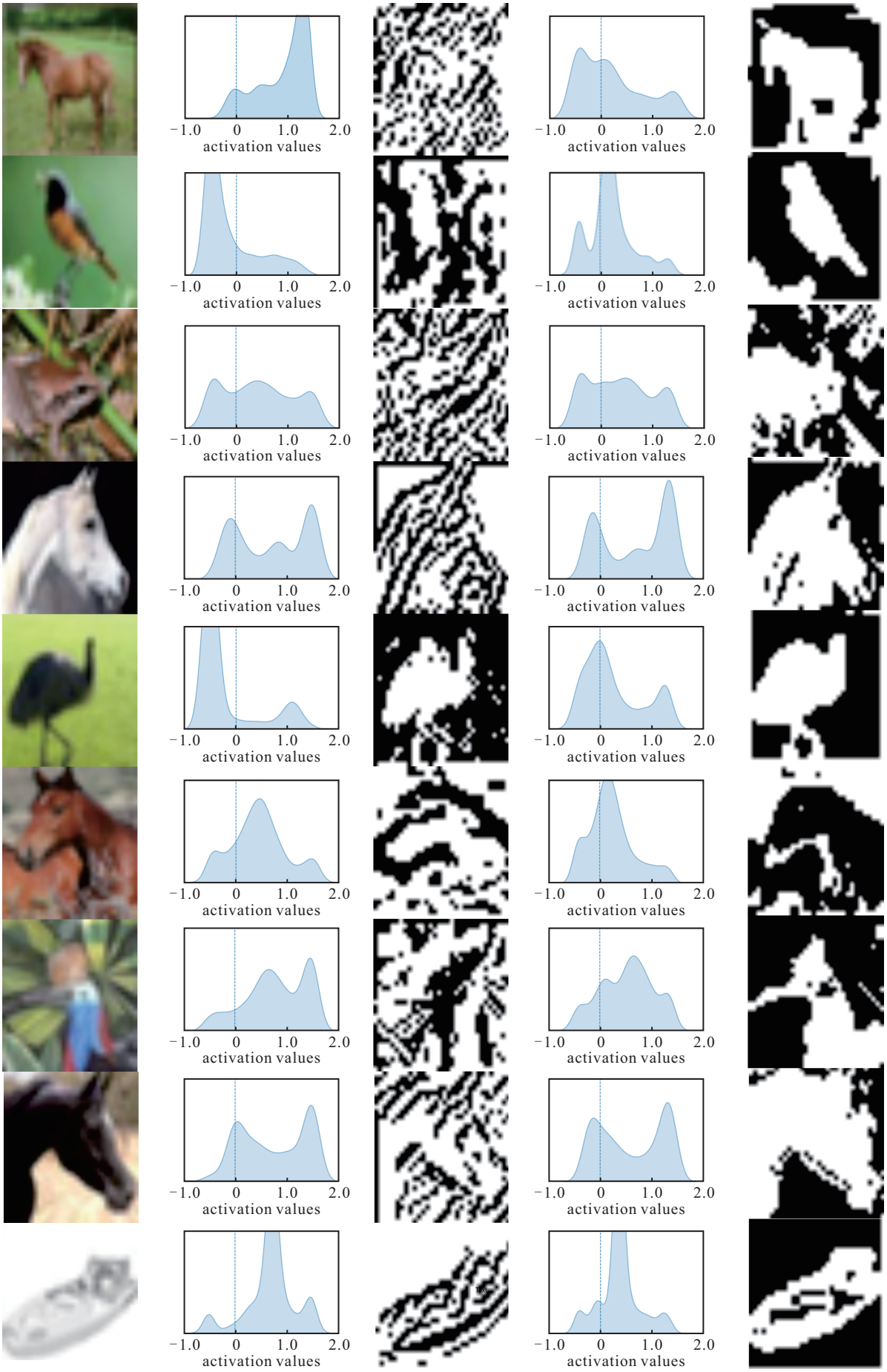
| 方法 | CIFAR-10 | | TinyImageNet | |
|---|---|---|---|---|
| | 参数量/M | 速度/FPS | 参数量/M | 速度/FPS |
| ReActNet | 11.20 | 20.39 | 11.29 | 20.28 |
| BNN-BN | 11.18 | 36.78 | 11.28 | 35.31 |
| RBNN | 11.17 | 20.05 | 11.18 | 22.64 |
| IR-Net | 11.17 | 35.36 | 11.28 | 35.50 |
| ReCU | 11.17 | 32.65 | 11.28 | 31.68 |
| SD-BNN | 11.32 | 40.42 | 11.41 | 40.21 |
| FDG-BNN | 11.25 | 40.19 | 11.34 | 40.21 |

5) 可视化分析.

图7给出了SD-BNN和FDG-BNN的可视化效果,(a)为CIFAR-10数据集随机选择图片,(b)、(c)为SD-BNN中特征分布与特征图,(d)、(e)为FDG-BNN中特征分布与特征图. 观察可知,从CIFAR-10数据集中随机选取不同图像输入到两者的ResNet-18 BNN中,其特征分布直方图显示了第1层卷积的激活分布,垂直线表示量化零点,注意到FDG-BNN比SD-BNN的特征分布更为均衡,且零点更接近于激活峰值. 在SD-BNN中,足够的鉴别语义信息几乎消失了,然而FDG-BNN保留了有用的语义特征,可清晰地观测到图片对象的主体轮廓.

## 4 结 论

本文提出了一种基于特征分布调整的二值量化方法,以解决二值量化中权重和激活的离散性和分布优化不当造成的准确率损失及语义信息消失的问题. 所提出的FDG-BNN针对激活权重分布进行调整,均衡其正负分布比例,进一步优化分布方式和阈值分割. 不同于当前主流先进方法针对全部特征进行统一调整,FDG-BNN使用分组激励与特征精调模块动态地调整激活分布的均值和方差,同时减小计算量. 此外,大量实验表明了所提出方法可调整输入特征图,更好地保留语义信息,在各种网络中均有较好的效果.




(a) 随机图片　(b) SD-BNN特征分布　(c) SD-BNN特征图　(d) FDG-BNN特征分布　(e) FDG-BNN特征图

**图 7　可视化效果对比**

## 参考文献(References)


[1] 郑香平, 梁循. 基于剪枝优化的深层胶囊网络[J]. 计算机学报, 2022, 45(7): 1557-1570.
(Zheng X P, Liang X. Deep capsule network based on pruning optimization[J]. Chinese Journal of Computers, 2022, 45(7): 1557-1570.)

[2] Liu Z, Wang Y, Han K, et al. Instance-aware dynamic neural network quantization[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE/CVF, 2022: 12434-12443.

[3] 潘瑞东, 孔维健, 齐洁. 基于预训练模型与知识蒸馏的法律判决预测算法[J]. 控制与决策, 2022, 37(1): 67-76.
(Pan R D, Kong W J, Qi J. Legal judgment prediction based on pre-training model and knowledge distillation[J]. Control and Decision, 2022, 37(1): 67-76.)

[4] 程旗, 李捷, 高晓利, 等. 基于深度稀疏低秩分解的深度神经网络轻量化方法[J]. 控制与决策, 2023, 38(3): 751-758.
(Cheng Q, Li J, Gao X L, et al. Lightweight method of deep neural network based on deep sparse low rank decomposition[J]. Control and Decision, 2023, 38(3): 751-758.)

[5] 余文勇, 张阳, 姚海明, 等. 基于轻量化重构网络的表面缺陷视觉检测[J]. 自动化学报, 2022, 48(9): 2175-2186.
(Yu W Y, Zhang Y, Yao H M, et al. Visual inspection of surface defects based on lightweight reconstruction network[J]. Acta Automatica Sinica, 2022, 48(9): 2175-2186.)

[6] Howard A G, Zhu M L, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J/OL]. 2017, arXiv: 1704.04861.

[7] Courbariaux M, Bengio Y, David J P. Binary connect: Training deep neural networks with binary weights during propagations[J/OL]. 2015, arXiv: 1511.00363.

[8] Rastegari M, Ordonez V, Redmon J, et al. XNOR-net: ImageNet classification using binary convolutional neural networks[M]. Computer Vision — ECCV 2016. Cham: Springer International Publishing, 2016: 525-542.

[9] Qin H T, Gong R H, Liu X L, et al. Forward and backward information retention for accurate binary neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 2250-2259.

[10] Xue P, Lu Y, Chang J F, et al. Self-distribution binary neural networks[J]. Applied Intelligence, 2022, 52(12): 13870-13882.

[11] Liu Z C, Shen Z Q, Savvides M, et al. Reactnet: Towards precise binary neural network with generalized activation functions[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 143-159.

[12] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7132-7141.

[13] Zhang D Q, Yang J L, Ye D, et al. LQ-nets: Learned quantization for highly accurate and compact deep neural networks[C]. Computer Vision — ECCV 2018. Cham: Springer International Publishing, 2018: 373-390.

[14] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[15] Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks[C]. Advances in Neural Information ProcessingSystems29. Barcelona: Curran Associates, 2016: 4107-4115.

[16] Ding R Z, Chin T W, Liu Z Y, et al. Regularizing activation distribution for training binarized deep networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 11400-11409.

[17] Kim H, Kim K, Kim J, et al. Binaryduo: Reducing gradient mismatch in binary activation network by coupling binary activations[C]. Proceedings of the International Conference on Learning Representations. Addis Ababa, 2020: 2118-2137.

[18] Xu Z H, Lin M B, Liu J Z, et al. ReCU: Reviving the dead weights in binary neural networks[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 5178-5188.

[19] Gong R H, Liu X L, Jiang S H, et al. Differentiable soft quantization: Bridging full-precision and low-bit neural networks[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 4851-4860.

[20] Xu Y X, Han K, Xu C, et al. Learning frequency domain approximation for binary neural networks[J/OL]. 2021, arXiv: 2103.00841.

[21] Rozen T, Kimhi M, Chmiel B, et al. Bimodal-distributed binarized neural networks[J]. Mathematics, 2022, 10(21): 4107-4121.

[22] Lin M, Ji R, Xu Z, et al. Rotated binary neural network[C]. Advances in Neural Information Processing Systems 33. Vancouver: Curran Associates, 2020: 7474-7485.

[23] Liu Z C, Wu B Y, Luo W H, et al. Bi-real net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm[J/OL]. 2018, arXiv: 1808.00278.

[24] Chen T L, Zhang Z Y, Ouyang X, et al. “BNN − BN =?”: Training binary neural networks without batch normalization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, 2021: 4614-4624.


## 作者简介


刘畅(1997−), 女, 硕士生, 从事模型压缩的研究, E-mail: 6201905006@stu.jiangnan.edu.cn;

陈莹(1976−), 女, 教授, 博士生导师, 从事计算机视觉与模式识别的研究, E-mail: chenying@jiangnan.edu.cn.