软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(10):4880−4892 [doi: 10.13328/j.cnki.jos.007363] [CSTR: 32375.14.jos.007363]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

用于二值神经网络的加宽和收缩机制*

韩 凯[1,2], 刘传建[3], 吴恩华[1,4]

[1](计算机科学国家重点实验室 (中国科学院 软件研究所), 北京 100190)

[2](中国科学院大学, 北京 100049)

[3](华为诺亚方舟实验室, 北京 100085)

[4](澳门大学 科技学院, 澳门 999078)

通信作者: 吴恩华, E-mail: ehwu@um.edu.mo

摘 要: 二值神经网络 (binary neural network, BNN) 因其较少的计算和存储开销而对业界非常有吸引力, 但其准确率仍然比全精度参数的网络差. 大多数现有方法旨在通过利用更有效的训练技术来提高二值神经网络的性能. 然而, 通过实验发现量化后特征的表示能力远弱于全精度的特征. 因此, 提出一种加宽和收缩机制来构建高精度而紧凑的二值神经网络. 首先, 通过将原始全精度网络中的特征投影到高维量化特征来解决量化特征表示能力弱的问题. 同时, 冗余的量化特征将被消除, 以避免某些特征维度的过度增长. 进而建立一个紧凑但具有足够表示能力的量化神经网络. 基准数据集上的实验结果表明, 该方法能够以更少的参数量和计算量建立高精度二值神经网络, 其准确率与全精度基线模型几乎相同, 例如, 二值量化的 ResNet-18 在 ImageNet 数据集上达到了 70% 的准确率.

关键词: 神经网络; 模型量化; 图像分类; 目标检测

中图法分类号: TP301

中文引用格式: 韩凯, 刘传建, 吴恩华. 用于二值神经网络的加宽和收缩机制. 软件学报, 2025, 36(10): 4880–4892. http://www.jos.org.cn/1000-9825/7363.htm

英文引用格式: Han K, Liu CJ, Wu EH. Widening and Squeezing Mechanism for Binary Neural Networks. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4880–4892 (in Chinese). http://www.jos.org.cn/1000-9825/7363.htm

Widening and Squeezing Mechanism for Binary Neural Networks

HAN Kai[1,2], LIU Chuan-Jian[3], WU En-Hua[1,4]

[1](State Key Laboratory of Computer Science (Institute of Software, Chinese Academy of Sciences), Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

[3](Huawei Noah's Ark Lab, Beijing 100085, China)

[4](Faculty of Science and Technology, University of Macau, Macao 999078, China)

Abstract: Binary neural networks (BNNs) are highly appealing to the industry due to their significantly reduced computation and storage requirements. However, their accuracy still lags behind that of networks with full-precision parameters. Most existing methods focus on improving the performance of BNNs through advanced training techniques. Empirical findings reveal that the representation capability of quantized features is considerably weaker than that of full-precision features. To address this limitation, a widening and squeezing mechanism is proposed to construct high-accuracy yet compact BNNs. Specifically, features from the original full-precision networks are projected into high-dimensional quantized features to mitigate the representation gap. Meanwhile, redundant quantized features are pruned to prevent the over growth of feature dimensions. As a result, a compact yet sufficiently expressive quantized neural network is constructed. Experimental results on benchmark datasets demonstrate that the proposed method achieves high-accuracy BNNs with significantly fewer parameters and computations while delivering performance comparable to full-precision baseline models. For instance,

* 基金项目: 国家自然科学基金 (62332015)
收稿时间: 2023-11-09; 修改时间: 2024-06-17, 2024-09-18; 采用时间: 2024-11-26; jos 在线出版时间: 2025-07-17
CNKI 网络首发时间: 2025-07-18

the binary ResNet-18 achieves a top-1 accuracy of 70% on the ImageNet dataset.

Key words: neural network; model quantization; image classification; object detection

深度神经网络, 尤其是卷积神经网络, 在各种计算机视觉应用中获得了优异的性能, 如图像分类[1−3]和目标检测[4,5]、语义分割[6,7]等任务. 部署在移动设备上的应用具有低时延、更好的隐私和离线操作等优势. 然而, 由于高内存和计算成本, 在资源受限的移动设备上部署深度学习模型具有很大挑战性. 出于这一需求, 研究人员提出了许多模型压缩和加速方法, 以提高学习的深度模型的适用性, 包括剪枝[8]、量化[9−11]、知识蒸馏[12]和轻量化模型[13,14]等方法.

模型量化[10,15]是最广泛使用的模型压缩方法之一. 为了降低深度神经网络的复杂性, 最近涌现了许多量化权重或激活值的工作. 其中, 权重和激活量化为+1 或−1 的二值神经网络[10]具有许多优点. 与非二值网络相比, 二值网络需要更小的内存, 并用逻辑运算如 XNOR 和 POPCOUNT 取代大多数浮点运算, 这提高了工作效率, 大大减少了推理时的内存大小和计算量.

尽管很多方法已经做出了诸多改进, 通过调整二值神经网络的结构 (如改变激活函数和批归一化的顺序[16]) 和添加更多的正则化[17]来提高二值神经网络的性能. 事实上, 大多数二值神经网络的算法都可以被视为二值化特征嵌入任务, 其中二值化特征和原始全精度特征的维数完全相同. 如果原始特征没有明显的冗余, 那么特征在二值空间中的表示能力肯定会低于特征在原始空间中的表示能力. 因此, 我们有必要将二值化卷积核的数量增加到合适的量, 以获得具有相同表示能力的二值化特征.

为此, 我们首先提供了两个实验来展示二值化特征的表示能力. 首先, 为了通过二值化特征来发现深度特征的内在表示, 我们使用正交变换将给定全精度神经网络中的特征投影到高维二值空间中, 以此保持住它们的成对欧几里得距离不变. 然后, 通过学习的选择掩码来识别原始特征中的冗余. 基于所获得的紧凑二值化特征, 我们以可接受的卷积核数量增量重新配置神经网络. 在基准数据集上的实验表明, 使用所提出的方法建立的二值神经网络能够实现与全精度基准模型相似的性能, 并且具有显著较低的内存占用和计算量.

1 模型压缩相关工作

本文通过使用更多的特征来解决量化网络的极限表示能力问题. 先前的研究主要集中在设计新的网络架构或新的量化函数以及找到量化值的最佳分布. 在本节中, 我们将回顾现有的建立紧凑模型的方法, 包括模型量化、剪枝和知识蒸馏.

1.1 模型量化

神经网络一般使用 FP32 的数值格式进行训练和存储, 我们称其为全精度神经网络. 模型量化将神经网络的权值或激活值量化为更低比特的离散值, 减少存储和计算量, 同时在量化过程中保持高精度 (达到或接近全精度神经网络在测评时的准确程度). 在 Deep compression 方法[18]中, 使用剪枝、量化和霍夫曼编码来压缩模型. DoReFa-Net[9]提出使用不同的位宽量化权重、激活和梯度. 梯度通过基于全精度权重的平均绝对值的自定义形式进行近似. BinaryConnect[19]将权值 W 替换为 $Sign(W)$, 直接优化网络的整体损失函数, 并在后向过程中用硬双曲正切函数逼近 *Sign* 函数来避免零梯度问题.

二值神经网络是一种极致的量化方法, 它将神经网络的权值和激活值都量化成单比特的数值, 从而将神经网络的计算转化为逻辑运算, 大幅降低存储和提升计算效率. BinaryNet[10]在二值化过程中为权重添加比例因子帮助优化. XNOR-Net[20]提出将实值缩放因子添加到二值化卷积的每个输出通道. Bi-Real Net[21]通过插入恒等映射方式将真实激活连接到二值卷积的输出上 (在量化函数之前), 以增强表示能力. ABC-Net[22]和二值集成方法[23]在每层使用更多的卷积运算来提高精度. 尽管这些工作已经取得了很大的进展, 但低比特量化网络, 特别是二值神经网络的性能仍然比全精度网络差得多.

1.2 模型剪枝和结构优化

网络修剪是一种压缩和加速神经网络的有效技术, 能够在存储和计算资源有限的硬件设备上更高效地部署网

络. 结构化剪枝方法[8,24,25]针对卷积核或整个层进行剪枝, 因此剪枝后的网络可以很容易地应用到实际场景. 例如, Liu 等人[26]利用尺度因子上的 L_1 正则化来选择通道. He 等人[27]利用基于几何中值的准则来剔除不重要的卷积核. 在本文中, 我们借鉴网络剪枝的思想来精简加宽后的二值神经网络, 以实现更低的内存和计算成本.

二值神经网络的结构和全精度网络有所不同, 因此针对二值网络设计更优的结构是一个重要方向. UBNAS[28]提出减少梯度误差并加速二值网络结构搜索的策略, 同时还探索了新的二值搜索空间. BATS[29]提出二值导向搜索空间和加速搜索过程的方法. 不过通过搜索获取二值网络结构的搜索过程需要大量的计算资源, 这在现实应用中不够实用.

1.3 知识蒸馏

知识蒸馏通过将大模型 (教师模型) 提供的知识迁移到小模型来提高小模型 (学生模型) 的性能. Hinton 等人[12]最早提出了知识蒸馏方法, 将较大模型的知识通过 KL 散度损失函数来指导小型神经网络. 从那时起, 知识蒸馏开始被广泛采用, 并涌现出许多新方法. 例如, Romero 等人[30]提出了 FitNet, 它提取了中间层的特征图以及最终输出, 以指导学生网络学习. 之后, Zagoruyko 等人[31]基于注意力图定义了注意力转移方法, 以提高学生网络的性能. 在本文中, 我们将剪枝、蒸馏与量化网络相结合. 使用网络剪枝方法来找到有效的量化网络. 然后, 我们使用知识蒸馏来提高紧致量化网络的精度.

2 基础知识

本文所提方法主要进行神经网络二值量化, 下面就量化基本知识予以介绍.

2.1 模型量化

二值神经网络作为一种极度极致的量化方式, 将神经网络内部的权值和激活值均量化至单比特的数值范畴, 由此促使神经网络的计算转化为逻辑运算形态, 大幅度削减了存储量, 并显著增强了计算效率. 这里我们以神经网络中的一个卷积层为例, 对权重和激活值均进行二值量化. 在不失一般性的前提下, 我们假设 W 为实数权重, A 为实数激活值, 原始全精度卷积的计算为 $W*A$. 除了二值化权值用 B_W 表示外, Q_W 为量化后的权值, 量化后的激活值用 Q_A 表示, 量化后卷积的计算为 Q_W*Q_A. 对于权值二值化, 采用 XNOR-Net[20]中常用的二值方法, 量化过程为 $B_W=Sign(W)$ 其中 $Sign(\cdot)$ 是符号函数, 反量化过程为 $W\approx\alpha B_W$ 其中 $\alpha=\overline{\|W\|_1}$ 是权重的绝对值均值. 该实现使用直通估计器 (STE)[32]反向传播梯度.

对于更高比特 (n-bit, $n>1$) 的量化, 权重量化的过程包含 3 个步骤.

1) $W=\dfrac{\tanh(W)}{2\cdot\max(|\tanh(W)|)}+0.5$, 权重被映射到 0 至 1 之间.

2) $W=\dfrac{round(W\cdot scale)}{scale}$, 其中 $scale=2^n-1$, 量化后的值取自范围 $\left\{0,\dfrac{1}{2^n-1},\dfrac{2}{2^n-1},\ldots,1\right\}$; 在该步骤中, 我们也使用 STE 进行梯度反向传播.

3) $Q_W=2\cdot W-1$, 量化值被映射到−1 至 1 之间.

对于激活值 X 的 n-bit 量化, 我们首先将 X 中的值截断到 0 至 1 之间, 然后使用 $\dfrac{round(X\cdot scale)}{scale}$ 来获得量化的值, 其中 $scale=2^n-1$. 这里反向传播时也使用了 STE 算法.

量化神经网络中典型的模块 (block) 结构如后文图 1 所示. 一般情况下, 第 1 层和最后一层不进行量化处理, 因为如果对输入图像或输出进行量化处理, 会造成过多的信息损失. 在训练过程中, 分别对批归一化层后的特征和卷积核权重进行量化. 然后它们被输入到卷积层. 训练后只有二值权值会被保存. 在进行推理时, 只需要对特征进行重新量化即可计算结果.

2.2 加宽和加深的优劣

为了验证加宽或加深二值神经网络的优劣性, 我们在 CIFAR-10 上对不同宽度和深度的二值化网络进行了实验. 我们选择全精度 ResNet-20 作为基线. 我们对二值化的 ResNet-20 进行 1、2、3、4、8 倍的加宽, 并选择二值

化的 ResNet-32、ResNet-56 和 ResNet-110 作为不同深度的网络.

精度结果展示在图 2 中. 可以看出, 参数较少的 2 倍宽的 ResNet-20 的性能优于 ResNet-110. 3 倍宽的二值化 ResNet-20 使用更少的参数, 实现了和全精度 ResNet-20 几乎相同的准确率. 随着每层通道的增加, 4 倍和 8 倍宽的模型准确率进一步提高. 其原因在于二值化特征的表示能力低于全精度特征, 因此我们希望使用更多的二值化特征来提高性能. 综合来说, 加宽的二值化网络能够比加深取得更好的效果, 因此接下来我们探索如何找到量化神经网络每一层最优的宽度.

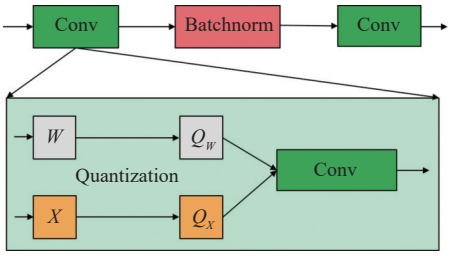

图 1 量化模块示意图

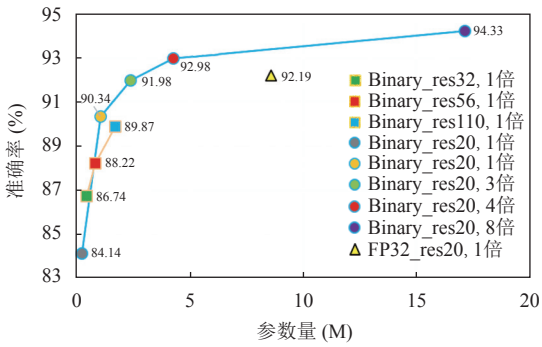

图 2 二值化网络加宽和加深的结果对比

在已有神经网络架构的基础上, 加宽或加深二值神经网络可以提升二值网络的精度. 不过, 加宽或加深也存在一定的局限性. 加宽或加深二值网络的精度有上限, 会被基础的网络架构所约束, 如果要突破上限就需要探索新的网络架构, 这是加宽或加深所达不到的.

3 基于加宽和收缩机制的量化神经网络

首先, 二值神经网络中往往需要更长的二值特征编码来保持来自原全精度输入空间的判别能力. 因此, 我们尝试找到二值神经网络中单层特征图的二值特征长度的下界. 接着, 我们提出一种端到端学习的网络宽度优化方法, 来找到量化网络中每一层最优的宽度.

3.1 逐层二值特征宽度求解方法

我们的目标是找到二值神经网络的最小维数 (即卷积核的数量), 以保持住原始全精度神经网络的性能. 对于预训练深度神经网络中的任意卷积层, 其卷积运算可以表示为:

$$X^{\top}F+b=Y \tag{1}$$

其中, $X\in\mathbb{R}^{wh\times ck^2}$ 是输入数据 (即前一层的特征或激活值) 根据本层的卷积核大小转换成的矩阵, $F\in\mathbb{R}^{ck^2\times n}$ 是 n 个卷积核权重, $Y\in\mathbb{R}^{wh\times n}$ 是输出特征图. w 和 h 分别是特征图的宽度和高度. n 是输出通道数, b 是偏置项, 为了简单起见, 通常会省略.

对于神经网络二值化问题, 我们将近似的二值特征图表示为 $\tilde{Y}=\phi(Y)\in\mathbb{R}^{wh\times m}$, 其中, m 是二值化层中卷积核的数量, $\phi(\cdot)$ 是线性[33]或者非线性变换[34,35]. 通常, 我们可以利用线性变换 P 来完成[33,36], 即 $\tilde{Y}=YP^{\top}$, 其中 $P\in\mathbb{R}^{m\times n}$. 因此, 对特征图 Y 的二值化可以表示为:

$$\min_{P,B}\frac{1}{2}\|YP^{\top}-B\|_F^2 \tag{2}$$

其中, $B\in\{-1,+1\}^{wh\times m}$ 是二值化后的特征图, $\|\cdot\|_F$ 是矩阵的 Frobenius 范数.

上述函数只强制给定卷积层中的特征是二值化的, 它没有继承在大量训练数据上学习到的特性. 因此, 我们建议保留每两个样本的特征之间的关系, 这通常是图像分类、检测和分割等视觉识别任务中的重要的特性, 即:

$$\min_{P,B}\frac{1}{2}\|YP^{\top}-B\|_F^2+\frac{\gamma}{2}\|\mathcal{D}(Y)-\mathcal{D}(B)\|_F^2 \tag{3}$$

其中, $\mathcal{D}(\cdot)$ 计算训练集中所有样本特征之间的欧氏距离, γ 是用于平衡上述函数中的两项的超参数. 由于训练集中的样本数量通常非常大 (如 ImageNet 中有百万量级的图片[37]), 所以 $\mathcal{D}(Y)$ 是一个非常大的矩阵 ($10^6\times10^6$), 这在实际操作中难以优化. 不过, 如果 P 是一个方阵且正交, 也就是 $P^{\top}P=I$, 任意两个样本特征之间的欧氏距离可以完全保持. 简单证明如下, 给定使用原始网络生成的任意两个样本 i 和 j 对应的特征 Y_i 和 Y_j, 我们有:

$$\begin{aligned}\|Y_iP^{\top}-Y_jP^{\top}\|_F^2&=\mathrm{Tr}(Y_iP^{\top}PY_i^{\top})-2\mathrm{Tr}(Y_iP^{\top}PY_j^{\top})+\mathrm{Tr}(Y_jP^{\top}PY_j^{\top})\\&=\mathrm{Tr}(Y_iY_i^{\top})-2\mathrm{Tr}(Y_iY_j^{\top})+\mathrm{Tr}(Y_jY_j^{\top})\\&=\|Y_i-Y_j\|_F^2\end{aligned}\tag{4}$$

进而我们可以把公式 (3) 转化为:

$$\min_{P,B}\frac{1}{2}\|YP^{\top}-B\|_F^2+\frac{\gamma}{2}\|P^{\top}P-I\|_F^2\tag{5}$$

这样, 我们可以利用公式 (5) 作为目标函数对给定的全精度网络进行二值化, 以保持其性能.

为了进一步探索二值化网络最高效的结构, 我们想找到特征图的二值化嵌入的下界, 这意味着二值化表示 B 的列稀疏性. 因此我们引入掩码 $M\in\{0,1\}^m$, 用来选择 B 中的特征. $M\circ B$ 表示 M 中的元素与 B 中的列的乘法. 我们在目标函数中添加掩码 M 的 L1 范数来求解其下界, 即:

$$\min_{P,B,M}\frac{1}{2}\|YP^{\top}-M\circ B\|_F^2+\frac{\gamma}{2}\|P^{\top}P-I\|_F^2+\beta\|M\|_1\tag{6}$$

我们使用交替迭代优化算法[38]来求解目标函数 (6), 通过循环迭代如下 3 步即可得到目标函数的解.

1) 求解 B. 由于二值化特征 B 中的元素是独立的, 可以简单地由如下等式得到:

$$B=Sign(YP^{\top})\tag{7}$$

其中, $Sign(\cdot)$ 是符号函数.

2) 求解 M. 对于固定的二值化变量 B 和投影矩阵 P, M 的优化目标可表示为:

$$\mathcal{L}(M)=\frac{1}{2}\|YP^{\top}-M\circ B\|_F^2+\alpha\|M\|_1\tag{8}$$

其目的是消除一些重构误差较大的列.

3) 求解 P. 根据解出的掩码 M 和二值化变量 B, 优化投影矩阵 P 的损失函数可写成:

$$\mathcal{L}(P)=\frac{1}{2}\|YP^{\top}-M\circ B\|_F^2+\frac{\gamma}{2}\|P^{\top}P-I\|_F^2\tag{9}$$

我们在 VGG-small 网络结构的实验表明, 上述算法能够得到二值神经网络每层所需通道数的下限, 并保持住原全精度网络的准确率. 虽然这种求解方法提供了一种寻找二值化特征数量的方法, 但其很难优化, 因为我们必须逐层优化所有卷积特征. 所以我们希望找到一种更简单而有效的方法.

3.2 端到端学习的网络宽度优化方法

根据上述分析, 我们希望使用尽可能少的量化特征来获得高精度的模型. 因此, 我们在加宽的量化网络基础上接着使用收缩方法来搜索有效的结构. 在第 2.2 节, 我们实验分析了二值特征的表示能力并不强, 最好的解决方案是加宽二值网络, 那么就出现了一个问题, 我们应该把网络加宽多少遍呢?

网络剪枝被广泛用于在低资源环境下降低深度神经网络模型的推理代价. 剪枝后的结构具有低计算量和高精度的特点. 此外, 剪枝方法可以看作是一种架构搜索范式, 用来寻找最优、最高效的架构. 为了利用网络剪枝的优势, 我们选择加宽 4 倍的量化网络为基础, 以获得更好的精度. 然后利用网络瘦身[26]方法对网络进行剪枝, 得到高效的模型. 具体地, 我们对批归一化层中的比例因子 γ 施加了稀疏性惩罚. 剪枝过程中的训练目标为:

$$\mathcal{L}_{\text{pruning}}=\mathcal{L}_0+\lambda\|\gamma\|_1\tag{10}$$

其中, $\mathcal{L}_0$ 为特定任务的原始损失函数, 例如分类任务的交叉熵损失和回归任务的均方误差 (MSE) 损失, λ 为稀疏正则化超参数. 在训练过程中, 不重要的通道被自动识别并去除, 产生具有相当精度的紧凑模型. 然后对压缩模型

进行再训练或微调以获得较高的精度.

知识蒸馏是一种提升模型精度的方法, 在这种方法中, 一个小模型 (学生模型) 被训练来模仿一个预训练的、更大的模型 (教师模型). 在蒸馏过程中, 通过最小化损失函数将知识从教师模型传递给学生模型, 损失函数的目标是教师模型预测的类别概率分布. 将教师模型和学生模型的 *Softmax* 之前的输出 (即 logits) 分别记为 o_{T} 和 o_{S}, *Softmax* 之后的输出表示为 $p_{\mathrm{T}}=Softmax(o_{\mathrm{T}})$ 和 $p_{\mathrm{S}}=Softmax(o_{\mathrm{S}})$. 知识蒸馏损失函数可表示为

$$\mathcal{L}_{\mathrm{KD}}=\mathcal{H}(y,p_{\mathrm{S}})+\mu\mathcal{H}(p_{\mathrm{T}}^{\tau},p_{\mathrm{S}}^{\tau}) \tag{11}$$

其中, $\mathcal{H}(\cdot,\cdot)$ 是交叉熵损失, y 是独热码标签向量, μ 是平衡这两个项的超参数. 此外, p_{T}^{τ} 和 p_{S}^{τ} 是教师模型和学生模型的软化后预测:

$$\begin{cases}p_{T}^{\tau}=Softmax\left(\dfrac{o_{\mathrm{T}}}{\tau}\right)\\ p_{S}^{\tau}=Softmax\left(\dfrac{o_{\mathrm{S}}}{\tau}\right)\end{cases} \tag{12}$$

其中, τ 为温度系数. 我们利用知识蒸馏来提高紧凑模型的精度. 教师模型可以是全精度模型, 也可以是加宽后的二值网络.

现有的神经网络要么是加宽来提升精度[39], 要么是收缩来减少参数量或计算量[8,24,25], 我们的方法首次提出通过先加宽再收缩的机制来优化二值神经网络结构. 在加宽和收缩机制的具体实现中, 首先我们提出逐层二值特征宽度求解方法, 然后引入端到端学习的网络宽度优化方法, 前者可以保证逐层的最优解, 后者可以更快速地求解整网的结构. 通过我们所提出的加宽和收缩机制得到的二值神经网络, 能够在保证精度的同时最小化参数量和计算量.

4 实验分析

在本节中, 我们将在多个图像分类基准数据集和一个目标检测任务上进行实验, 验证所提出的量化方法的有效性. 我们对实验结果进行详细的分析, 以进一步帮助了解所提出的方法的好处.

4.1 实验数据集

为了验证所提出的量化方法的有效性, 我们在多个基准视觉数据集上进行了实验, 包括 CIFAR-10[40]、CIFAR-100[40]、ImageNet ILSVRC 2012 数据集[37]和 PASCAL VOC0712 目标检测基准[41]. 表 1 给出了数据集所对应的详细信息.

表 1 实验数据集

类型	数据集	训练集数量	验证集数量	类别数
图像分类	CIFAR-10	60 000	10 000	10
	CIFAR-100	60 000	10 000	100
	ImageNet ILSVRC 2012	1.2M	50k	1 000
目标检测	PASCAL VOC0712	16 500	4 952	20

我们首先利用 CIFAR-10 数据集分析所提出方法的特性, 该数据集由属于 10 种类别的 60 000 彩色图像组成, 其中有 50 000 张训练图像和 10 000 张验证图像. CIFAR-100 数据集具有相同数量的图像和训练集验证集数量, 只是它有 100 类. CIFAR-10 和 CIFAR-100 采用了一种常用的数据增强方案, 包括随机裁剪和镜像. ImageNet 是一个大规模的图像数据集, 它包含了 1 000 类的 1.2M 张训练图像和 50k 张验证图像. 在训练过程中采用了常用的数据预处理策略, 包括随机裁剪和翻转. 我们还在 PASCAL VOC0712 数据集上进行了目标检测实验. 按照通常的做法, 我们在训练集 (16 500 张图像) 上训练模型, 并使用 4 952 张图像的验证集进行评估.

4.2 训练设定及基准模型

我们对每个 n 比特量化网络在{1, 2, 3, 4, 5, 8}范围内的一个或几个宽度上进行了加宽实验. 当宽度为 1 时, 量

化网络的宽度与标准网络相同. 对于所有基线实验, 我们将权重衰减设为 5E–5. 对于 CIFAR-10 和 CIFAR-100, 选择 ResNet-20[3]作为基准网络结构. 而 ResNet-18[3]和 VGG16 [2]网络结构用于测试 ImageNet 的性能. 在目标检测任务上, 我们采用以 VGG16 为骨干网络的 SSD[42]检测模型作为基础模型.

4.3 CIFAR-10 实验

(1) 逐层二值特征宽度求解方法的有效性

我们在 CIFAR-10 上使用 VGG-small 结构, 逐层优化每层卷积特征, 结果如表 2 所示. SGD 用于求解 P 和 M. 为了充分挖掘表征能力, 我们初始设置 m 为 n 的 8 倍. 第 3 行显示了优化后的通道数. 在较低的层中, 需要更多的特征, 而在较深的层中需要更少的特征. 然后对优化后的二值网络进行再训练, 得到 92.44% 精度. 这表明了我们的逐层二值特征宽度求解方法的有效性, 但是其求解过程较为复杂, 后续我们使用端到端学习的加宽和收缩机制进行网络优化.

表 2 CIFAR-10 数据集上逐层二值特征宽度求解方法的结果

方法	第2层	第3层	第4层	第5层	第6层	准确率 (%)
全精度网络	128	256	256	512	512	93.94
二值化网络 (逐层)	410	332	614	420	25	92.44

(2) 不同宽度的结果

我们在 1 比特和 4 比特量化网络上进行了实验. 在表 3 中, 基线网络 (n=32) 的 top-1 准确率为 92.19%. 对于 1 比特二值网络, 准确率随着网络宽度的增加而提高. 当宽度为 4 时, 二值网络的准确率超过基线. 对于 4 比特量化网络, 当宽度为 2 时, 准确率超过基线. 结果表明, 当量化网络的比特数较少时, 需要更多的特征来达到原始网络一样的准确率.

表 3 CIFAR-10 数据集上加宽不同宽度的结果

比特数	宽度	参数量 (M)	准确率 (%)
32	1	0.27	92.19
1	1	0.27	84.14
1	2	1.07	90.34
1	3	2.41	91.98
1	4	4.28	92.98
1	8	17.12	94.22
4	1	0.27	90.23
4	2	1.07	93.01
4	4	4.28	94.39

(3) 加宽和收缩机制的结果

虽然加宽可以让量化网络达到原始网络的准确率, 不过我们希望使用尽可能少的量化特征来获得更高的精度. 我们提出的加宽和收缩机制可以获得小而精确的量化网络. 图 3 为 CIFAR-10 数据集上, ResNet-20 不同正则化系数 λ 收缩后每层的通道数结果. 在本实验中, 我们使用二值化网络, 并设置网络宽度为 4, 并将通道去除的阈值设置为 0.01. 从表 4 和图 3 可以看出, 随着正则化系数 λ 的增大, 被去除的通道越多, 准确率就越低. 对于每个残差块, 有 2 个卷积层和 1 个残差连接. 第 1 个卷积层能够去除更多的通道, 原因是残差连接防止了过多的信息被丢弃.

由于正则化系数为 0.000 65 的结果与加宽 2 倍的量化网络具有相同数量的参数, 我们尝试用知识蒸馏的方法来提高其准确性. 教师模型分别选择全精度网络 (92.19% 的准确率) 和宽度为 8 的二值网络 (94.22% 的准确率). 我们在全连接层后使用 KD 知识蒸馏[12]. 结果如表 5 所示, 其中 τ 为温度系数, μ 为知识蒸馏与交叉熵的平衡系数. 可以看出宽度为 8 的二值网络作为教师的结果和全精度网络作为教师相比准确率更高. 通过知识蒸馏, 将收缩后的二值网络精度提高 1%–91.39%, 接近全精度的 92.19%. 因此, 通过使用 2 倍的参数数量, 使用我们方法的二值化 ResNet-20 的性能可以非常接近全精度 ResNet-20.

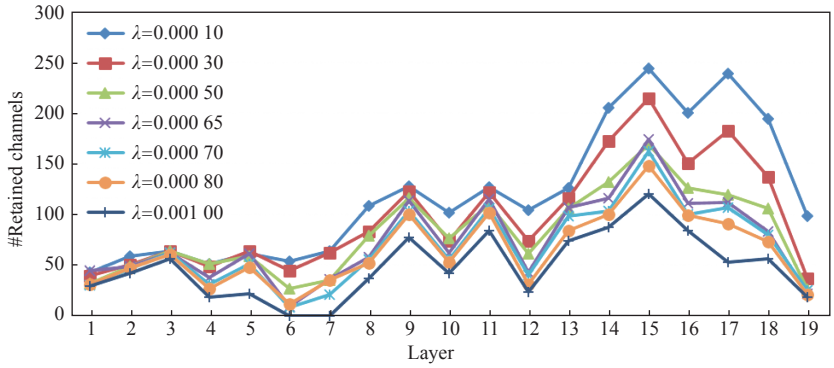

图 3 CIFAR-10 数据集上二值网络在不同正则化系数 λ 下的通道数结果

表 4 CIFAR-10 数据集上不同正则化系数 λ 的结果

正则化系数 λ	通道去除率 (%)	参数量 (M)	准确率 (%)
0.000 1	17.48	2.98	92.71
0.000 3	32.56	1.93	91.94
0.000 5	43.93	1.29	90.93
0.000 65	48.58	1.07	90.42
0.000 7	53.38	0.89	89.85
0.000 8	55.89	0.78	89.41
0.001	66.50	0.49	87.78

表 5 CIFAR-10 数据集上知识蒸馏后的结果

教师模型	τ	μ	准确率 (%)
全精度网络	3	0.2	90.72
全精度网络	5	0.3	91.0
全精度网络	10	0.2	91.12
宽度为8的二值网络	3	0.2	90.91
宽度为8的二值网络	5	0.3	91.22
宽度为8的二值网络	10	0.2	91.39

4.4 CIFAR-100 实验

(1) 不同宽度的结果

我们在 1 比特和 4 比特量化网络上进行了实验, 以 ResNet-20 为基础网络架构. 在表 6 中, 基线全精度网络的 top-1 的精度为 69.78%. 对于 1 比特二值网络, 精度随着网络宽度的增加而提高. 当宽度为 4 时, 二值网络的准确率 70.45% 超过了基线. 对于 4 比特量化网络, 当宽度为 2 时, 准确率为 70.25%, 也超过基线. 结果表明, 当量化网络的比特数较少时, 需要更多的特征.

表 6 CIFAR-100 数据集上加宽不同宽度的结果

比特数	宽度	参数量 (M)	准确率 (%)
32	1	0.28	69.78
1	1	0.28	50.44
1	2	1.08	62.62
1	3	2.43	67.61
1	4	4.31	70.45
1	8	17.17	74.68
4	1	0.28	63.35
4	2	1.08	70.25
4	4	4.31	73.85

(2) 加宽和收缩机制的结果

我们采用加宽和收缩机制对二值网络进行结构优化. 图 4 为 CIFAR-100 数据集上, ResNet-20 不同正则化系数 λ 收缩后每层的通道数结果. 图 5 和表 7 展示了我们的方法在不同正则化系数 λ 下得到的网络结构及其分类准确率. 与 CIFAR-10 相比, CIFAR-100 包含更多的类别. 因此, 在收缩机制过程中需要更多的参数和特征. 所有实验均将通道去除的阈值设置为 0.01. 当正则化系数 λ 为 0.000 5 时, 我们的方法得到的二值化 ResNet-20 的准确率为 67.98%, 比加宽 3 倍的二值化 ResNet-20 的准确率更高, 而且我们网络的参数更少.

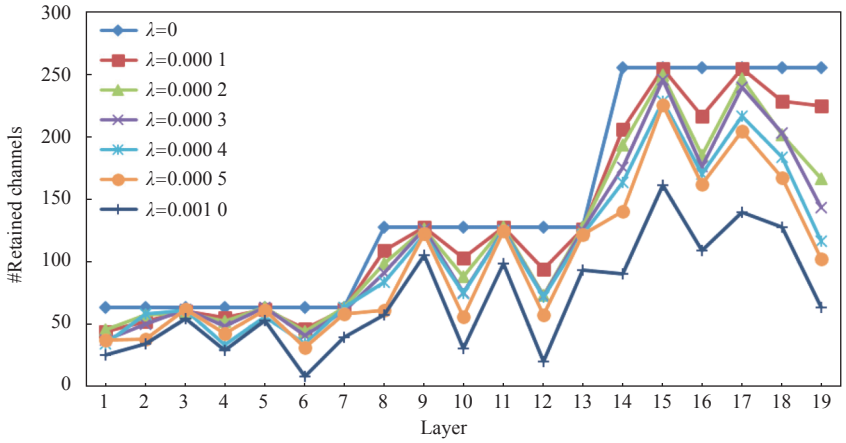

图 4　CIFAR-100 数据集上二值网络在不同正则化系数 λ 下的通道数结果

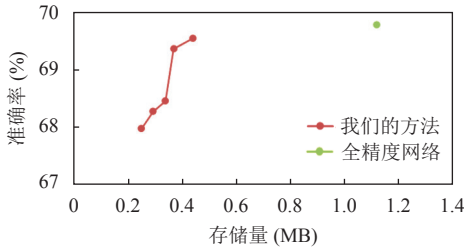

图 5　CIFAR-100 数据集上我们的方法得到的二值神经网络和全精度网络对比

表 7　CIFAR-100 数据集上不同正则化系数 λ 的结果

正则化系数 λ	通道去除率	参数量 (M)	准确率 (%)
0.000 1	10.32	3.54	69.54
0.000 2	17.11	2.98	69.36
0.000 3	20.97	2.73	68.46
0.000 4	26.31	2.37	68.28
0.000 5	31.36	2.03	67.98
0.001	50.84	0.99	62.07

我们利用知识蒸馏进一步提高了我们模型的性能. 宽度为 8 的二值化 ResNet-20 被用作教师网络, 其准确率为 74.68%. 温度 τ 和平衡系数分别设置为 3.0 和 0.8. 我们模型的准确率提高到 70.52%, 高于全精度基线的准确率 69.78%.

为了更直观地对比我们的方法和全精度网络的效果, 我们在图 5 中对比了准确率、存储开销的关系曲线. 计算开销和存储开销成正比, 这里不再列出. 从图中可以看出, 我们的方法只需要 40% 以内的存储量就能达到和全精度网络接近的准确率.

4.5　ImageNet 实验

(1) 不同宽度的结果

对于 ImageNet, 我们在二值化 ResNet-18 和 VGG16 网络上做了实验. 首先, 网络加宽后的结果如表 8 和表 9 所示. 对于 ResNet-18, 加宽 5 倍的二值网络的准确率超过了全精度网络. 对于 VGG16, 加宽 4 倍的二值网络的准确率与全精度网络相当. 这些实验也验证了更多的量化特征有利于量化网络性能的提高.

表 8　ImageNet 数据集上 ResNet-18 加宽不同宽度的结果

比特数	宽度	top-1准确率 (%)	top-5准确率 (%)
32	1	70.79	89.5
1	1	52.6	76.84
1	2	63.73	85.3
1	3	68.07	87.92
1	4	69.74	89.05
1	5	71.08	89.74

表 9　ImageNet 数据集上 VGG16 加宽不同宽度的结果

比特数	宽度	top-1准确率 (%)	top-5准确率 (%)
32	1	71.41	90.47
1	1	65.99	86.57
1	2	69.85	89.33
1	4	71.01	90.02

(2) 加宽和收缩机制的结果

为了得到高效准确的量化网络, 我们采用加宽和收缩机制对二值化 ResNet-18 网络进行结构优化, 并使用知识蒸馏提升准确率. 在这些实验中, 初始宽度设置为 5. 图 6 和表 10 给出了正则化系数 λ 下的结果. 当正则化系数 λ 为 0.000 5 时, 剩余的通道约为基线的 3.3 倍. 我们选用准确率为 70.79% 的全精度 ResNet-18 作为教师进行知识蒸馏. 温度 τ 和平衡系数分别设置为 3.0 和 0.3. 前 1 准确率提高到 70.04%, 与全精度基线网络相当.

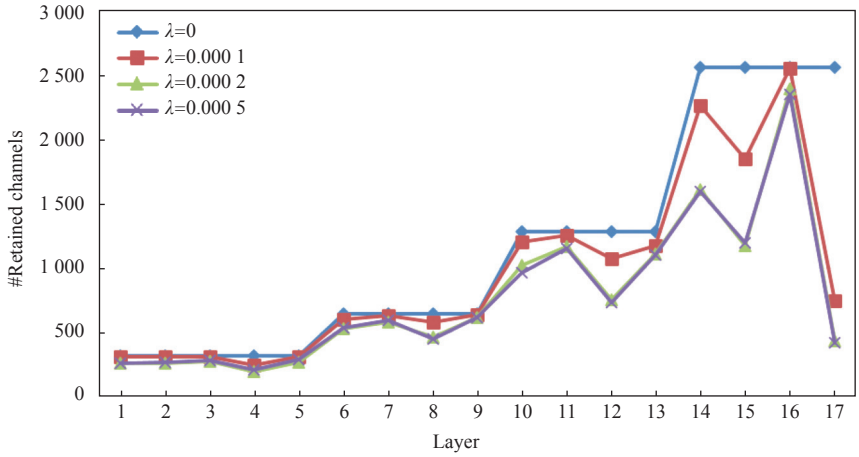

图 6 ImageNet 数据集上二值 ResNet-18 网络在不同正则化系数 λ 下的通道数结果

表 10 ImageNet 数据集上二值 ResNet-18 网络不同正则化系数 λ 的结果

正则化系数 λ	通道去除率 (%)	top-1准确率 (%)
0.000 1	17.95	68.32
0.000 2	33.06	69.19
0.000 5	33.44	69.18

4.6 目标检测实验

为了验证我们方法的泛化性, 我们进一步在目标检测任务进行了实验. SSD[42]是一种高效、高精度的单阶段目标检测方法. 我们使用二值神经网络代替骨干网络以及检测头. 结果如表 11 所示, 其中采用平均精度 (*mAP*) 作为评价指标. 我们采用全精度 VGG16 骨干网络的 SSD 作为基线, 数据集为 PASCAL VOC0712. 方法 A 和 C 仅将全精度 VGG16 骨干网络替换为宽度为 1 和宽度为 2 的二值 VGG16 骨干网络. 对于方法 B 和 D, 我们将除第 1 个输入卷积层和最后一个分类和定位层外的所有全精度卷积层替换为二值化卷积层. 从表 11 中看出, 即使只使用宽度为 1 的二值骨干网络的方法 A 也可以得到 68.74% 的 *mAP*. 对于 B, *mAP* 只减少 1.3%, 变为 67.44%. 使用宽度为 2 的二值化 VGG16 骨干网络的 SSD 获得 72.79% 的 *mAP*. 这些实验证明, 更多的量化特征仍然有助于提高目标检测任务的性能.

表 11 COCO 数据集上二值化 SSD 的结果

模型	*mAP* (%)
全精度SSD基线	76.81
A: 宽度为1的二值VGG16 & 全精度检测头	68.74
B: 宽度为1的二值VGG16 & 二值检测头	67.44
C: 宽度为2的二值VGG16 & 全精度检测头	72.79
D: 宽度为2的二值VGG16 & 二值检测头	71.75

4.7 和前沿方法的对比实验

我们的方法可以和不同的量化方法进行结合, 本节我们将我们的方法和 ReCU 二值量化方法[43]结合, 进一步提升二值神经网络的精度. 我们和近年来的前沿二值神经网络方法进行了对比 (如表 12 所示), 其中, 2024 年提出的 UBNAS[28]是当前二值神经网络结构优化的性能最好的方法. 根据已有前沿方法的模型规模, 我们将模型规模划分为 1M 左右参数量和 2M 以上参数量两种规模进行公平比较.

为了更直观地对比我们的方法和前沿方法的效果, 我们在图 7 中对比了准确率、存储开销的关系曲线. 这里存储量指的是模型参数使用 1 比特数值格式进行保存时所需要的字节数 (单个字节可以存储 8 个 1 比特的参数). 计算开销和存储开销成正比, 这里不再列出. 从图中可以看出, 我们的方法显著优于已有前沿方法, 在存储量相同的情况下, 准确率比已有方法高 1–3 个百分点.

表 12 CIFAR-10 数据集上和前沿方法的结果对比

规模	模型	参数量 (M)	准确率 (%)
1M左右参数量	RBCN[44]	0.59	85.5
	UBNAS[28]	0.76	88.5
	UBNAS[28]	1.2	89.6
	我们的方法	0.99	90.1
2M以上参数量	BNAS[45]	4.6	91.7
	IR-Net[46]	11.7	90.4
	BATS[29]	2.8	93.7
	UBNAS[28]	3.1	91.3
	我们的方法	2.02	93.8

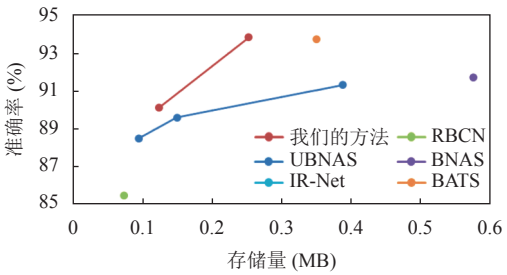

图 7 CIFAR-10 数据集上我们的方法得到的二值神经网络和全精度网络对比

5 总 结

为了提高二值神经网络的性能, 本文提出了一种有效的二值网络结构优化和训练方法. 全精度特征的量化会导致关键信息的丢失和精度的降低. 为了改善这一问题, 我们在加宽后网络的基础上采用网络收缩的方法来寻找高效、准确的量化模型. 此外, 还采用知识蒸馏的方法进一步提高了模型的性能. 在基准模型和数据集上进行的实验验证了该方法的有效性, 实验结果与全精度模型的性能相当. 此外, 该方法还可以与新的量化激活函数和网络架构等不同方法相结合, 得到更有效、更精确的量化网络.

References:

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2015.

[3] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[4] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[5] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

[6] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

[7] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[8] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient ConvNets. arXiv:1608.08710, 2017.

[9] Zhou SC, Wu YX, Ni ZK, Zhou XY, Wen H, Zou YH. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv:1606.06160, 2018.

[10] Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. Binarized neural networks. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 4114–4122.

[11] Gong C, Lu Y, Dai SR, Liu FX, Chen XW, Li T. Ultra-low loss quantization method for deep neural network compression. Ruan Jian Xue Bao/Journal of Software, 2021, 32(8): 2391–2407 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6189.htm [doi: 10.13328/j.cnki.jos.006189]

[12] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[13] Sandler M, Howard A, Zhu ML, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520. [doi: 10.1109/CVPR.2018.00474]

[14] Han K, Wang YH, Tian Q, Guo JY, Xu CJ, Xu C. GhostNet: More features from cheap operations. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586. [doi: 10.1109/CVPR42600.2020.00165]

[15] Hu J. Architecture optimization and quantization acceleration of convolutional neural networks [Ph.D. Thesis]. Beijing: Institute of Software, Chinese Academy of Sciences, 2023 (in Chinese with English abstract).

[16] Zhang DQ, Yang JL, Ye DQ, Hua G. LQ-Nets: Learned quantization for highly accurate and compact deep neural networks. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 373–390. [doi: 10.1007/978-3-030-01237-3_23]

[17] Lin MB, Ji RR, Xu ZH, Zhang BC, Wang Y, Wu YJ, Huang FY, Lin CW. Rotated binary neural network. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 7474–7485.

[18] Han S, Mao HZ, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv:1510.00149, 2016.

[19] Courbariaux M, Bengio Y, David JP. BinaryConnect: Training deep neural networks with binary weights during propagations. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 3123–3131.

[20] Rastegari M, Ordonez V, Redmon J, Farhadi A. XNOR-Net: ImageNet classification using binary convolutional neural networks. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 525–542. [doi: 10.1007/978-3-319-46493-0_32]

[21] Liu ZC, Wu BY, Luo WH, Yang X, Liu W, Cheng KT. Bi-Real net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 747–763. [doi: 10.1007/978-3-030-01267-0_44]

[22] Lin XF, Zhao C, Pan W. Towards accurate binary convolutional neural network. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 344–352.

[23] Zhu SL, Dong X, Su H. Binary ensemble neural network: More bits per network or more networks per bit? In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4918–4927. [doi: 10.1109/CVPR.2019.00506]

[24] He YH, Zhang XY, Sun J. Channel pruning for accelerating very deep neural networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1398–1406. [doi: 10.1109/ICCV.2017.155]

[25] He Y, Kang GL, Dong XY, Fu YW, Yang Y. Soft filter pruning for accelerating deep convolutional neural networks. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 2234–2240. [doi: 10.24963/ijcai.2018/309]

[26] Liu Z, Li JG, Shen ZQ, Huang G, Yan SM, Zhang CS. Learning efficient convolutional networks through network slimming. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2755–2763. [doi: 10.1109/ICCV.2017.298]

[27] He Y, Liu P, Wang ZW, Hu ZL, Yang Y. Filter pruning via geometric median for deep convolutional neural networks acceleration. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4335–4344. [doi: 10.1109/CVPR.2019.00447]

[28] Tan MH, Gao WF, Li H, Xie J, Gong MG. Universal binary neural networks design by improved differentiable neural architecture search. IEEE Trans. on Circuits and Systems for Video Technology, 2024, 34(10): 9153–9165. [doi: 10.1109/TCSVT.2024.3398691]

[29] Bulat A, Martinez B, Tzimiropoulos G. BATS: Binary architecture search. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 309–325. [doi: 10.1007/978-3-030-58592-1_19]

[30] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv:1412.6550, 2015.

[31] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928, 2017.

[32] Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv: 1308.3432, 2013.

[33] He XF, Niyogi P. Locality preserving projections. In: Proc. of the 17th Int'l Conf. on Neural Information Processing Systems. Whistler: MIT Press, 2004. 153–160.

[34] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323–2326. [doi: 10.1126/science.290.5500.2323]

[35] Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500): 2319–2323. [doi: 10.1126/science.290.5500.2319]

[36] Yu FX, Kumar S, Gong YC, Chang SF. Circulant binary embedding. In: Proc. of the 31st Int'l Conf. on Machine Learning. 2014. 946–954.

[37] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma SA, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

[38] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 2011, 3(1): 1–122. [doi: 10.1561/2200000016]

[39] McDonnell MD. Training wide residual networks for deployment using a single bit for each weight. arXiv:1802.08530, 2018.

[40] Krizhevsky A. Learning multiple layers of features from tiny images [MS. Thesis]. Toronto: Department of Computer Science, University of Toronto, 2009.

[41] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

[42] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot MultiBox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

[43] Xu ZH, Lin MB, Liu JZ, Chen J, Shao L, Gao Y, Tian YH, Ji RR. ReCU: Reviving the dead weights in binary neural networks. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 5178–5188. [doi: 10.1109/ICCV48922.2021.00515]

[44] Liu CL, Ding WR, Xia X, Hu Y, Zhang BC, Liu JZ, Zhuang BH, Guo GD. RBCN: Rectified binary convolutional networks for enhancing the performance of 1-bit DCNNs. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao, China: AAAI Press, 2019. 854–860.

[45] Chen HL, Zhuo L, Zhang BC, Zheng XW, Liu JZ, Ji RR, Doermann D, Guo GD. Binarized neural architecture search for efficient object recognition. Int'l Journal of Computer Vision, 2021, 129(2): 501–516. [doi: 10.1007/s11263-020-01379-y]

[46] Qin HT, Gong RH, Liu XL, Shen MZ, Wei ZR, Yu FW, Song JK. Forward and backward information retention for accurate binary neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 2247–2256. [doi: 10.1109/CVPR42600.2020.00232]

附中文参考文献:

[11] 龚成, 卢冶, 代素蓉, 刘方鑫, 陈新伟, 李涛. 一种超低损失的深度神经网络量化压缩方法. 软件学报, 2021, 32(8): 2391–2407. http://www.jos.org.cn/1000-9825/6189.htm [doi: 10.13328/j.cnki.jos.006189]

[15] 胡杰. 卷积神经网络的结构优化与量化加速研究 [博士学位论文]. 北京: 中国科学院软件研究所, 2023.

韩凯(1993－), 男, 博士生, CCF 学生会员, 主要研究领域为深度学习, 计算机视觉.

吴恩华(1947－), 男, 博士, 研究员, CCF 会士, 主要研究领域为计算机图形学, 虚拟现实, 机器学习.

刘传建(1986－), 男, 博士, 主要研究领域为深度学习, 计算机视觉.