航空学报
Acta Aeronautica et Astronautica Sinica
ISSN 1000-6893,CN 11-1929/V

《航空学报》网络首发论文

题目：　二值卷积神经网络综述
作者：　丁文锐，刘春蕾，李越，张宝昌
收稿日期：　2020-07-07
网络首发日期：　2020-09-29
引用格式：　丁文锐，刘春蕾，李越，张宝昌．二值卷积神经网络综述．航空学报.
https://kns.cnki.net/kcms/detail/11.1929.V.20200928.1729.010.html

航 空 学 报
Acta Aeronautica et Astronautica Sinica ISSN 1000-6893 CN 11-1929/V
http://hkxb.buaa.edu.cn hkxb@buaa.edu.cn

DOI:10.7527/S1000-6893.2020.24504

二值卷积神经网络综述

丁文锐[1,*]，刘春蕾[2]，李越[2]，张宝昌[3]

1. 北京航空航天大学 无人系统研究院，北京 100191

2. 北京航空航天大学 电子信息工程学院，北京 100191

3. 北京航空航天大学 自动化科学与电气工程学院，北京 100191

摘 要：二值卷积神经网络（Binary Neural Network，BNN）占用存储空间小、计算效率高，然而由于网络前向的二值量化与反向梯度的不匹配问题，使其与同结构的全精度深度卷积神经网络（Convolutional Neural Network，CNN）之间存在较大的性能差距，影响了其在资源受限平台上的部署。至今，研究者已提出了一系列网络设计与训练方法来降低卷积神经网络在二值化过程中的性能损失，以推动二值卷积神经网络在嵌入式便携设备发展中的应用。因此，本文对二值卷积神经网络进行综述，主要从提高网络表达能力与充分挖掘网络训练潜力两大方面，给出了当前二值卷积神经网络的发展脉络与研究现状。具体而言，提高网络表达能力分为二值化量化方法设计、结构设计两方面，充分挖掘网络训练潜力分为损失函数设计与训练策略两方面。最后，我们对二值卷积神经网络在不同任务与硬件平台的实验情况进行了总结和技术分析，并展望了未来研究中可能面临的挑战。

关键词：二值卷积神经网络；全精度卷积神经网络；二值化；量化；模型压缩；轻量化；深度学习

中图分类号：TP37 **文献标志码：**A

近年来，以深度卷积神经网络[37][47][48]为代表的人工智能技术得到了学术与工业领域的广泛关注，被视为一次具有重大意义的技术革新。目前深度卷积神经网络在多个领域，诸如计算机视觉、语音识别以及自然语言处理等，得到了大量的应用。当前，移动互联、物联网等技术与各个产业深度融合，以各类无人系统为平台载体的移动便携终端设备在识别应用的需求方面不断增加。然而，一方面，高性能的深度网络模型往往较大，这对将其装备到小内存的移动端无疑是一巨大挑战。另一方面，深度卷积神经网络的一大弊端即计算复杂度高，在运行较大的卷积网络模型时，为了实现网络中最常见的点积运算需要进行大量的计算。因此，复杂深度神经网络的劣势，即构成其的大量权重参数会导致相当大的存储空间、内存带宽以及计算资源上的消耗，使得在资源受限的移动端难以进行部署，从而其在实际使用中仍存在着很大的局限性。

基于以上问题，我们就需要对网络进行模型压缩以获得轻量化模型，使其可以更方便地部署到小内存的移动端设备上。现有研究对深度网络模型压缩已经做出较多综述性研究[51-56]。模型压缩[45][46]主要分为高效结构设计、模型剪枝、网络量化等方法。而本文则主要针对于网络量化中的极致量化，即二值卷积神经网络(Binary Neural Network，BNN，下文简称二值网络)，进行全面综述。所谓二值网络，其目标是将激活和权重同时量化为二值，二值化后的网络具有以下几个优

收稿日期：2020-07-07；退修日期：2020-08-03；录用日期：2020-09-07； 网络出版时间：
网络出版地址：
*通讯作者 E-mail: ding@buaa.edu.cn

引用格式：DING W, LIU C, LI Y, ZHANG B. Summary of Binary Convolutional Neural Network[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(4): 324504. 丁文锐，刘春蕾，李越，张宝昌. 二值卷积神经网络综述[J]. 航空学报, 2021, 42(4): 324504.

点：第一，内存更少。对于嵌入式移动设备来说，通常无法部署较大内存的网络模型。而网络量化减少了网络所需要的内存，使得网络模型更容易部署。第二，计算速度更高。当今典型的卷积神经网络模型在训练时通常需要较大的训练数据集和较多的迭代次数，巨大的计算量会导致较长的网络训练时间，而网络量化可以使得网络的计算成本相对减少，比如，模型当中的二值量化可以将浮点 32 位的数据转化为 1 位的数据，从而将浮点运算转化为位运算，使得计算速度大大提高。因此开展卷积神经网络二值化技术研究，不仅是对相关理论基础的进一步丰富和扩展，更对整个深度学习领域有重要的实际应用和理论价值，二值卷积神经网络应用优势示意图如图 1 所示。

近年来研究者们已经提出了一系列卷积神经网络二值化算法和训练方法来降低二值化过程中的性能损失。比如，XNOR-Net[1]通过重建具有单个比例因子的全精度滤波器，提供了卷积运算的有效实现。文献[3]提出了一种新的结构变体 Bi-Real Net，通过增加网络的便捷连接(shortcut)来大大增强网络的表达能力。文献[4]提出一种损失感知的方法，将二值量化损失考虑到端到端的网络中。文献[5]提出了使用残差网络进行二值量化，可在精确性和复杂性之间做出权衡。PCNN[2]通过离散反向传播对多个投影进行扩展来学习一组不同的量化核。然而现有技术研究虽然取得了较大的进展，但仍未解决二值网络与全精度网络之间巨大的性能差异问题。因此，二值卷积神经网络的更多潜力有待进一步被挖掘。

本文主要从提高网络的表达能力和充分挖掘网络的训练潜力方面出发，将现有研究二值卷积神经网络的方法进行全面综述。具体来讲，提高网络的表达能力可以从量化方法和结构设计两方面出发，充分挖掘网络训练潜力可以从损失函数设计、训练策略等角度出发。此外，我们还介绍了二值卷积神经网络在不同任务和硬件平台的发展情况，并总结了未来发展可能面临的挑战。本文章节组织如下：第一部分介绍了二值卷积神经网络的基础描述；第二部分从提高网络表达能力和充分挖掘网络训练潜力出发介绍现有的方法；第三部分介绍不同方法在不同任务和硬件平台中的性能以及分析；第四部分介绍影响及发展趋势，第五部分总结全文。

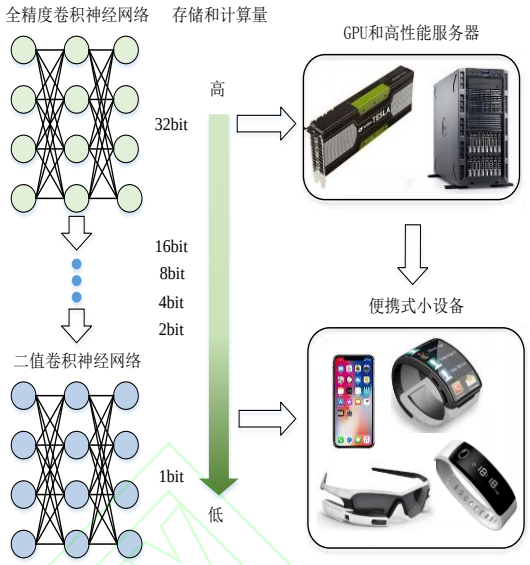

图 1 二值卷积神经网络应用优势示意图。相比于全精度卷积神经网络和其他 16bit、8bit、4bit、2bit 量化，1bit 量化具有最低的存储和计算量，从而使得配备在 GPU 和高性能服务器的全精度卷积神经网络模型能够在量化为 1bit 后配备在小型便携设备中。

1 二值卷积神经网络基础描述

在全精度卷积神经网络中，卷积基本运算可以表示为

$$X^{l+1} = X^l * W^l \tag{1}$$

其中 X^l 和 W^l 分别代表第 l 层的特征图与权重。$*$ 代表卷积算子。大量的浮点乘加运算造成了卷积神经网络推理过程中效率低下的问题。并且浮点型权重需要大量的存储空间。因此，我们希望采用二值化的方式来减少内存，降低计算复杂度。

1.1 前向过程二值化

二值卷积神经网络指的是具有二值权重和二值激活的深度网络模型，特别是通过 $\mathrm{sign}(\cdot)$ 函数来进行二值化，

$$\hat{x} = \mathrm{sign}(x) = \begin{cases} -1 & x > 0 \\ +1 & \text{otherwise} \end{cases} \tag{2}$$

$$\hat{w} = \mathrm{sign}(w) = \begin{cases} -1 & w > 0 \\ +1 & \text{otherwise} \end{cases} \tag{3}$$

其中 x 和 w 分别为矩阵 X 和 W 中的元素，$\hat{x}$ 和 $\hat{w}$ 为对应的二值化值。因此，公式(1)中的前向传播过程可以更改为

$$X^{l+1} = (\hat{X}^l * \hat{W}^l)\alpha^l \tag{4}$$

其中 α^l 代表 l 层的尺度因子。

1.2 反向传播

反向传播过程可以表示为：

$$\frac{\partial L}{\partial x}=\frac{\partial L}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial x},\qquad \frac{\partial L}{\partial w}=\frac{\partial L}{\partial \hat{w}}\frac{\partial \hat{w}}{\partial w} \tag{5}$$

其中 L 代表网络损失。由于反向传播过程中，量化器 $\mathrm{sign}(\cdot)$ 的导数为冲击波形式，零点处梯度无穷，非零点梯度处处为 0，即会在更新过程中造成梯度消失或梯度爆炸。因此，必须设计合适的梯度来代替 $\mathrm{sign}(\cdot)$ 原始的梯度来进行反向传播。现有研究中，研究者通常采用如图 2 所示的方波或者三角波的形式或采用其他近似 $\mathrm{sign}(\cdot)$ 函数的导数来代替反向传播中的量化器梯度。

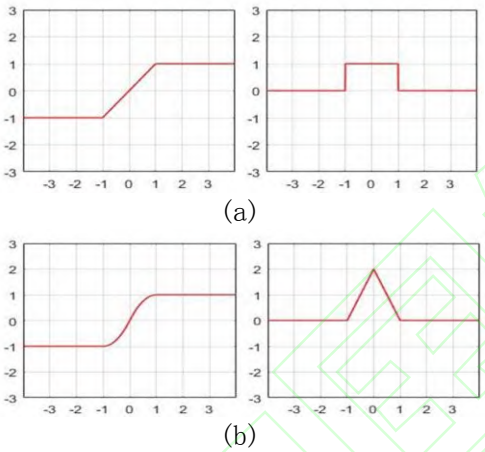

图 2 量化函数 sign 的梯度近似示意图。(a)代表方波近似，(b)代表三角波近似。其中第一列为近似 sign 函数，第二列为对应的近似梯度。

2 二值卷积神经网络方法概述

近年来，二值卷积神经网络领域得到颇多关注，催生了众多种类的二值卷积神经网络方法，从发展之初的使用预定义函数直接对权重和输入进行量化的朴素二值化方法，到目前使用基于多种角度和技术的基于优化的二值化方法，研究者们在二值网络领域已进行了诸多探索。在目前的基于优化的二值方法中，优化的角度与技术是多种多样的，其中包括了通过设计二值优化算法来近似全精度值、通过设计网络结构来增大网络的表达能力，与通过改进网络损失函数来限制权重等。然而，即使在二值网络中配置上述方法，二值卷积神经网络相比于其对应的全精度网络，仍然会产生相当大的精度损失，不利于其在很多具有高精度需求的设备上的应用。因此如何优化二值卷积神经网络，使其在具有节省资源优势的同时保持较高的网络精度，仍是一个具有挑战性的问题。本文认为，二值卷积神经网络性能损失的原因主要可归结为两点：其一是有限的表达能力，其二是不充分的训练。因此，基于这两点，本文从结构设计和量化方法两方面出发阐述当前提升二值网络表达能力的方法，从损失函数设计、训练策略等角度出发归纳挖掘网络训练潜力的方法。最后，再对针对其他任务平台所提出的方法进行简单介绍。

2.1 提高网络表达能力的二值卷积神经网络

2.1.1 二值优化方法

朴素的二值网络计算直接将激活函数和权重量化为 1 和-1[26]，这种映射方式虽然简单，但对全精度特征和权重所包含的丰富信息造成了巨大损失，极大降低了量化后网络的表达能力。作为考虑量化误差的早期研究，Rastegari 等人提出了 XNOR-Net[1]，将权重和激活都进行二值化。与先前的研究不同，该工作通过引入二值参数的比例因子对浮点数值进行更准确地近似。对于权重部分，比例因子的计算过程如下式所示

$$\begin{aligned} J(\hat{W},\alpha) &= \| W-\alpha\hat{W} \|^2 \\ \alpha^*,\hat{W}^* &= \arg\min_{\alpha,\hat{W}} J(\hat{W},\alpha) \end{aligned} \tag{6}$$

式中 α 即为权重部分的尺度因子，该尺度因子为逐通道级，通过最小化量化误差计算而得。尺度因子 α 的引入可实现二值参数与对应全精度参数的近似误差最小，进而对浮点数值进行更准确地近似。激活函数的量化与公式(6)类似。该方法可以大幅度降低由直接量化所带来的性能损失，并且首次在大型图像识别数据集 ImageNet 上进行了实验验证，为接下来的二值网络研究奠定了基础。

XNOR-Net 优化算法掀起了二值网络研究的热潮。为了进一步减少量化误差，高阶残差量化(HORQ)[5]采用基于量化残差的全精度激活的递归近似，而不是 XNOR-Net 中使用的一步近似。它通过递归地执行残差量化操作来获得递减尺度的二值激活，并通过对这些二值激活进行线性组合来得到最终的量化激活。高阶残差量化的引入在提升了网络表达能力的同时，不可避免的带来了计算量升高的问题。此后文献[6]引入了一种比例因子获取的新方式——数据驱动。其通过学习带有参数的门函数，来从未量化的激活中预测通道级的比例因子。该方法在增加不足 1%计算量的情况下大大提高了二值卷积神经网络的性能。

XNOR-Net[1]优化算法与 HORQ[5]算法均从网

络前向传播入手，然而反向传播算法对于二值卷积神经网络的训练亦是至关重要。在对二值网络进行反向传播部署时，由于 sign 函数的梯度不连续可导，通常采用直通估计器（Straight Through Estimator,STE）的方法来对梯度进行近似。但由于 sign 的实际梯度与 STE 之间存在明显的梯度不匹配，极易导致反向传播误差积累的问题，致使网络训练偏离正常的极值点，使得二值网络优化不足，从而严重降低性能。对于近似 sign 函数梯度的方波梯度，除了[-1，+1]范围之内的参数梯度不匹配外，还存在[-1，+1]范围之外的参数将不被更新的问题。直观来看，精心设计的二值化的近似函数可以缓解反向传播中的梯度失配问题。Bi-Real Net[3]提供了一个自定义的近似 sign 函数(ApproxSign)来替换 sign 函数以进行反向传播中的梯度计算，该梯度以三角波形式近似 sign 函数的梯度，相比于传统的 STE 其相对于冲击波的相似度更高，因而更贴近于 sign 函数梯度的计算。BNN+[7]直接提出用 swishsign 函数对 sign 函数进行近似来获取更优近似梯度。这些梯度近似方法能进一步对二值卷积神经网络的反向更新过程进行适度优化。

2.1.2 网络结构设计

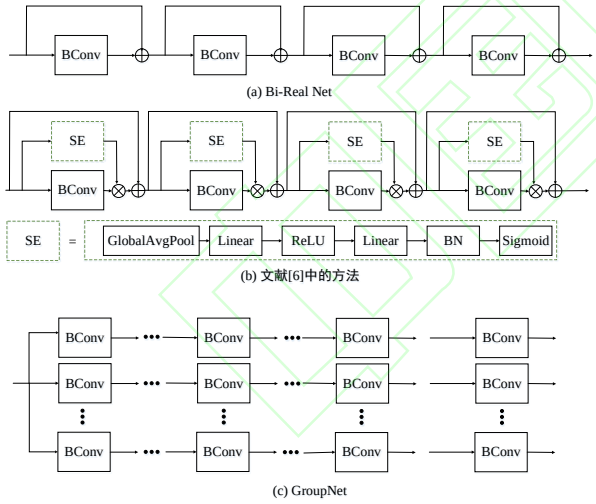

图 3 网络结构示意图，其中(a)(b)(c)分别表示 Bi-Real Net[3]，文献[6]种的方法和 GroupNet[9]的网络结构。

除了直接优化量化方法外，也有很多研究从网络结构设计方面出发，通过优化网络结构提升网络的表达能力。Liu 等人[8]从二值滤波器的重设计出发，设计了循环二值卷积神经网络（Circulant Binary Convolutional Network, CBCN），提出了循环滤波器和循环二值卷积，通过多角度旋转二值滤波器来增强二值化卷积特征的表达能力；与此同时，该循环滤波器也可提升网络的旋转不变性，提升二值网络对于旋转物体的识别鲁棒性。为优化该循环滤波器，该文还提出了相应的循环反向传播用以对网络进行训练。此外，为了提升网络的表达能力，Bi-Real Net[3]将每层卷积的输入特征图连接到后续网络中，这种方法实质上是通过结构转换来调整数据分布，对网络的提升效果十分明显。除此之外，Zhuang 等人[9]提出了组网络（GroupNet），即将网络分成若干组，在这些组中，通过聚集一组均匀的二元分支可以有效地重构全精度网络，该策略显示出对不同任务（包括分类和语义分割）的强大性能优势，在准确性和节省大量计算资源方面均有一定优越性。此外，文献[6]中比例因子获取的方式也可以看作对结构的改进，通过增加显著模块（Squeeze and Excitation，SE）来提高网络的表达能力。如图 3 所示，我们列出了 Bi-Real Net [3]，文献[6]和 GroupNet[9]的网络结构。相比于 XNOR-Net，以上几种方法在提升表达能力的同时，通常需要额外的存储或者计算量，但与全精度网络相比仍然具有较大的理论加速比。对于 Bi-Real Net 而言，由于只增加了 shortcut 的个数（加法计算量），该计算量相对于整体的 flops 而言是微乎其微的；对于 GroupNet 而言，其同时增加了存储和计算量，但是由于增加的是二值计算，相比于全精度而言，理论上仍然存在较大的加速比；对于文献[6]中的网络而言，其仅仅增加了不足 1%的计算量，但却取得了 ResNet18 在 ImageNet 上正确分类率 65.4%的结果。

2.2 提高网络训练潜力的二值卷积神经网络

Binaryduo[10]提出，二值卷积神经网络损失并非由于其表达能力有限，而是极度有限的两个状态使得模型难以被优化。因此，在提高模型表达能力的同时，仍然有一系列算法并行挖掘网络的训练潜力。本节从损失函数设计和训练策略的角度对这些训练方法进行归纳。

2.2.1 损失函数设计

从网络的训练方面来看，仅关注某一层很难保证经过一系列层后最终输出的精确性。因此，网络的训练需全局考虑二值化以及特定的任务目标。近来，大量研究工作对于网络训练中的损失函数进行探究，以期在二值化带来的限制下损失函数仍能较为准确地引导网络参数的学习过程。

通常来讲，一般的二值化方案仅关注对浮点数值的精确局部逼近，而忽略二值参数对全局损失的影响。Hou 等人提出了损失感知二值化方法

[5]，使用拟牛顿算法与对角哈希近似直接将与二值权重相关的总损失最小化，并求得了近端步骤的有效闭环解。该文证明了除了从量化角度考虑与任务相关的损失外，设计额外的量化感知损失项也是可行的。另一方面，激活的分布对于整个网络的优化也是至关重要的。Ding 等人[14]总结了二值卷积神经网络中由前向二值化和反向传播引起的问题，包括激活分布的“退化”，“饱和”和“梯度不匹配”。为解决这些问题，Ding 等人提出了一系列对于二值特征图的约束，联合任务中的损失函数共同指导网络训练，尽可能降低这些问题带来的不利影响。此外，全精度模型也可以损失函数的形式为二值网络提供引导信息，用以指导、优化二值网络的训练。Liu 等人[15]认为如果二值网络能学习到与全精度网络相似的分布，则其表现可获得一定程度的提升。因而提出了基于“分布损失”的方法，通过计算二值网络与全精度网络输出之间的 KL 散度来衡量二者分布间的差距，进而引导二值卷积神经网络去学习与全精度网络输出相似的分布。同样地，文献[6]从特征图级的约束出发，以全精度网络的特征图为引导信息，通过注意力匹配损失函数的引入来对二值网络的特征图与全精度特征图进行匹配，提升二值网络的训练潜能。另外，在卷积核级的约束上，Gu 等人[11]提出 Bayesian 损失，将全精度核和特征的先验分布合并到二值网络中，通过全精度卷积核的引导提升二值卷积核的表达能力。

总结来讲，利用构建损失函数来提高二值卷积网络的训练潜能，一般可以在不增加网络推理时间的同时提升网络性能，是当下较为流行的训练手段之一。

2.2.2 训练策略

由于二值网络所具有的高度离散性，其训练过程常需要引入特殊的训练方法，以使得训练过程稳定且获得更高的收敛精度。因此一类广为研究的方法，即对二值网络的训练方法进行重新设计，以构建出高效的二值网络。

传统的二值方法同时对激活函数和权重进行优化，文献[13]认为这对反向传播本就使用近似梯度的低 bit 网络来说是较为困难的，为此该文提出使用两阶段优化策略来逐步找到良好的局部最小值。具体来说，首先构建仅具有量化权重的网络并对此进行优化，然后将该优化所得的网络作为预训练模型，将其激活也进行量化，对权重与激活均量化的网络进行训练。与此同时，文献[13]还提出另一种渐进式优化方法，通过在训练过程中逐渐减少网络中数值表示的位宽，实现从高精度网络到低精度网络的逐渐转换。这种渐进式训练策略可以为低位宽模型提供适宜的初始化条件，有助于减轻低 bit 网络的训练难度。与逐渐降低网络位宽相似，文献[27]提出平滑渐进量化器，在训练过程中，将量化器逐渐地逼近 sign 函数，从量化函数的角度出发进行网络的渐进训练。

与此同时，模型蒸馏的方法被广泛应用在二值网络的训练之中。一般来讲，模型蒸馏的方法是通过大型教师模型提供引导信息，指导小型学生模型的训练。在二值网络的应用中，全精度网络或高精度网络一般被视作教师模型，二值网络或低精度网络一般被视作学生模型，以实现二值网络的引导性训练。文献[13][12]等多篇论文都提出以模型蒸馏的思想，通过高精度模型所生成的特征图对低精度模型的训练过程加以指导，从而使得低精度特征图接近于高精度特征图以获取更高的训练精度。Zhuang 等人[13]对于二值网络中量化函数不可微问题，提出一种基于全精度辅助模块辅助训练的方案。该方案通过在原始二值网络基础上添加全精度模块构建混合网络，并对这种混合网络进行优化，使得二值网络的更新能够同时借助全精度网络提供的信息。实验证明该方法可有效提升二值网络的训练性能。与纯粹的蒸馏思想不同，文献[16]提出了一种借助生成对抗模型来引导二值网络训练的方法。该文利用对抗学习中生成器与判别器相互对抗、共同获得性能提升的思想，将二值网络视作生成器，生成“假”特征图，由相应的全精度网络生成“真”特征图，通过引入一个判别器对真假特征图进行鉴别，使得二值网络生成特征图分布更接近于全精度特征图，从而提高二值网络性能。

另外，一种基于“耦合-分解”思想的训练策略在 BinaryDuo[10]中被提出。该文通过利用梯度失配估计器进行实验发现，对于二值网络中存在的梯度不匹配问题，采用更高的激活精度比修改激活函数的可微近似更为有效。基于该发现，Binaryduo 在训练过程中将两个二值激活耦合为三元激活，对该三元耦合网络进行优化，并将优化所得的网络解耦为二值网络，通过微调进一步提升网络性能。同时，文献[56]中提出了一种实现梯度量化的 DoReFa-Net。由于向前/向后遍历期间的卷积可以分别在低位宽权重和激活/梯度上运行，因此 DoReFa-Net 可以使用位卷积内核来加快训练和推理速度。

此外，在训练二值卷积神经网络时选择适当的超参数和特定的优化器有助于提高二值网络的性能，使二值网络的收敛更为迅速并最终达到

较高的收敛精度。大多数现有的二值卷积神经网络模型选择了自适应学习率优化器，例如 Adam 优化器，使用 Adam 优化器可以使训练过程收敛更快更好[38]。同时，批量归一化处理的设置对于网络的训练也很关键，在网络训练过程中通过批量归一化处理对网络的特征图分布进行调整，能够使网络训练的更加充分，有助于提升二值网络的整体性能。另外，一些研究者试图从信息论的角度对二值网络进行解释，并得出了相关的正则化训练技巧。Qin 等人[17]指出量化函数的使用使得二值网络的前向与反向传播中都不可避免的产生了信息损失。为了降低这种损失，Qin 等人从最大化信息熵的角度出发来最小化前向传播中的信息损失，通过简单地正则化操作使得二值网络的训练更为鲁棒。

2.3 面向其他应用的二值卷积神经网络

除图像分类任务之外，目标检测与语义分割也是视觉领域的常见任务之一，且相比于图像分类其复杂度更高、难度更大，其性能对于二值化的敏感度也更高。目前领域内存在少量研究，其二值网络是专门为这两种复杂任务而设计。

在目标检测方面，由于常规的网络二值化方法通常在具有受限表达能力的一级或两级检测器中直接量化权重和激活，这会造成信息冗余，从而导致大量误报并严重降低性能。针对这一问题，二值检测器 BiDet[18]提出充分利用二值卷积神经网络的表达能力，通过冗余消除进行目标检测，通过减少误报来提高检测精度。具体来说，将信息瓶颈(Information Bottleneck,IB)原理推广到目标检测领域，对高级特征图中的信息量进行限制，并且使得特征图与对象检测之间的互信息最大化。与此同时，BiDet 学习稀疏对象先验，以便后验者可专注于具有误报消除的信息检测预测。BiDet 是第一个提出将目标检测任务中的主干网络和检测网络同时量化的二值网络，然而其结果显示，二值化后的网络产生了较大性能损失，二值卷积神经网络在目标检测任务中仍然任重而道远。

对于语义分割任务而言，其对于网络特征图的多尺度信息要求更高。在这种要求下，GroupNet 提出二值并行卷积(Binary Parallel Atrous Convolution, BPAC)，该算法将丰富的多尺度上下文嵌入到 BNN 中以进行准确的语义分割。与仅使用 Groupnet 相比，具有 BPAC 的 Group-Net 可以在保持复杂度不变的情况下显著提高模型性能。

除此之外的大部分应用方法均是以分类为主，在其他应用上进行迁移和测试，很少有针对任务本身设计的二值化方法。

2.4 其他方法

近年来网络结构搜索（Neural Architecture Search, NAS）[40]在深度学习领域取得了令人振奋的成绩，这种方法通常自动设计针对于各种任务的最佳神经网络体系结构。二值网络结构搜索（Binary Neural Architecture Search, BNAS）[20]提出将通道采样和降低搜索空间引入到 NAS 中，以显著降低搜索成本，通过基于性能的策略来放弃有效性低的操作。Shen 等人[19]提出了一个用于自动搜索紧凑而准确的二值卷积神经网络的新框架。具体而言，基于该框架的二值网络将每层中的通道数编码到搜索空间中，并在训练过程中利用进化算法进行优化。实验表明，该方法搜索得到的二值卷积神经网络模型在模型大小和计算增量都可以接受的情况下，可以实现与全精度模型完全匹配的性能。

除 NAS 之外，还有一些研究从优化角度重新考虑二值网络的优化问题。文献[21]认为在二值网络中，不能仅将训练中的全精度权重直接类似于实值网络中的权重。相反，它们的主要作用是在训练过程中为二值权重的更新提供惯性。因此，文献[21]为二值网络的优化提供了新颖的见解，根据惯性来解释当前二值网络优化方法，并设计了一个专用于二值网络的优化器 Bop。根据将 Bop 用于二值网络优化其在 CIFAR-10 和 ImageNet 数据集上的性能表现，文献[20]很好地证明了该种二值网络优化新视角的可行性。而将训练中的全精度权重重新从惯性的角度加以定义以及引入 Bop 在一起，也可以帮助研究者们对二值网络优化有更深入地理解，并为进一步改进二值网络的训练方法开辟了新的道路。

此外，考虑到在二值网络中梯度下降法并不适用于量化函数，文献[22]提出可以将二值网络的优化看作一个离散的优化问题，为量化函数设置新的目标以最小化损失。对于离散优化问题，其目标是找到一组目标，以使每个单元（包括输出）都有线性可分离的问题要解决。给定这些目标，网络将分解为单独的感知器，因而可以使用标准凸方法进行学习。在此基础上，文献[22]开发了一种用于学习深阈值网络的递归微型批处理算法。该方法的提出为量化领域开辟了一个新的研究方向，并在分类数据集 ImageNet 上进行了验证。

3 实验

3.1 用于二值网络的实验数据集

二值卷积神经网络量化应用主要集中在目标分类任务上，同时在目标检测与语义分割任务上也有少部分工作。本章节我们将分别介绍不同应用中常用的数据集。

3.1.1 分类数据集

对于分类任务，常用的数据集主要包括 MNIST 手写字体数据集[34]、SVHN 数据集[39]、CIFAR10/100 数据集[35]以及 ImageNet 大规模图像数据集[36]。

MNIST：该数据集来自美国国家标准与技术研究所，由不同人所手写的数字图片构成。数据集包含 60,000 个用于训练的样本和 10,000 个用于测试的样本。这些样本均已经过尺寸标准化处理使数字位于图像中心。每个样本为大小固定的 28×28 像素，其像素值范围为 0~1。

SVHN：该数据来源于谷歌街景图像中门牌号码，每张图片中包含一组‘0~9’的阿拉伯数字。数据集分成了训练集、测试集与附加集 3 个子集。其中训练集中包含 73257 个数字，测试集中包含 26032 个数字，附加集有 531131 个数字。其图像为大小固定的 32×32 像素，像素值范围为 0 到 1。相比于同为数字识别数据集的 MNIST，SVHN 由于标记数据更多、来自自然场景因而识别难度更大。

CIFAR10/100：CIFAR10 与 CIFA100 均为彩色图片数据集。其中 CIFAR10 由包含 10 个类别的 60000 个彩色图像样本组成，并被分成了训练集与测试集两个子集。训练集和测试集分别包含 50000 张与 10000 张图像，每张图像分辨率为 32×32。该数据集覆盖了包括飞机、汽车、鸟类、猫、鹿、狗、青蛙、马、船和卡车在内的 10 个类别，类别之间完全互斥。与 CIFAR10 数据组成结构一致，CIFAR100 则包含具有更为细致分类的 100 个类别，每个类别包含了 600 个图像样本，为 500 个训练样本与 100 个测试样本的组合。由于需要进行更精细的识别，CIFAR100 的识别难度比 CIFAR10 更大。

ImageNet：ImageNet 是一个用于视觉对象识别研究的大型可视化数据库。相比于前面介绍的数据集，ImageNet 不管在图片数量还是图片分辨率上都有数量级上的提升。其由 1000 个类别组成，包括了约 120 万张训练图像和 5 万张验证图像。ImageNet 对深度学习的浪潮起了巨大的推动作用，也是当前神经网络量化在分类数据集上验证的最常用数据集。

表 1 不同新型的二值卷积神经网络在 ImageNet 分类数据集上的验证，FP(Full Precision)表示全精度网络。

		XNOR-Net	Bi-Real Net	XNOR-Net++	PCNN	BONN	IR-Net	CI-Net	FP
ResNet18	Top-1	51.2	56.4	57.1	57.3	59.3	58.1	59.9	69.3
ResNet18	Top-5	73.2	79.5	79.9	80.0	81.6	80.0	84.2	89.2
ResNet34	Top-1	-	62.2	-	-	-	62.9	64.9	73.3
ResNet34	Top-5	-	83.9	-	-	-	84.1	86.6	91.3
ResNet50	Top-1	63.1	62.6	-	-	-	-	-	74.7
ResNet50	Top-5	83.2	83.9	-	-	-	-	-	92.1
		BNN	ABC-Net	R-to-B (base)	R-to-B	CBCN	Group	BNN+	Bina-ryduo
ResNet18	Top-1	42.2	65.0	60.9	65.4	61.4	64.4	46.1	60.9
ResNet18	Top-5	69.2	85.9	83.0	86.2	82.8	85.6	75.7	82.6
ResNet34	Top-1	-	-	-	-	-	-	-	-
ResNet34	Top-5	-	-	-	-	-	-	-	-
ResNet50	Top-1	-	-	-	-	-	-	-	-
ResNet50	Top-5	-	-	-	-	-	-	-	-

3.1.2 目标检测与语义分割数据集

相比于图像分类数据集，目标检测与语义分割数据集由于标注工作量巨大，因而其建立过程更为复杂。目前领域内常用的检测、分割数据集有 PASCAL VOC2012[32]数据集与 COCO[33]数据集。

PASCAL VOC2012：该数据集源于 PASCAL 视觉目标检测比赛，用于评估计算机视觉领域中包括语义分割、目标检测等在内的多种任务上模型

的性能。该数据集包含人、常见动物、交通车辆、室内家具用品在内的 4 个大类并进一步细分为 20 小类。对于检测任务，VOC2012 包含了 11540 张图片在内的 27450 个物体，而对于分割任务，VOC2012 则包含了 2913 张图片在内的 6929 个物体。

COCO：该数据集是 Microsoft 团队提供的用于图像识别和目标检测的数据集，是一个大型、丰富的物体检测与分割数据集。该数据集以场景理解为目标，主要从复杂的日常场景截取图像。数据集由 80 个类别构成，涵括了超过 33 万张图片，其中 20 万张有标注，整个数据集中个体的数目超过了 150 万个。现有研究使用 COCO trainval 35k（115K 图像）进行分割训练，并使用 minival（5K 图像）进行分割验证。

3.2 图像分类实验结果

对于 MNIST、SVHN,由于其类别较小，数据量也较少，现有二值卷积神经网络在这些数据集上往往取得接近于全精度的结果，对于二值方法性能评估的意义不大。因此，近年来很少有工作报告该数据集上的测试精度。CIFAR10/100 介于 MNIST、SVHN 与 ImageNet 之间，但由于数据量有限，在 CIFAR 数据集上进行测试容易造成过拟合，但由于其相对难度和训练时间适中，也多为研究者所采用。

在本小节中，我们列出了近年来关于二值卷积神经网络比较经典和先进方法的结果，其所有数据都是直接引用对应原始论文中的结果。在此，我们选取了 XNOR-Net[1]、Bi-Real Net[3]、XNOR-Net++[23]、PCNN[2]、BONN[11]、IR-Net[17]、CI-Net[25]、BNN[26]、ABC-Net[24]、BNN+[7]、CBCN[8]、GroupNet[9]、文献[6]中的 baseline、文献[6]以及 Binaryduo[10]等方法来进行对比，用以显示当前二值网络在目标分类任务上的性能水平。

从表 1 的结果中，我们可以看出，由于训练时的 GPU 资源有限，目前大部分研究在展示基于 ImageNet 数据集的实验结果时，都选用了模型较小、对 GPU 资源需求较少的 ResNet18[37]结构。基于二值优化的 XNOR-Net 方法比朴素量化的 BNN 高出 9%（51.2%~42.2%），这也说明了其 XNOR-Net 方法中优化所得到的尺度因子在量化过程中大大提高了二值化模型的表达能力。此外，基于结构设计的方法，Bi-Real Net、CBCN、GroupNet、文献[6]中的方法进一步提高了二值网络的表达能力。通过一个简单的特征图短接的加法操作，Bi-Real Net 对二值卷积后进行了信息补偿，从而获得了超过 5%的性能提升。GroupNet 则采用牺牲计算量的方法来扩展二值分支，当扩展到 4 倍时，其 ResNet18 的 Top-1 性能高达 64.4%，这也说明了从结构设计角度对二值网络进行适当调整，可以大幅增加其表达能力。但是如何在结构设计中同时考虑效率与分类精度两个方面，以获得这两个度量上的平衡，是目前存在的一个关键问题。

关于训练策略而言，Binaryduo 通过对较高精度的三值网络的解耦，在不增加推理过程中存储和计算量的情况下大大提高了二值网络的性能。此外，文献[6]方法使用了模型蒸馏、渐进量化等多种训练策略，很大程度上挖掘了模型的训练潜力。其 ResNet18 结构在 ImageNet 分类数据集上可以达到 65.4%的精度，在增加不足 1%计算量的情况下，将其与全精度网络的性能差距缩小到了 3.9%。

3.3 目标检测和图像分割实验结果

文献[9]认为在目标检测任务中，对检测网络的主干部分和特征金字塔均进行二值化处理对性能的影响较小。但是，对于网络的其他部分，如检测头，当进行量化时情况并不乐观。实验表明对检测头进行二值化会导致十分明显的检测性能上的下降。这种下降可以根据检测网络的特性得到解释。一般认为，检测网络的检测头部分是负责将提取的多级特征适配到分类和回归目标，其表达能力对于检测器的性能至关重要。但是，当多级信息进行二值化操作而被强制约束为{-1，1}时，其信息会遭到破坏而影响检测性能。同时这也表明除主干外的其他检测模块对量化都很敏感，需要得到更多的研究以减轻其量化困难的问题，因而这也很可能是未来工作的一个有希望的方向。

此外，对于语义分割任务而言，文献[9]提出的基于 ResNet50 骨干网的 Group-Net 性能下降相对最大。它进一步表明，广泛使用的瓶颈结构对于二值网络并不友好。

3.4 其他应用实验结果

除了目标分类、目标检测和语义分割等一些主流应用验证，还有一些研究也在其他应用上进行过实验验证，比如文献[16]在目标跟踪任务中的 Got10k、OTB50、OTB100 及 UAV123 等数据集上进行验证，实验表明在跟踪任务的精度和成功率两大指标方面，二值网络与全精度网络仍然具有一定的差距。此外，Bi-Real Net 将二值模型应用在深度估计应用领域，其是自动驾驶和无

人导航的重要任务，压缩深度估计 CNN 对于将强大的 CNN 部署到内存和计算资源有限的移动设备至关重要。在深度估计任务上进行验证时，Bi-Real Net 采用了 KITTI 数据集[49]。结果表明，在深度估计实验中，Bi-Real Net 二值卷积神经网络几乎能达到和全精度网络相近的性能，仅仅产生了 0.3%的微量性能损失。另外，还有研究[28]在人脸识别、行人重识别、手势分类[44]等应用上进行过实验验证，实验结果显示直接迁移的二值网络在其他任务应用上还与全精度网络有一定的差距。

3.5 嵌入式设备应用实验结果

为了能够使二值网络得到实际中的应用，目前针对二值网络的嵌入式设备开发也有相应研究。Bi-Real Net 基于嵌入式应用的 Vivado 设计套件[50]估算了 18 层 Bi-Real Net 及其在现场可编程门阵列（Field-Programmable Gate Aray, FPGA）上的全精度网络的执行时间。与 FPGA 板上的全精度网络相比，二值卷积神经网络 Bi-Real Net 使用相同或更少的资源可实现执行速度上 6.07 倍的加速。与全精度卷积相比，二值卷积层的速度提高了 15.8 倍。通过累加所有操作的执行时间，最终经过测试，18 层 Bi-Real Net 比具有相同结构的全精度网络能够实现 7.38 倍的加速。

另外如文献[38]所述，在二值网络领域已经有一些推断框架，例如 BMXNet[41]、BitStream[42]、BitFlow [43]。其中，BitStream 和 BitFlow 只进行了论文发表，而没有建立源代码或二值库。BMXNet 虽然开源，但在 Google Pixel 1 手机上进行的测试显示，其运行速度甚至比全精度推断框架 TensorFlow Lite 还要慢。因此，为了填补二值网络推理框架缺失的空白，京东 AI 开源了一个针对 ARM 指令集高度优化的二值网络推理框架 dabnn[29]，这也是第一个高度优化的针对二值网络的开源推理框架。和 BMXNet 相比，dabnn 的速度得到了一个数量级的提升。一些研究者也实际部署与应用该推理框架，用以测试所设计二值网络的实际推理速度。如 IR-Net 便使用 dabnn 框架计算了算法部署到实际移动设备中时的效率。与现有的高性能推理（包括 NCN[30]和 DSQ[31]）比较的结果如表 2[17]所示，从中可以看出相比于全精度网络与其他低位宽网络，二值卷积神经网络 IR-Net 的推理速度要快得多。与此同时,IR-Net 的模型大小也可以大大减小，且在 IR-Net 中引入的移位尺度几乎不会带来额外的推理时间和存储消耗。

表 2 不同二值卷积神经网络在推理过程中的存储量和推理时间。FP 表示全精度网络。

	带宽	模型大小(Mb)	推理时间(ms)	精度（%）
FP	32/32	46.77	1418.94	69.3
NCNN	8/8	-	935.51	-
DSQ	2/2	-	551.22	65.17
IR-Net	1/1	4.21	261.98	58.1

3.6 二值化网络理论上的效率分析

表 3[38]显示了各种方案中 BNN 的参数大小和计算成本。计算成本（Floating Point Operations, FLOPs）是根据 Bi-Real Net 中的方法计算而得，即浮点数乘法的数量与 1/64 倍 1 位乘法的数量之和。从表中的结果可以看出，对于大部分基于训练方法或者优化近似的二值卷积神经网络而言，其并未改变其网络的基本结构配置，因此在推理过程中，具有相近的参数和浮点计算量。对于一些扩展结构的二值卷积神经网络而言，其参数和计算量会相对增加，以 GroupNet 为例，当扩展倍数为 1 时，GroupNet 的参数量和计算量与普通二值网络相同，而当相应扩展倍数时，其卷积层的计算量和参数量也会成倍增加。相似的，CBCN 由于保持可学习滤波器不变，旋转卷积核，所以其存储的参数量保持不变，但是推理时由于旋转卷积核的引入将需要更多的计算成本，实际上其与使用 4 位激活和 1 位权重的多位网络类似的参数和计算量等价。

表 3 各种新型二值卷积神经网络在 ImageNet 分类数据集上的参数量和 FLOPS

	参数量	FLOPS	精度（%）
XNOR-Net	33.3Mbit	164M	51.4
BNN+	33.3Mbit	164M	46.1
Bi-Real Net	33.3Mbit	164M	56.4
ABC-Net	33.3Mbit	164M	65.0
Group-Net (base=1)	33.3Mbit	164M	64.4
CBCN	33.3Mbit	656M	61.4
PCNN	33.7Mbit	164M	57.3
BONN	33.7Mbit	164M	59.3
Binaryduo	31.9Mbit	164M	60.9

4 影响及发展趋势

二值卷积神经网络旨在解决深度学习技术的

效率瓶颈，这将会对社会产生积极影响。特别是，由于小型智能设备广泛的商业价值和令人兴奋的前景，全世界将配备数十亿个小型，联网和智能设备。这些设备中的许多设备将嵌入我们的家庭，车辆，工厂和城市中。此外，可穿戴设备正变得越来越流行。低功耗计算设备的激增将推动工业领域乃至整个社区的发展，并在下一波个人计算浪潮中发挥至关重要的作用。重要的是，我们认为这些设备将在很大程度上依赖于现代深度学习，从而在感知和决策方面变得“智能”。通过将数据从移动设备上载到云来依靠云计算可能会遇到许多问题，由于延迟，隐私问题，更是难以实现，因而变得不可行。因此，迫切需要在移动设备本身上执行深度学习推理。但是，深度学习方法主要是为“重”平台（例如 GPU）设计的，而大多数移动设备都没有配备强大的 GPU，也没有足够的内存来运行和存储庞大的深度模型。解决这些瓶颈将使我们能够设计和实施有效的深度学习系统，这将帮助我们解决各种实际应用，例如具有高度隐私的个人计算。因此，二值卷积神经网络作为解决在提高深度学习推理效率以使其更具可扩展性和实用性方面的核心技术挑战，将会为整个社会带来巨大利益。

此外，目前二值网络应用主要面向目标分类，一些文章也专门设计了针对于语义分割和目标检测任务的二值网络。而对于视频、语音、其他时序信号方面，网络量化技术的发展和应用仍处于较为空白的阶段。而面对这些应用，若要取得较好的量化结果，必须充分考虑应用的特点，比如视频信号具有强帧间相关性特点。语音、通信信号为一维信号，量化后可能会显示出与二维图像信号完全不同的特点，要考虑其时域、频域特点，结合任务特征进行量化处理。

尽管二值卷积神经网络中现今已取得很大进展，但相对于全精度网络而言，仍然面临巨大的性能损失，特别是对于大型网络和数据集。 主要原因可能是包括：（1）尚不清楚哪种网络结构是适合二值化的，即未能总结出网络哪些组成即使进行二值化之后，也可以保留网络中较为充足的信息。（2）即使我们具有用于二值化的梯度估计器或近似函数，在离散空间中优化二进制网络也是一个难题。但是通过现有的研究明确了提高二值网络表达能力以及充分挖掘其训练潜力都将对提高二值网络性能产生积极影响。

5 结论

本文对二值卷积神经网络进行全面的综述，主要从提高网络表达能力与充分挖掘网络训练潜力的角度出发，给出当前二值卷积神经网络的发展脉络与研究现状。具体而言，从二值化量化方法设计和结构设计两方面进行提高网络表达能力方法的概述，从损失函数设计和训练策略两方面进行充分挖掘网络训练潜力方法的概述。最后，我们将二值卷积神经网络在不同任务与硬件平台的应用情况进行总结和讨论，并展望了未来研究可能面临的挑战。对于当前二值卷积神经网络的研究，本文总结如下：

1）二值卷积神经网络占用存储空间小、计算效率高，研究的主要挑战是其与全精度网络之间巨大的性能差异。

2）二值网络的研究，主要以提高网络的表达能力和挖掘网络的训练潜力为主。

3）针对于挖掘网络训练潜力的方法，在一般情况推理模式下模型存储和计算量都不发生改变，而模型性能提高。

4）部分对于以提高网络表达能力为主的方法会增加网络的存储或计算量，并容易获得显著的性能提升。

5）现在二值网络的应用主要以 ImageNet 数据集上的目标分类为主，少量研究也在目标检测、语义分割、目标跟踪、深度估计、人脸识别及行人重识别等应用上进行验证，表明二值网络的适用范围较广。

6）目前将二值网络配置在硬件设备上实现实际加速已有一些研究，主要集中在 ARM、FPGA 上，但开源研究仍较少，由于各种结构设计和辅助模块，在真实硬件设备上并不一定有较大的加速比，因此，实际硬件加速也将是二值网络研究的一个重点方向。

综上所述，二值卷积神经网络的研究将对未来嵌入式小型便携设备的发展产生不可忽略的作用，并可极大推动深度学习技术的发展和应用。

参 考 文 献

[1] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision, pp. 525-542, 2016.

[2] Jiaxin Gu, Ce Li, Baochang Zhang, Jungong Han, Xianbin Cao, Jianzhuang Liu, and David Doermann. Projection convolutional neural networks for 1-bit cnns via discrete back propagation. In Proceeding of the Conference of Association for the Advance of Artificial In-

telligence, volume 33, pp. 8344–8351, 2019.

[3] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In Proceedings of the European Conference on Computer Vision, pp. 722–737, 2018.

[4] Lu Hou, Quanming Yao, Kwok James T. Loss-aware Binarization of Deep Convolutional Networks. In Proceedings of the International Conference on Learning Representations, 2017.

[5] Li Z , Ni B , Zhang W , et al. Performance Guaranteed Network Acceleration via High-Order Residual Quantization. In Proceedings of IEEE International Conference on Computer Vision. IEEE Computer Society, 2017.

[6] B. Martinez, J. Yang, A. Bulat, G. Tzimiropoulos. Training binary neural networks with real-to-binary convolutions. In Proceedings of the International Conference on Learning Representations, 2020.

[7] S. Darabi, M. Belbahri, M. Courbariaux, V. P. Nia. Bnn+: Improved binary network training. NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2019.

[8] Chunlei Liu, Wenrui Ding, Xin Xia, Baochang Zhang, Jiaxin Gu, Jianzhuang Liu, Rongrong Ji, and David Doermann. Circulant binary convolutional networks: Enhancing the performance of 1-bit dcnns with circulant back propagation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2691–2699, 2019.

[9] B. Zhuang, C. Shen, M. Tan, L. Liu, I. Reid. Structured binary neural networks for accurate image classification and semantic segmentation. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–422, 2019.

[10] H. Kim, K. Kim, J. Kim, J.-J. Kim. Binaryduo: Reducing gradient mismatch in binary activation network by coupling binary activations. In Proceedings of the International Conference on Learning Representations, 2020.

[11] J. Gu, J. Zhao, X. Jiang, B. Zhang, J. Liu, G. Guo, R. Ji. Bayesian optimized 1-bit cnns, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4909–4917.

[12] Mishra, Asit, and Debbie Marr. Apprentice: Using Knowledge Distillation Techniques To Improve Low-Precision Network Accuracy. International Conference on Learning Representations, 2018.

[13] B. Zhuang, C. Shen, M. Tan, L. Liu, I. Reid. Towards effective low-bitwidth convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7920–7928, 2018.

[14] Ruizhou Ding, Ting-Wu Chin, Zeye Liu, and Diana Marculescu. Regularizing activation distribution for training binarized deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11408–11417, 2019.

[15] Z. Liu, Z. Shen, M. Savvides, K.-T. Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In Proceedings of the European Conference on Computer Vision, 2020.

[16] Liu C, Ding W, Xia X, Hu Y et al. RBCN: Rectified Binary Convolutional Networks for Enhancing the Performance of 1-bit DCNNs. In Proceeding of International Joint Conference on Artificial Intelligence, 2019.

[17] Qin, Haotong, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and Backward Information Retention for Accurate Binary Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2250-2259. 2020.

[18] Wang, Ziwei, et al. BiDet: An Efficient Binarized Object Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[19] Shen, Mingzhu, et al. Searching for accurate binary neural architectures. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[20] Chen, Hanlin, et al. Binarized neural architecture search. In Proceeding of the Conference of Association for the Advance of Artificial Intelligence, 2020.

[21] K. Helwegen, J. Widdicombe, L. Geiger, Z. Liu, K.-T. Cheng, R. Nusselder. Latent weights do not exist: Rethinking binarized neural network optimization. In proceeding of the Conference of Advances in Neural Information Processing Systems, pp. 7531–7542, 2019.

[22] Friesen, Abram L., and Pedro Domingos. Deep Learning as a Mixed Convex-Combinatorial Optimization Problem. In proceeding of the International Conference on Learning Representations, 2018.

[23] Bulat, Adrian, and Georgios Tzimiropoulos. XNOR-Net++: Improved binary neural networks. In proceeding of the British Machine Vision Conference, 2019.

[24] Lin, Xiaofan, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. In Proceeding of the Conference of Advances in Neural Information Processing Systems. 2017.

[25] Z. Wang, J. Lu, C. Tao, J. Zhou, Q. Tian. Learning channel-wise interactions for binary convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 568–577, 2019.

[26] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, Y. Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or-1, ArXiv Preprint ArXiv:1602.02830.

[27] Bulat, Adrian, et al. Improved training of binary networks for human pose estimation and image recognition. ArXiv Preprint ArXiv:1904.05868.

[28] C. Liu, W. Ding, Y.Hu, et al. Circulant Binary Convolutional Networks for Object Recognition. IEEE Journal of Selected Topics in Signal Processing, PP(99):1-1, 2020.

[29] J. Zhang, Y. Pan, T. Yao, H. Zhao, and T. Mei. dabnn: A super fast inference framework for binary neural networks on ARM devices. In Proceeding of ACM Multimedia Conference, 2019.

[30] nihui, BUG1989, Howave, gemfield, Corea, and eric612.ncnn. https://github.com/Tencent/ncnn.

[31] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[32] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2), 303-338, 2010.

[33] STsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, pp. 740–755, 2014.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[35] N. Krizhevsky and Hinton. The cifar-10 dataset. online: http://www. cs. toronto. edu/kriz/cifar. html.

[36] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[37] H. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[38] Qin, Haotong, et al. Binary neural networks: A survey. Pattern Recognition, 2020.

[39] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In Proceeding of Neural Information Processing Systems Workshop, 2011.

[40] Liu, Hanxiao, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. ArXiv Preprint ArXiv:1806.09055, 2018.

[41] H. Yang, M. Fritzsche, C. Bartz, C. Meinel. Bmxnet: an open-source binary neural network implementation based on mxnet. In Proceedings of the ACM Multimedia Conference, 2017.

[42] T. Zhao, X. He, J. Cheng, J. Hu. Bitstream: Efficient computing architecture for real-time low-power inference of binary neural networks on cpus, In Proceedings of the ACM Multimedia Conference, pp. 1545–1552, 2018,.

[43] Yuwei Hu, Jidong Zhai, Dinghua Li, Yifan Gong, Yuhao Zhu, Wei Liu, Lei Su, and Jiangming Jin. BitFlow: Exploiting Vector Parallelism for Binary Neural Networks on CPU. In Proceeding of IEEE International Parallel and Distributed Processing Symposium, pp. 244–253, 2018.

[44] 胡骏飞，文志强，谭海湖. 基于二值化卷积神经网络

的手势分类方法研究. 湖南工业大学学报, 031(1):75-80, 2017.

[45] 蔡瑞初, 钟椿荣, 余洋等. 面向"边缘"应用的卷积神经网络量化与压缩方. 计算机应用, 038(9):2449-2454, 2018.

[46] 袁庆祝. 基于CNN卷积神经网络的图像压缩技术[D]. 2019.

[47] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. Faster rcnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pp. 91-99. 2015.

[48] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848, 2017.

[49] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meetsrobotics: The kitti dataset. The International Journal of Robotics Research, 32(11):1231 – 1237, 2013.

[50] Tom Feist. Vivado design suite.White Paper, 5:30, 2012.

[51] 纪荣嵘, 林绍辉, 晁飞,等. 深度神经网络压缩与加速综述. 计算机研究与发展, 55(9):1871-1888, 2018.

[52] 雷杰, 高鑫, 宋杰,等. 深度网络模型压缩综述. 软件学报, 029(2):251-266, 2018.

[53] 李江昀等. 深度神经网络模型压缩综述. 工程科学学报 41.10: 1229-1239, 2019.

[54] 曹文龙, 芮建武, 李敏. 神经网络模型压缩方法综述. 计算机应用研究, 3 (2019): 3, 2019.

[55] 耿丽丽, 牛保宁. 深度神经网络模型压缩综述. 计算机科学与探索, (2020): 1-16, 2020.

[56] 张弛, 田锦, 王永森, 刘宏哲. 神经网络模型压缩方法综述. 中国计算机用户协会网络应用分会 2018 年第二十二届网络新技术与应用年会论文集, 2018.

[57] Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., & Zou, Y. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. ArXiv preprint ArXiv:1606.06160.

作者简介：

丁文锐 女，博士，研究员，博士生导师。主要研究方向：无人系统智能信息处理。

E-mail:ding@buaa.edu.cn

刘春蕾 女，博士研究生。主要研究方向：计算机视觉和模式识别。

E-mail: liuchunlei@buaa.edu.cn

李越 女，硕士研究生。主要研究方向：模型压缩，深度网络量化。

E-mail: ppy@buaa.edu.cn

张宝昌 男，博士，副教授。主要研究方向：计算机视觉和模式识别。

E-mail: bczhang@buaa.edu.cn

Summary of Binary Convolutional Neural Network

DING Wenrui [1,*], LIU Chunlei[2], LI Yue[2], ZHANG Baochang[3]
1. Unmanned System Research Institute, Beihang University, Beijing 100191, China
2. School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
3. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Abstract:
In recent years, binary neural networks (BNNs) have attracted much attention due to their low storage and high computational efficiency. However, the mismatch between forward and backward makes a huge performance gap between the BNN and the full-precision convolutional neural network, which affects BNN's deployment on resource-constrained platforms. So far, researchers have proposed a series of algorithms and training methods to reduce the performance gap during the binarization process, thereby promoting the application of BNNs to be deployed in embedded portable devices. Therefore, this paper makes a comprehensive review of BNNs, mainly from the perspective of improving network representative capabilities and fully exploring the network training potential. Specifically, improving network representative capabilities includes the design of binary quantization method and structure design, and fully exploring the network training potential includes loss function design and training strategy. Finally, we discuss the performance of BNNs in different tasks and hardware platforms, and summarize the challenges in future research.

Key words: binary convolutional neural network; full-precision convolutional neural network; binarization; quantization; model compression; lightweight; deep learning

[*] Received：2020-07-07；Revised：2020-08-03；Accepted：2020-09-07；Published online：
URL：
*Corresponding author. E-mail: ding@buaa.edu.cn