基于自适应梯度优化的二值神经网络

王子为[1,2],鲁继文[1,2],周 杰[1,2]
(1. 清华大学自动化系,北京 100084; 2. 北京信息科学与技术国家研究中心,北京 100084)

摘 要: 二值神经网络由于在储存空间和计算上的高效性,在视觉任务中被广泛运用. 为了训练不可导的二值网络,直通近似(Straight-Through Estimator)和 S 型近似(Sigmoid)等多种松弛优化方法被用来拟合量化函数. 但是,这些方法存在两个问题:(1)由于松弛函数和量化算子的差异导致的梯度失配;(2)由于激活值饱和引起的梯度消失. 量化函数自身的特性使二值网络梯度的准确性和有效性无法同时保证. 本文提出了基于自适应梯度优化的二值神经网络(Adaptive Gradient based Binary Neural Networks,AdaBNN),其通过自适应地寻找梯度准确性和有效性之间的最佳平衡来解决梯度失配和梯度消失的问题. 具体而言,本文从理论上证明了梯度准确性和有效性之间的矛盾,并通过比较松弛梯度的范数和松弛梯度与真实梯度之间的差距,构建了这一平衡的度量标准. 因此,二值神经网络能根据所提出的度量调整松弛函数,从而得到有效训练. 在 ImageNet 数据集上的实验表明,本文的方法相较于被广泛使用的 BNN 网络将 top-1 准确率提升了 17.1%.

关键词: 二值神经网络;梯度饱和;梯度失配;自适应梯度;图像分类
基金项目: 国家重点研发计划(No.2017YFA0700802);国家自然科学基金(No.62125603,No.61822603,No.U1813218,No.U1713214)
中图分类号: TP391.4;TP29 **文献标识码:** A **文章编号:** 0372-2112(2023)02-0257-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211084

Learning Adaptive Gradients for Binary Neural Networks

WANG Zi-wei[1,2], LU Ji-wen[1,2], ZHOU Jie[1,2]
(1. *Department of Automation*, *Tsinghua University*, *Beijing* 100084, *China*;
2. *Beijing National Research Center for Information Science and Technology*, *Beijing* 100084, *China*)

Abstract: Binary neural networks are widely employed in visual tasks due to the computation acceleration and storage shrinkage compared with the float counterparts. In order to train the non-differentiable networks, some continuous relaxation methods were proposed to approximate the quantizer including straight-through estimator (STE) and Sigmoid. However, these methods cause: (1) gradient mismatch due to the discrepancy between the quantizer and the relaxed function, (2) gradient vanishing due to the activation saturation. Because of the nature of quantization, the accuracy and validity of the gradient cannot be obtained for binary neural networks at the same time. In this paper, we propose AdaBNN that simultaneously solves the gradient mismatch and vanishing by adaptively achieving the optimal trade-off. Specifically, we theoretically prove the contradiction between gradient accuracy and validity, and formulate the evaluation measure for the trade-off by comparing the relaxed gradient norm and the discrepancy with true gradients. Therefore, the binary neural networks are trained effectively by changing the relaxation function based on the measure. Compared with the widely adopted BNN, experiments on ImageNet show that our method increases the top-1 classification accuracy by 17.1%.

Key words: binary neural networks; gradient saturation; gradient mismatch; adaptive gradients; image classification
Foundation Item(s): National Key Research and Development Program of China (No.2017YFA0700802); National Natural Science Foundation of China (No.62125603, No.61822603, No.U1813218, No.U1713214)

1 引言

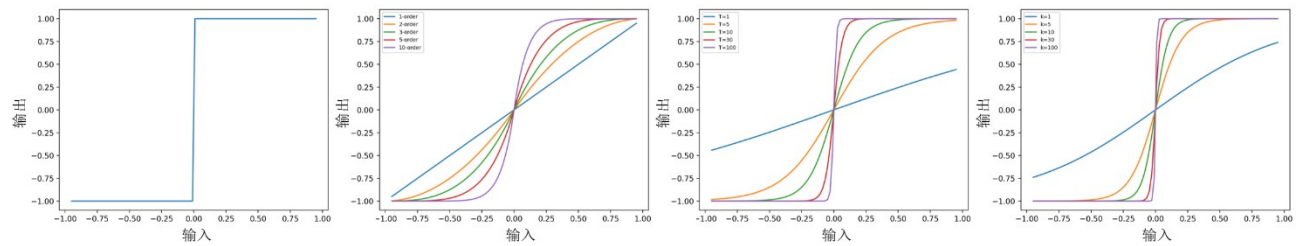
二值神经网络[1,2]被广泛运用于目标检测[3~6]和图像分类[7~11]等多项视觉任务中. 由于二值网络将传统实值网络中的浮点权重和激活值替换为二值,并将乘加操作替换成同或计数运算,所以神经网络存储空间和计算复杂度都显著下降. 然而,二值网络中不可导的

收稿日期:2021-08-13;修回日期:2021-09-28;责任编辑:王天慧

符号函数阻碍了梯度的反向传播. 因此,找到一个高效的替代优化方式来训练二值网络成为迫切的需求.

离散的搜索方法[12,13]尽管可以用于优化二值网络从而得到最优解,但是因其NP难带来的过高计算量,在实际应用中往往不可行. 一种简洁但有效的方法是直通近似(Straight-Through Estimator)[1],其用线性函数来近似符号函数. 虽然直通近似被证明是有效的[14],但其性能仍然因其梯度的失配而受到限制. 为了更精确地近似符号函数,S形函数(sigmoid)[15]、双曲正切(tanh)[16]和多项式函数[7]等可微函数也被用来替代符号函数. 但是这些函数在二值神经网络的训练过程中,由于激活值饱和从而无法提供有效的梯度,导致了梯度消失的问题. 为了平衡梯度的准确性和有效性,手动调整松弛函数陡峭程度的方法被广泛研究[15,16],但往往难以得到最优平衡.

本文提出了基于自适应梯度优化的二值神经网络,从而同时解决二值网络中梯度失配和梯度消失的问题. 不同于现有方法中直接用固定的松弛函数或者手动调整松弛函数的方法,本文根据所提出的评估度量动态自适应地平衡训练过程中梯度的准确性和有效性. 具体来说,本文在理论上证明了二值网络松弛梯度的准确性和有效性之间的矛盾,并使用松弛梯度的范数和松弛梯度与真实梯度之间的差距作为评价准确性和有效性平衡的度量. 因此,二值网络可以根据输入样本自适应选择最优的梯度松弛函数进行训练. 图1展示了在各种超参数设定下,不同阶松弛函数和符号函数之间的差距. 在CIFAR-10[17]和ImageNet[18]数据集上的实验表明,我们提出的方法在多种网络结构中大幅超过了现有二值网络训练方法.



(*a*) 符号函数 (*b*) 多项式函数 (*c*) S形函数 (*b*) 双曲正切函数

注: 松弛函数不同的超参数影响着其对符号函数的近似程度. 平缓的松弛函数会导致梯度失配,而陡峭的松弛函数会因为输入饱和导致梯度消失. 传统方法使用固定的松弛函数或手动调整松弛函数的陡度,导致训练过程中梯度准确性和有效性之间的平衡只能获得次优解. 而我们提出的基于自适应梯度优化二值神经网络可以在训练阶段根据评估度量,动态自适应地寻找最优松弛函数.

图1 符号函数、多项式函数、S形函数和双曲正切函数之间的比较

2 相关工作

在本节中,我们主要回顾在量化策略设计和量化网络优化领域的研究方法.

2.1 量化策略设计

网络量化[19~25]由于对实值网络计算和存储效率的优化,在计算机视觉领域引起了高度关注. 现有的量化网络分为两大类:权重量化网络的和权重激活量化网络. 前者将实值权重量化成低比特权重以节约存储空间,网络中的乘加运算也被替换成加法运算来节省计算量. Courbariaux等人[27]使用符号函数量化了浮点数权重,并在小数据集上取得了不错的性能. Rastegari等人[2]将尺度因子加入量化后的权重以减少信息损失. Zhang等人[30]通过训练与位间操作相容的自适应量化器来最小化量化损失. 因为实验表明更大的权重位宽可以使权重量化网络取得更好的性能[26],因此研究者提出了三值[26]和多比特[29~32]的权重量化方案. 然而,实值激活值因为需要累加操作的存在而无法进一步降低计算量. 对于权重激活量化网络而言,实值网络的乘加运算被替换为异或计数操作,从而计算量大幅度降低. Hubara等人[1]和Rastegari等人[2]通过符号函数量化了权重和激活从而获得了可观的加速. Lin等人[33]为网络权重和激活的量化设计了多组基,并在大规模数据集上很显著地提升了性能. Liu等人[7]在相邻的卷积模块间加入了额外的短连以提升二值网络的表示能力. Bethge等人[34]为了增强二值网络的信息流动,提出了更稠密的连接和更宽的网络结构,并且达到了目前较优的性能. 然而,量化网络中的不可导问题始终阻碍了训练过程中的梯度反传. 因此,有效地优化二值网络是获得高性能紧致网络的必要环节.

2.2 量化网络优化

由于量化网络中存在不可微的符号函数,使用近似方法替代神经网络中传统的梯度反向传播算法最近被广泛研究[35~38]. 现有的优化方法可以分为两类:离散优化和可微松弛. 离散优化方案使用有效的搜索策略以获得最优的量化权重. Hou等人[12]使用近端牛顿算法,使用对角黑塞矩阵估计的方式来直接最小化由二

值权重得到的损失函数. Leng等人[13]将连续参数从离散约束中解耦,并使用交替方向乘子算法(Alternating Direction Method of Multipliers)[39]迭代式解决子问题. 然而,由于离散搜索是NP难问题,其训练成本在实际应用中无法负担. 对于可微松弛,线性和非线性函数都被用于近似符号函数. 直通估计器[1, 40]因其简洁性和高效性,被广泛地用于线性松弛. 直通估计器假设$\mathrm{sgn}(x)=x$,其中sgn和x分别表示符号函数和输入. 由于符号函数与线性函数之间较大的差别,梯度失配影响了量化神经网络的性能. 诸如S形函数[15]、双曲正切函数[16]和多项式函数[7]等非线性函数缓解了梯度失配的问题,但会导致量化网络的梯度消失和收敛困难. 而通过手动选择的方式调整松弛函数[15,16]会导致梯度准确性和有效性的平衡陷入局部最优.

3　方法

本节中,我们先从理论上证明二值网络中梯度准确性和有效性的矛盾,然后提出基于自适应梯度优化二值神经网络,从而在训练过程中根据所提出的评估度量来动态自适应调整梯度准确性和有效性.

3.1　梯度下降的收敛性

梯度下降方法旨在最小化训练样本中的经验损失函数. 由于无法得到解析解,梯度下降方法通常被用来迭代更新神经网络的权重,从而得到最优解. 步长为η时的更新公式为

$$x_{t+1}=x_t-\eta\nabla f(x) \tag{1}$$

我们要求神经网络的优化过程是收敛的,收敛的定义如下:

定义1(收敛序列的定义)　如果满足

$$\lim_{t\to\infty}\frac{x_{t+1}-x^*}{x_t-x^*}=\mu \tag{2}$$

则序列$\{x_t\}$收敛到x^*,$\mu\in(0,1)$为收敛率.

由于我们用梯度下降的方式优化二值神经网络,因此我们假设符号函数是一个在自变量为0附近很陡峭的可导函数,之后我们根据该假设代替符号函数进行讨论. 全局损失函数f使得神经拟合数据的分布,其满足利普希茨连续条件. 进一步而言,f关于权重的梯度也满足具有利普希茨常数β的利普希茨连续条件,即f被称为具有β-平滑的性质.

定义2　若$f(x)$的梯度满足利普希茨常数为β的利普希茨连续条件,即满足

$$\left\|\nabla f(x)-\nabla f(y)\right\|\leqslant\beta\left\|x-y\right\| \tag{3}$$

则称它为β-平滑.

β-平滑函数梯度的差距被自变量的差距和利普希茨常数限制在有界范围. 我们接下来给出两条引理,用于证明在使用松弛梯度进行梯度下降时二值网络的收敛条件.

引理1　任意β-平滑函数$f(x)$均满足以下不等式:

$$\begin{aligned}&\left|f(x)-f(y)-\nabla f_r(y)(x-y)\right|\\&\leqslant\frac{\left\|x-y\right\|}{2}\left(\beta\left\|x-y\right\|+k\right)\end{aligned} \tag{4}$$

其中,f_r表示松弛后的f,$k=\left\|\nabla f(y)-\nabla f_r(y)\right\|$表示真实梯度和松弛梯度之间的差距.

引理2　二值网络f和其对应的符号函数松弛近似后的二值网络f_r,满足如下不等式:

$$f(x)-f(y)\leqslant\nabla f_r(x)(x-y)-\frac{p}{2\beta}(p-k) \tag{5}$$

其中$p=\left\|\nabla f_r(x)-\nabla f_r(y)\right\|$表示$x$和$y$松弛梯度之间的距离. 引理1和引理2的证明见附录. 其中,引理2是引理1的推论. 最后,我们证明使用松弛函数时,通过梯度下降优化二值网络收敛的充要条件. 根据定义1,二值网络在当且仅当第$(t+1)$次的权重更新值x_{t+1}比第t次的值x_t离最优值更近的条件下收敛. 应用式(1)中的更新规则,我们需要保证以下不等式成立以确保二值网络收敛:

$$\left\|x_t-\eta\nabla f_r(x_t)-x^*\right\|^2\leqslant\left\|x_t-x^*\right\|^2 \tag{6}$$

其中,权重用松弛梯度∇f_r进行优化. 将式(6)的左边展开,从而得到以下不等式:

$$\nabla f_r(x_t)(x_t-x^*)\geqslant\frac{\eta}{2}\left\|\nabla f_r(x_t)\right\|^2 \tag{7}$$

将式(5)中的x和y分别赋值为x_t和x^*,并且由于最优权重x^*的损失函数比x_t的损失函数更小,我们可以得到以下约束

$$0\leqslant f(x_t)-f(x^*)\leqslant\nabla f_r(x_t)(x_t-x^*)-\frac{p}{2\beta}(p-k) \tag{8}$$

由于松弛函数被证明能够提供有效的梯度[41],因此最优权重x^*处的梯度值与x_t处相比是可以忽略不计的. 将式(7)左边代换为上述约束,并忽略$\nabla f_r(x^*)$,我们得到二值网络收敛的充要条件为

$$\left\|\nabla f(x_t)-\nabla f_r(x_t)\right\|\leqslant(1-\beta\eta)\left\|\nabla f_r(x_t)\right\| \tag{9}$$

由于在训练过程中,学习率趋于0,而利普希茨常数是有界的,故$\beta\eta\to0$. 最终我们得到二值网络的实际收敛条件是

$$\left\|\nabla f(x_t)-\nabla f_r(x_t)\right\|\leqslant\left\|\nabla f_r(x_t)\right\| \tag{10}$$

该条件也作为评价梯度准确性和有效性平衡的度量. 式(10)的物理含义是,松弛梯度和真实梯度之间的差距,必须小于松弛梯度的模长. 用接近真实符号函数的松弛函数近似时,两者的差值和松弛值的模长都会变小;反之则松弛梯度和真实梯度之间的差距以及松

弛梯度的模长也会变大. 我们提出的基于自适应梯度优化的二值神经网络根据输入样本动态自适应地选择松弛函数,从而达到梯度准确性和有效性的最优平衡.

3.2 基于自适应适应松弛函数二值网络优化

二值网络通常在梯度反传过程中用可微函数代替原先的符号函数,从而消除不可导性. 一般的松弛运算有如下形式:

$$\boldsymbol{x}_{\text{out}}=\alpha\mathcal{T}_{\beta}\left(\boldsymbol{x}_{\text{in}}\right) \tag{11}$$

其中$\boldsymbol{x}_{\text{in}}$和$\boldsymbol{x}_{\text{out}}$分别是松弛函数$\mathcal{T}_{\beta}$的输入和输出,$\alpha$是用于稳定训练的尺度参数,$\beta$是松弛函数陡峭程度的参数. 松弛函数有多种选择:

S形函数(Sigmoid) S形函数在各处都有平滑且非零的梯度,因此可以使用梯度反传的方法更新网络权重. 为了控制其与符号函数的差距,我们用β调节陡峭程度:

$$\mathcal{T}_{\beta}(x)=\frac{1}{1+\exp\left(-\beta x\right)} \tag{12}$$

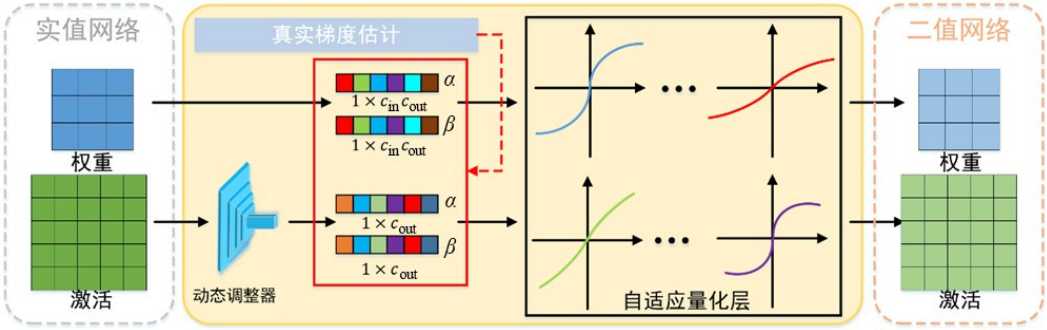
双曲正切函数(Tanh) 双曲正切激活函数和S形函数相似,不同之处在于双曲正切与符号函数的差距更小. 类似地,我们用β调节陡峭程度:

$$\mathcal{T}_{\beta}(x)=\frac{\exp\left(\beta x\right)-\exp\left(-\beta x\right)}{\exp\left(\beta x\right)+\exp\left(-\beta x\right)} \tag{13}$$

多项式函数 分段多项式函数也可以很好地估计符号函数,而且计算复杂度更低. 通过控制最高阶次β,可以得到各种不同陡峭程度的多项式松弛函数. 不同于前两者,多项式函数的β只能取正值:

$$\mathcal{T}_{\beta}(x)=\begin{cases}-1, & x<-1\\ (x+1)^{\beta}-1, & -1\leqslant x<0\\ -(1-x)^{\beta}+1, & 0\leqslant x<1\\ 1, & x\geqslant 1\end{cases} \tag{14}$$

由于二值网络梯度的准确性和有效性是矛盾的,最优化两者之间的平衡可以显著提高性能. 同时,松弛函数应该根据不同的输入来动态自适应地调整尺度参数α和陡峭参数β以获得最优平衡. 图2展示了我们提出的基于自适应梯度量化的二值网络的训练流程.



注: 各层全精度的权重和激活值通过自适应量化层(AdaBin)量化成二值,其中,尺度和陡度参数α和β决定了松弛函数的形式. 对于权重,这两个参数在网络优化的过程中进行更新;对于激活值,我们利用动态调整器计算α和β. 我们的方法使得梯度准确性和有效性的平衡能够在训练阶段动态保持最优. 在模型推理阶段,我们移除动态调整器,并将自适应量化层替换为符号函数.

图2 基于自适应梯度优化的二值神经网络训练流程图

传统的符号函数被替换为自适应量化函数,后者使用S形函数、双曲正切函数或多项式函数进行松弛量化. 权重的尺度参数α和陡峭参数β用梯度反传的方式直接训练,而激活值的这两个参数用动态调整器进行训练. 动态调整器将各层的特征作为输入,并输出对应的尺度参数α和陡峭参数β. 当使用多项式函数近似符号函数时,陡峭参数β被离散化到其最近的整数用于网络的前向和反向传播. 骨干网络和动态调整器的联合训练目标函数如下所示:

$$\min_{W,W_a}\ \mathcal{J}=f\left(x,y,\alpha,\beta\right)+\frac{\gamma}{2}\left(\left\|\nabla f-\nabla f_r\right\|_2^2-\left\|\nabla f_r\right\|_2^2\right) \tag{15}$$

其中,W和W_a分别表示骨干网络的和动态调整器的权重,$\|\cdot\|_2$表示二范数,γ是平衡两项的超参数. 我们使用基于陡峭可微量化的二值网络作为真实梯度的估计器,来获取训练目标中的$\left\|\nabla f\right\|$. 该估计器与实际训练中的骨干网络共享权重和网络结构. 目标函数的第一项旨在最小化二值网络分类任务的损失,来获得最优的分类效果;第二项是为了让二值网络达到梯度准确性和有效性的最佳平衡,从而收敛到最优解. 因此,二值网络的训练稳定性和判别性都得到提升.

我们还利用梯度裁剪的方法来保证训练稳定性. 动态调整器用一个1×1的卷积层和一个全连接层,来获取表示α和β的向量. 在推理阶段,我们移除动态调整器,并且用符号函数量化各层权重和激活,使实际部署时的存储和计算成本不会增加.

4 实验

本节描述上述方法在CIFAR-10和ImageNet这两

个图片分类数据集上的实验结果. 我们首先介绍实现细节,并在方法的各方面性能进行了详细分析,然后还探究了方法中各个技术的有效性. 最后,我们将本方法与现有最优的二值网络训练方法在多种网络结构下进行比较,并分析比较了推理阶段的部署效率.

4.1　实现细节

针对 CIFAR-10 数据集,我们在 VGG-small[27] 和 ResNet20[30, 42] 结构上训练了自适应二值网络;针对 ImageNet,我们在 ResNet18 和 ResNet34 上进行了训练. 为了避免大幅的准确率下降,我们按照 XNOR-Net[2] 中提出的方案,保留第一层和最后一层网络权重和激活值为实值. 在训练二值网络的过程中,我们分别使用 S 形函数、双曲正切函数和多项式函数进行松弛量化. 自适应量化层中的可训练参数,α 和 β 分别初始化为 0.8 和 1.25. 为了放大 β 值的方差从而使松弛量化层更快地逼近符号函数,我们使用一个随着训练过程逐渐增大的超参数 T,令 $T\beta$ 为实际使用的陡峭参数. 在真实梯度估计器中,我们选用 $\boldsymbol{x}_{\text{out}}=\text{sigmoid}\left(100\cdot\boldsymbol{x}_{\text{in}}\right)$ 作为量化函数. 骨干网络和动态调整器的权重采用三阶段的方式进行端到端训练. 首先,固定动态调整器的权重以及自适应量化层中可训练的 α 和 β,只更新骨干网络中的权重. 接着,固定骨干网络的权重,只更新 α 和 β 和动态调整器的权重. 最后,我们联合训练骨干网络、动态调整器权重和可训练的 α 和 β,来获得最优解.

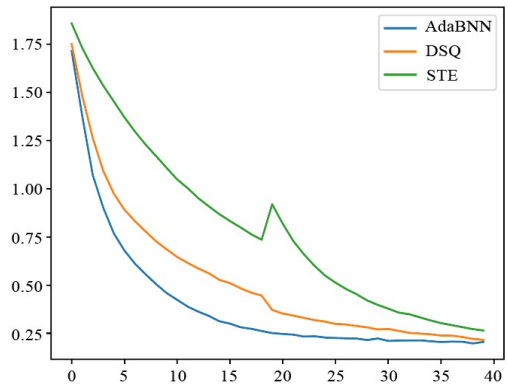
在实验中,我们采用 Adam 优化器[43] 和尺寸为 96 的批数据量进行训练. 对于 CIFAR-10 数据集上的实验,我们分别将上述三阶段训练的轮数设置为 60/40/40 轮. 在每个阶段中,学习率初始值为 0.005,并且在第 15 和第 30 轮分别衰减为之前的 0.1 倍. 在 ImageNet 上训练时,我们使用实值网络的权重作为预训练权重,并在三阶段训练的各个阶段分别训练 30 轮. 学习率初始值设为 0.001,并在 15 和 25 轮分别衰减到 1e−4 和 1e−5. 在训练结束后,我们将自适应量化层替换为符号函数,并在固定其他层的情况下用直通估计器重新训练批标准层(BatchNorm),来吸收尺度和陡度参数.

4.2　性能分析

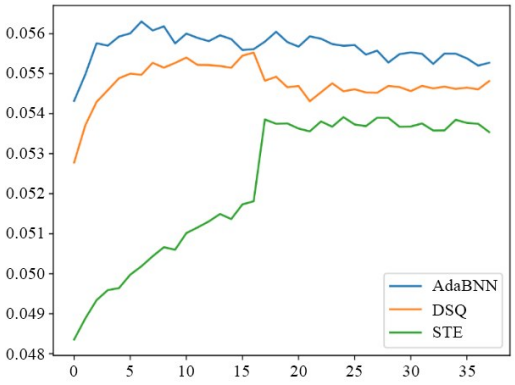
我们从网络收敛性和量化演进的角度,用实验评估了自适应梯度优化的二值神经网络在平衡梯度准确性和有效性方面的优越性. 在本节的实验中,我们在 CIFAR-10 数据集上采用 ResNet-20 结构评价我们的方法.

4.2.1　网络收敛性

传统二值网络在梯度反传时使用直通估计器,这会导致近似梯度与真实梯度有很大的不一致性. 尽管非线形松弛函数可以更准确地估计符号函数,但输入饱和的问题会导致梯度消失,使二值网络权重在每次更新时几乎没有改变. 梯度失配和梯度消失问题会导致训练不稳定以及收敛困难. 手动设置松弛函数中的尺度参数 α 和陡度参数 β 难以获得梯度准确性与有效性平衡的最优解. 为了解决以上问题,我们提出的自适应梯度优化二值网络可以根据输入样本动态地对梯度准确性和有效性的平衡进行动态调整以保证最优,从而显著提升网络收敛性. 我们将自适应梯度优化二值网络与 Xnor-Net[2] 和 DSQ[15] 进行比较. 其中,Xnor-Net 用直通估计器进行训练,而 DSQ 使用双曲正切的松弛函数近似符号函数. 图 3(a) 展示了 $\mathcal{A}=\|\nabla f_r\|_2^2-\|\nabla f-\nabla f_r\|_2^2$,表示松弛梯度的模长相比于它与真实梯度之间的差异的领先量. 图 3(b) 展示了在训练过程中,分类损失函数值的变化. 从中可以看出,自适应二值网络的训练比其他松弛方法更加稳定,因为更大的 $\mathcal{A}$ 能让网络更快收敛. 我们的方法能通过最优的平衡梯度准确性和有效性来稳定训练,并比直通估计器等传统的松弛方法能得到更高性能的二值网络.



(a) 松弛梯度的模长相比于它与真实梯度之间的差异的领先量



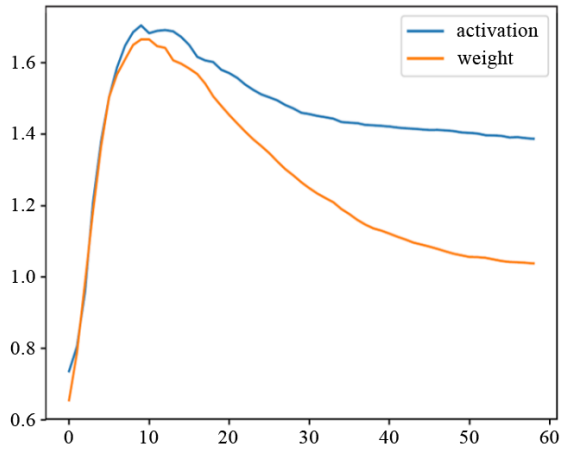
(b) 训练时的分类误差

图 3　不同优化方式下网络训练的演化进程

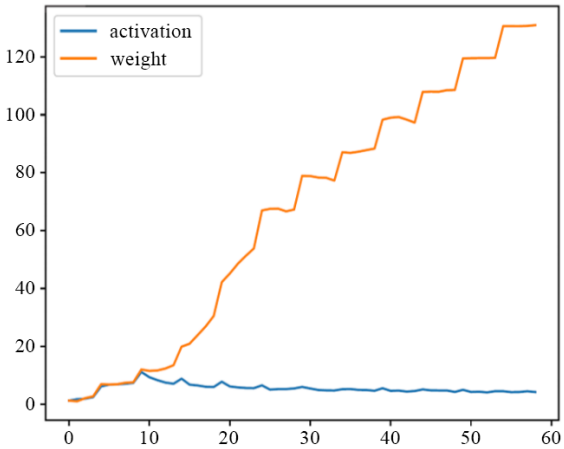
4.2.2　量化演进

非线性松弛函数中尺度参数 α 和陡度参数 β 的动态演进能根据输入动态保证梯度准确性和有效性.

图4分别画出了权重和激活值量化中，最后一个卷积层的α和β的演进过程. 由于激活值的α和β是随着输入改变的，因此我们画出所有输入对应的平均值. 陡度参数β随着训练逐渐增大并最终收敛到一个稳定值，说明自适应量化层在逐步逼近理想的符号函数. 同时，激活量化参数的β远比权重量化的β小，表明激活值比权重对量化更敏感. 尺度参数α在刚开始训练的时候快速增加，并随着训练逐渐衰减到收敛值，这样使松弛函数的极端值被限制到–1到1的区间内，可以最小化量化误差.



(a) 尺度参数α



(b) 陡度参数β

图4 权重激活中各参数的演化过程

4.3 方法有效性分析

为了进一步探究梯度准确性和有效性的平衡对二值网络性能的影响，我们对非线性松弛函数的形式、真实梯度估计器和α与β的自由度进行了消融实验. 我们采用ResNet-20的网络结构作为骨干，在CIFAR-10数据集上训练自适应梯度优化二值网络.

非线性函数的影响. 将符号函数替换为非线性松弛函数来量化权重和激活值，可以增强二值网络的准确率. 然而，不同形式的非线性函数会在不同程度上对训练过程产生影响. 表1中展示了各种非线性函数对最终性能的影响. 与直通估计器相比，所有的非线性函数都取得了更好的性能，而其中S形函数最优. 与其他非线性函数相比，因为S形函数的陡度随参数β变化得最为明显，也就更容易达到梯度准确性和有效性的最优平衡.

表1 选用不同的松弛函数时我们的方法在CIFAR-10数据集上的分类准确率

	松弛函数	ResNet20/%
全精度	-	92.10
二值	恒等	89.90
	S形函数	90.53
	双曲正切	90.17
	多项式	89.66

真实梯度估计器的影响. 真实梯度估计器提供了不可导二值网络反传时对于真实梯度的估计. 真实梯度估计器与骨干网络共享网络结构和权重，只是前者在量化层中使用了极陡的非线性可导函数.

然而，不同的函数类型和不同的陡度也会影响二值网络优化后的性能. 记该函数的形式为$\boldsymbol{x}_{\text{out}}=\mathcal{T}(m\boldsymbol{x})$，其中$\mathcal{T}$为S形函数或者双曲正切函数，$m$可以被赋值为50，100或200. 表2展示了二值网络在不同实验设置下的性能. 在各类非线性函数中，中等的陡度效果最好，因为较小的陡度导致梯度估计不准，而过大的陡度会使式(7)中$\beta\eta\to 0$的条件失效. 同时，在真实梯度估计器中使用S形函数或双曲正切函数的效果没有明显差距.

表2 不同的真实梯度估计器在CIFAR-10数据集上的分类准确率，其中网络结构是ResNet20

m	50	100	200
S形函数	89.99	90.53	89.97
双曲正切函数	90.09	90.49	90.18

***α*与*β*的自由度**. 尺度参数α和陡度参数β控制了梯度准确性和有效性之间的平衡. 在训练阶段，根据不同的输入来动态自适应地选取这两个参数可使得上述平衡达到最优. α与β不同的自由度设定会影响二值网络性能. α与β有三种自由度设定：固定值、静态可训练参数、动态可调整参数. 固定值分别设定为0.8和1.25；作为静态可训练参数时，两者都随着网络训练，但数值与输入样本无关；作为动态可调整参数时，动态调整器可以根据输入样本，输出不同的α与

β. α 与 β 的自由度对结果的影响展示在表3中,第三种方案因其可以动态地满足梯度准确性和有效性的平衡而取得了最优.

表3　CIFAR-10数据集上的分类准确率,选用不同的 α 和 β 自由度来作比较

	固定β/%	静态β/%	动态β/%
固定α	88.81	88.63	89.15
静态α	89.38	89.18	89.45
动态α	89.86	89.94	90.53

4.4　与现有方法的比较

本节在多种网络架构及包括CIFAR-10和ImageNet等图像分类任务的数据集上,比较自适应梯度优化二值网络和现有最佳的二值网络训练方法,后者包括BC[27], BNN[1], ABC-Net[33], Xnor-Net[2], Bi-Real-Net[7],QN[23]和DSQ[15],以及多bit量化网络如BWN[2], HWGQ[44],LQ-Nets[30],PACT[45].

CIFAR-10上的比较. CIFAR-10的训练集和测试集分别包含属于十个类别的50 000和10 000张图片. 我们将各图片的像素值归一化到[−1,1]的区间内,并通过填充和随机裁剪的方式,将图片的尺寸变换成32×32. 我们将VGG-small和ResNet20经过不同量化框架得到的分类准确性进行比较,得到表4中的结果. 对于具有VGG-small和ResNet20结构的二值神经网络,自适应梯度优化二值网络达到了最优性能,并在之前方法的基础上有较大的提升. 我们的方法甚至超过了2比特激活值的HWGQ和权重激活均为2比特的PACT.

表4　在CIFAR-10上与现有方法的分类准确率比较

方法	位宽	VGG-small/%	ResNet20/%
全精度	32/32	93.20	92.10
BC	1/32	90.10	-
TTQ	2/32	-	91.13
HWGQ	1/2	92.50	-
LQ-Net	1/2	93.40	88.40
PACT	2/2	-	89.70
BNN	1/1	89.90	-
DoReFa-Net	1/1	-	79.30
Xnor-Net	1/1	89.80	-
DSQ	1/1	91.72	84.11
AdaBNN	1/1	**92.52**	**90.53**

注:位宽表示权重/激活的比特数

ImageNet上的比较. ImageNet训练集包含约120万张图片,验证集包含5万张,分属1 000个类别. ImageNet的规模更大且类别更丰富,相比于CIFAR-10更具挑战性. 训练时,我们采用和CIFAR-10相同的像素归一化策略,并且将训练图片的短边长调整到256再随机裁剪出224×224的区域以供训练;测试时,使用中心裁剪的策略获得224×224的区域. 因为相邻卷积层之间额外的短连[7]可以增强二值网络的表示能力,所以为了提高性能我们也使用了额外短连的方法. 将自适应二值网络在ResNet18和ResNet34上分别和现有量化方法比较,表5中展示了top-1和top-5准确率. QN和DSQ使用非线性松弛函数取得了不错的效果,但其由于手动调整尺度和陡度参数,只能在梯度准确性和有效性的平衡上达到次优解. 而我们的自适应梯度优化二值网络可以根据输入动态自适应调整到最优解,提高训练效率从而增强了二值网络准确性.

表5　在ImageNet上与现有方法关于第一和前五分类准确率的比较

方法	位宽	ResNet18/%		ResNet34/%	
		top-1	top-5	top-1	top-5
全精度	32/32	69.3	89.2	7.3	91.3
BWN	1/32	60.8	80.3	-	-
HWGQ	1/2	59.6	82.2	64.3	85.7
LQ-Net	1/2	62.6	84.3	66.6	86.9
PACT	2/2	55.4	78.6	-	-
BNN	1/1	42.2	67.1	-	-
Xnor-Net	1/1	51.2	73.2	-	-
ABC-Net	1/1	42.7	67.6	52.4	76.5
QN	1/1	53.6	75.3	-	-
AdaBNN	1/1	54.3	77.4	60.6	82.2
Bi-Real-Net	1/1	56.4	79.5	62.2	83.9
AdaBNN+SC	1/1	**59.3**	**81.2**	**63.8**	**84.8**

注:位宽表示权重/激活的比特数

4.5　部署效率

本节讨论自适应梯度优化二值网络、Xnor-Net、Bi-Real-Net和实值网络的存储代价和计算复杂度. 存储代价用网络参数所需的内存来衡量. 对于实值参数,存储代价为参数量的32倍;对于二值参数,存储代价为参数量的一倍. 计算复杂度用FLOPs[39]衡量,由于CPU可以并行64个二值计算,所以总FLOPs为实值计算的次数加上二值计算次数的1/64.

表6展示了不同方法在ResNet18中的计算复杂度和存储代价. 我们提出的自适应梯度优化二值网络相比实值网络节约了91.07%倍的存储空间,并加速11.91倍. 由于在推理阶段去掉了尺度参数α及相邻模块的

表6　在ResNet18结构上与现有二值网络在计算量和存储代价上的比较

		存储量	FLOPs(G)
ResNet18	全精度	374.1Mbit	1.81
	Xnor-Net	33.7Mbit	0.17
	Bi-Real-Net	33.6Mbit	0.16
	AdaBNN	33.4Mbit	0.15

短连,它的计算复杂度和存储量比Xnor-Net和Bi-Real-Net也要少. 简而言之,自适应梯度优化二值网络比现有的二值神经网络具有更低的计算和存储成本.

5 结论

本文中,我们提出了基于自适应梯度优化的二值神经网络,旨在学习优化二值网络的最佳梯度. 通过动态自适应调整梯度准确性和有效性之间的平衡,我们同时解决了梯度失配和梯度消失的问题,并显著提高二值网络的训练效率和性能. 实验结果表明,我们的方法在CIFAR-10和ImageNet数据集以及多种网络结构上的表现优于现有方法.

附录

一、引理1的证明

引理1 对于任何β-平滑的函数,有以下不等式:

$$\left|f(x)-f(y)-\nabla f_r(y)(x-y)\right|\leqslant\frac{\left\|x-y\right\|}{2}\left(\beta\left\|x-y\right\|+k\right)$$

其中,f_r表示二值网络中松弛后的f,$k=\left\|\nabla f(y)-\nabla f_r(y)\right\|$表示真实梯度和松弛梯度的差距.

证明 构造差值函数$g(t)=f\left[y+t(x-y)\right]$,展开上式的左边:

$$\begin{aligned}&\left|f(x)-f(y)-\nabla f_r(y)(x-y)\right|\\&=\int_0^1\left|g'(t)-\nabla f_r(y)(x-y)\right|\mathrm{d}t\\&=\int_0^1\left|\nabla f\left[y+t(x-y)\right](x-y)-\nabla f_r(y)(x-y)\right|\mathrm{d}t\\&\leqslant\int_0^1\left|\left\{\nabla f\left[y+t(x-y)\right]-\nabla f_r(y)\right\}(x-y)\right|\mathrm{d}t\end{aligned}$$

根据柯西-施瓦茨不等式和β-平滑函数的定义,得上式的上界:

$$\begin{aligned}&\int\left|\left\{\nabla f[y+t(x-y)]-\nabla f_r(y)\right\}(x-y)\right|\mathrm{d}t\\&\leqslant\int_0^1\sqrt{\left\|\nabla f[y+t(x-y)]-\nabla f_r(y)\right\|^2\left\|x-y\right\|^2}\,\mathrm{d}t\\&\leqslant\int_0^1(\left\|\nabla f[y+t(x-y)]-\nabla f(y)+\nabla f(y)\right.\\&\quad\left.-\nabla f_r(y)\right\|^2\cdot\left\|x-y\right\|^2)^{1/2}\mathrm{d}t\\&\leqslant\int_0^1[(\left\|\nabla f[y+t(x-y)]-\nabla f(y)\right\|+\left\|\nabla f(y)\right.\\&\quad\left.-\nabla f_r(y)\right\|)^2\cdot\left\|x-y\right\|^2]^{1/2}\mathrm{d}t\\&\leqslant\int_0^1[(2\beta' t\left\|x-y\right\|+k)\left\|x-y\right\|]\mathrm{d}t\\&=\frac{\left\|x-y\right\|}{2}(\beta\left\|x-y\right\|+k)\end{aligned}$$

证毕

二、引理2的证明

引理2 对于二值网络f和松弛符号函数后的二值网络f_r,有以下不等式:

$$f(x)-f(y)\leq\nabla f_r(x)(x-y)-\frac{p}{2\beta}(p-k)$$

其中,$p=\left\|\nabla f_r(x)-\nabla f_r(y)\right\|$表示$x$和$y$的松弛梯度之差.

证明 定义$z=y-\frac{1}{\beta}\left(\nabla f_r(y)-\nabla f_r(x)\right)$,重写上式的左边:

$$f(x)-f(y)=f(x)-f(z)+f(z)-f(y)$$

由于反传的有效性,我们假设松弛后的损失函数是凹函数,可以得到下式:

$$\begin{aligned}f(x)-f(z)&\leqslant\nabla f_r(x)(x-z)\\&=\nabla f_r(x)(x-y)+\nabla f_r(x)(y-z)\end{aligned}$$

根据引理1的结论,得到

$$\begin{aligned}f(z)-f(y)&\leqslant\nabla f_r(y)(z-y)\\&\quad+\frac{\left\|z-y\right\|}{2}\left(\beta\left\|z-y\right\|+k\right)\#\end{aligned}$$

联合上面两式,并用x和y代替z,得到下列不等式:

$$\begin{aligned}f(x)-f(y)&\leqslant\nabla f_r(x)(x-y)\\&\quad+(\nabla f_r(x)-\nabla f_r(y))(y-z)\\&\quad+\frac{\left\|z-y\right\|}{2}(\beta\left\|z-y\right\|+k)\\&=\nabla f_r(x)(x-y)-\frac{1}{\beta}p^2+\frac{1}{2\beta}p(p+k)\\&=\nabla f_r(x)(x-y)-\frac{p}{2\beta}(p-k)\end{aligned}$$

证毕

参考文献

[1] HUBARA I, COURBARIAUX M, SOUDRY D, et al. Binarized neural networks[C]//Advances in Neural Information Processing Systems. Barcelona: NIPS, 2016: 4107-4115.

[2] RASTEGARI M, ORDONEZ V, REDMON J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 525-542.

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

[4] 侯志强, 刘晓义, 余旺盛, 等. 使用GIoU改进非极大值抑制的目标检测算法[J]. 电子学报, 2021, 49(4): 696-705.
HOU Z Q, Liu X Y, YU W S, et al. Object detection algorithm for improving non-maximum suppression using GIoU[J]. Acta Electronica Sinica, 2021,49(4): 696-705. (in Chinese)

[5] 李雅倩, 盖成远, 肖存军, 等. 基于细化多尺度深度特征的目标检测网络[J]. 电子学报, 2020, 48(12): 2360-2366.
LI Y Q, GAI C Y, XIAO C J, et al. Objection detection networks based on refined multi-scale depth feature[J]. Acta Electronica Sinica, 2020, 48(12): 2360-2366. (in Chinese)

[6] 李维刚, 叶欣, 赵云涛, 等. 基于改进YOLOv3算法的带钢表面缺陷检测[J]. 电子学报, 2020, 48(7): 1284-1292.
LI W G, YE X, ZHAO Y T, et al. Strip steel surface defect detection based on improved YOLOv3 algorithm[J]. Acta Electronica Sinica, 2020, 48(7): 1284-1292. (in Chinese)

[7] LIU Z, WU B, LUO W, et al. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm[C]//European Conference on Computer Vision. Munich: Springer, 2018: 722-737.

[8] 江泽涛, 秦嘉奇, 张少钦. 参数池化卷积神经网络图像分类方法.电子学报[J], 2020, 48(9): 1729-1734.
JIANG Z T, QIN J Q, ZHANG S Q. Parameterized pooling convolution neural network for image classification[J]. Acta Electronica Sinica, 2020, 48(9): 1729-1734. (in Chinese)

[9] WEI Y, PAN X, QIN H, et al. Quantization mimic: Towards very tiny cnn for object detection[C]//European Conference on Computer Vision. Munich: Springer, 2018: 267-283.

[10] WANG Z, LU J, ZHOU J. Learning Channel-Wise Interactions for Binary Convolutional Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3432-3445.

[11] 葛疏雨, 高子淋, 张冰冰, 等. 基于核化双线性卷积网络的细粒度图像分类[J]. 电子学报, 2019, 47(10): 2134-2141.
GE S Y, GAO Z L, ZHANG B B, et al. Kernelized bilinear CNN models for fine-grained visual recognition[J]. Acta Electronica Sinica, 2019, 47(10): 2134-2141. (in Chinese)

[12] HOU L, KWOK J T. Loss-aware weight quantization of deep networks[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-11.

[13] LENG C, DOU Z, LI H, et al. Extremely low bit neural network: Squeeze the last bit out with admm[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 3466-3473.

[14] ALIZADEH M, FERNÁNDEZ-MARQUÉS J, LANE N D, et al. An empirical study of binary neural networks' optimization[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-10.

[15] YIN P, LYU J, ZHANG S, et al. Understanding straight-through estimator in training activation quantized neural nets[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019: 1-12.

[16] GONG R, LIU X, JIANG S, et al. Differentiable soft quantization: Bridging full-precision and low-bit neural networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4852-4861.

[17] KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features From Tiny Images[R]. Toronto: University of Toronto, 2009.

[18] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248-255.

[19] WANG Z, LU J, WU Z, et al. Learning efficient binarized object detectors with information compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3082-3095.

[20] 龚成, 卢冶, 代素蓉, 等. 一种超低损失的深度神经网络量化压缩方法[J]. 软件学报, 2021, 32(8): 2391-2407.
GONG C, LU Y, DAI S R, et al. Ultra-low loss quantization method for deep neural network compression[J]. Journal of Software, 2021, 32(8): 2391-2407. (in Chinese)

[21] WANG Z, XIAO H, LU J, et al. Generalizable mixed-precision quantization via attribution rank preservation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 5291-5300.

[22] WANG Z, ZHENG Q, LU J, et al. Deep hashing with active pairwise supervision[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 522-538.

[23] YANG J, SHEN X, XING J, et al. Quantization networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7308-7316.

[24] 汪海龙, 禹晶, 肖创柏. 基于点对相似度的深度非松弛哈希算法[J]. 自动化学报, 2021, 47(5): 1077-1086.
WANG H L, YU J, XIAO C B. Deep non-relaxing hashing based on point pair similarity[J]. Acta Automatica Sinica, 2021, 47(5): 1077-1086. (in Chinese)

[25] 董震, 裴明涛. 基于异构哈希网络的跨模态人脸检索方法[J]. 计算机学报, 2019, 42(1): 73-84.
DONG Zhen, PEI M T. Cross-modality face retrieval based on heterogeneous hashing network[J]. Chinese Journal of Computers, 2019, 42(1): 73-84. (in Chinese)

[26] ZHU C, HAN S, MAO H, et al. Trained ternary quantization[C]// International Conference on Learning Representations. Toulon: ICLR, 2017: 1-10.

[27] COURBARIAUX M, BENGIO Y, DAVID J P. Binaryconnect: Training deep neural networks with binary

weights during propagations[C]//Advances in Neural Information Processing Systems. Montréal: NIPS, 2015: 3123-3131.
[28] ZHOU S, WU Y, NI Z, et al. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients[C]// International Conference on Learning Representations. San Juan: ICLR, 2016: 1-13.
[29] LOUIZOS C, ULLRICH K, WELLING M. Bayesian compression for deep learning[C]//Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017: 3288-3298.
[30] ZHANG D, YANG J, YE D, et al. Lq-nets: Learned quantization for highly accurate and compact deep neural networks[C]//European Conference on Computer Vision. Munich: Springer, 2018: 365-382.
[31] ULLRICH K, MEEDS E, WELLING M. Soft weight-sharing for neural network compression[C]//International Conference on Learning Representations. Toulon: ICLR, 2017: 1-10.
[32] BANNER R, HUBARA I, HOFFER E, et al. Scalable methods for 8-bit training of neural networks[C]//Advances in Neural Information Processing Systems. Montréal: NIPS, 2018: 5145-5153.
[33] LIN X, ZHAO C, PAN W. Towards accurate binary convolutional neural network[C]//Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017: 345-353.
[34] BETHGE J, HYANG H, BORNSTEIN M, et al. Back to simplicity: How to train accurate bnns from scratch? [EB/OL]. (2019-06-19)[2021-08-13]. https://arxiv.org/abs/1906.08637.
[35] DUAN Y, LU J, WANG Z, et al. Learning deep binary descriptor with multi-quantization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41 (8): 1924-1938.
[36] TANG W, HUA G, WANG L. How to train a compact binary neural network with high accuracy?[C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 2625-2631.
[37] WANG P, HU Q, ZHANG Y, et al. Two-step quantization for low-bit neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4376-4384.
[38] GU J, LI C, ZHANG B, et al. Projection convolutional neural networks for 1-bit cnns via discrete back propagation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8344-8351.
[39] BOYD S, PARIKH N, CHU E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends in Machine learning, 2011, 3(1): 1-122.
[40] YIN P, LYU J, ZHANG S, et al. Understanding straight-through estimator in training activation quantized neural nets[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019: 1-12.
[41] ANDERSON A G, BERG C P. The high-dimensional geometry of binary neural networks[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-10.
[42] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.
[43] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//International Conference on Learning Representations. San Diego: ICLR, 2015: 1-11.
[44] CAI Z, HE X, SUN J, et al. Vasconcelos. Deep learning with low precision by half-wave gaussian quantization [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5918-5926.
[45] CHOI J, WANG Z, VENKATARAMANI S, et al. PACT: Parameterized clipping activation for quantized neural networks[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-11.

作者简介

王子为 男,1996年生,湖南益阳人. 清华大学自动化系博士研究生. 主要研究方向为特征学习和模型压缩.
E-mail: wang-zw18@mails.tsinghua.edu.cn

鲁继文(通讯作者) 男,1981年生,湖北武穴人. 清华大学自动化系长聘副教授,IAPR Fellow. 主要研究方向为计算机视觉和模式识别. 获国家杰出青年科学基金项目资助. 中国电子学会会员编号:E190014211M.
E-mail: lujiwen@tsinghua.edu.cn

周 杰 男,1968年生,河南信阳人. 清华大学自动化系教授,IAPR Fellow. 主要研究方向为计算机视觉和模式识别. 获国家杰出青年科学基金项目资助.
E-mail: jzhou@tsinghua.edu.cn