二值图像超分辨率重建网络

姜馨蕊¹, 王楠楠^{1*}, 辛经纬¹, 李柯宇¹, 杨曦¹, 高新波²

1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室, 西安 710071

2. 重庆邮电大学重庆市图像认知重点实验室, 重庆 400065

* 通信作者. E-mail: nnwang@xidian.edu.cn

收稿日期: 2020-11-04; 修回日期: 2021-01-08; 接受日期: 2021-02-23; 网络出版日期: 2021-10-15

2030 新一代人工智能重大专项 (批准号: 2018AAA0103202) 和国家自然科学基金 (批准号: 61922066, 61876142, 62036007) 资助项目

摘要 近年来, 深层卷积神经网络在图像超分辨率重建任务中取得了巨大成功, 然而复杂的深度神经网络会消耗大量存储空间以及计算资源, 严重限制了其在资源有限的移动设备上的部署. 因此降低模型的资源消耗将有助于扩展深度超分辨率网络的实际应用范围. 二值神经网络占用存储空间小、计算效率高, 激励我们将二值化算法应用于目前的深度超分辨率重建领域, 满足现有移动设备对于超分辨率的实际应用需求. 因此, 本文关注于二值图像超分辨率重建网络的研究. 为此, 我们首先总结了现有二值化方法, 并针对其技术细节和算法特点进行了详细介绍. 随后, 我们探索了目前二值化方法在超分辨率领域的实际应用效果, 并面向图像超分辨率重建任务提出一种新的二值化算法, 主要通过提高网络前向过程表达能力和减少网络反向过程训练损失提升二值超分辨率网络的性能. 实验表明, 无论对比现有基于分类任务的二值化算法还是对比基于超分辨率任务的二值化算法, 我们的方法均可以取得最优的性能.

关键词 二值卷积神经网络, 图像超分辨率重建, 二值化, 量化, 模型压缩

1 引言

近年来, 随着人工智能技术的不断发展, 卷积神经网络在众多领域取得了巨大成功, 特别是计算机视觉、语音识别、自然语言处理等领域. 大体上来讲, 卷积网络的学习能力与其模型容量有关, 增加网络的复杂程度可以提高网络的学习能力, 从而取得更优异的性能. 自 2012 年 Hinton 小组以 AlexNet 网络^[1] 拿下 ImageNet 竞赛冠军后, 深度神经网络在计算机视觉领域大放异彩, 随后涌现出更多更深的网络结构, 如 VGG^[2], ResNet^[3], DenseNet^[4] 等. 然而, 随着识别精度的不断提升, 网络结构也变得越来越复杂, 因此带来了过高的存储代价和计算需求. 例如, Alex 网络^[1] 包含约 61 M 模型参数, 需要

引用格式: 姜馨蕊, 王楠楠, 辛经纬, 等. 二值图像超分辨率重建网络. 中国科学: 信息科学, 2021, 51: 1690–1705, doi: 10.1360/SSI-2020-0346
Jiang X R, Wang N N, Xin J W, et al. Binary neural networks for image super-resolution (in Chinese). Sci Sin Inform, 2021, 51: 1690–1705, doi: 10.1360/SSI-2020-0346

		Operations used in convolution	Memory saving	Computation saving
Standard convolution	<div> <div>Full-precision activations</div> <div>Full-precision weights</div> </div>	$+, -, \times$	$1 \times$	$1 \times$
Binary weights	<div> <div>Full-precision activations</div> <div>Binary weights</div> </div>	$+, -$	$\sim 32 \times$	$\sim 2 \times$
Binary weights and activations	<div> <div>Binary activations</div> <div>Binary weights</div> </div>	XNOR, bitcount	$\sim 32 \times$	$\sim 58 \times$

图 1 (网络版彩图) 二值卷积神经网络示意图

Figure 1 (Color online) The pipeline of binary neural networks

超过约 249 M 的模型存储空间以及 1.5×10^9 次浮点型算术操作; VGG-16^[2] 网络包含约 138 M 模型参数, 需要超过约 500 M 的存储空间以及 1.6×10^{10} 次浮点型算术操作. 尽管这些深度网络性能优异, 但很难实现实时处理, 且依赖于高性能的计算设备如图形处理单元 (graphics processing unit, GPU) 或张量处理单元 (tensor processing unit, TPU), 严重限制了其在手机等便携式嵌入设备中的应用. 因此, 在不影响深度网络模型性能的前提下, 对于深度卷积网络的压缩和加速就变得尤为重要.

目前, 模型压缩是人工智能领域内一个重要的研究方向, 有效的压缩方法可以精简原本需要占用大量内存和计算资源的模型. 大体上来讲, 模型压缩方法可以分为 4 大类: 模型剪枝、低秩分解、模型量化和知识蒸馏. 其中, 模型剪枝^[5,6] 通过寻找一种有效的评判准则评价参数的重要程度, 之后去除网络中不重要的模型参数达到缩减模型尺寸的目的; 低秩分解^[7,8] 通过组合多个低秩矩阵来近似满秩矩阵, 随后将低秩矩阵分解为小规模矩阵乘积实现计算资源的节约; 模型量化^[9,10] 大多通过较低的推理精度损失将连续取值的浮点型参数近似为有限多个离散值, 从而达到减小模型尺寸、减少内存消耗的目标; 知识蒸馏^[11,12] 则试图提炼出一个更紧凑的小网络模型再现一个大网络模型的性能.

在现有的模型压缩算法中, 模型量化被视为最有效的模型压缩技术. 其中, 模型二值化作为最大限度的模型量化方法, 将网络运算需要的参数约束到 +1 和 -1, 具有最为突出的压缩性能. 如文献^[9] 中所述 (图 1 所示), 与全精度网络模型相比, 二值神经网络通过将模型权值约束为 1-bit 参数, 可以减少约 32 倍的模型尺寸, 极大程度上减少了模型的内存耗用. 同时, 二值权值网络中的卷积操作可以仅由加法和减法操作实现, 带来了二倍速度增益^[9]. 文献^[9] 指出当网络中的权值与运行计算时得到的激活值均被量化为 1-bit 数据时, 模型运算需要的乘法累加操作可以被替换为简单的位运算 (如图 2 所示), 极大程度上降低了网络运行的计算花费, 加快了模型的推理速度. 例如, 1 个 32 位浮点型的乘法操作花费约 200 块 Xilinx FPGA, 而 1-bit 的逻辑运算仅需要一块^[13], 且二值化只会引入一个额外的量化操作, 享有很大的加速收益. 因此, 二值神经网络非常适用于资源有限的移动设备. 通过对二值网络的探索, 可使得全精度深度网络能够配备在小型便携设备中.

图像超分辨率重建任务旨在通过一个低分辨率图像重建一个高分辨图像, 享有非常广阔的应用空间, 如医学图像处理、遥感图像处理、高清视频等^[14~16]. 近年来, 深层神经网络在图像超分辨率重建任务中取得了显著优异的结果, 涌现出了许多优秀的超分辨率重建算法^[17~19]. 然而, 在超分辨率领

图 2 (网络版彩图) 二值网络的卷积过程

Figure 2 (Color online) The convolution process of binary neural networks

域, 网络的性能提升同样依赖于增大模型的容量和复杂程度. 因此, 现有超分辨率重建网络依然面临着难以部署于资源有限的移动端设备的问题. 由于二值神经网络占用存储空间小、计算效率高, 受此激励, 我们致力于将模型二值化方法引入图像超分辨率重建领域, 旨在保证超分辨率网络性能的前提下, 实现超分辨率网络的完全二值化, 满足手机等便携式移动设备对于超分辨率过程的计算要求. 因此, 对于二值超分辨率重建网络的研究, 将对整个超分辨率领域有着非常重要的理论意义和实际应用价值.

本文关注于二值图像超分辨率重建网络的研究. 为此, 我们首先对现有二值神经网络进行总结 (第 2 节), 详细介绍现有二值化方法的技术细节与方法性能. 其次, 本文探索现有二值化算法在图像超分辨率重建领域的实际应用效果 (第 3 节), 从实验角度分析基于不同二值化方法的超分辨率网络性能. 在此基础上, 我们进一步挖掘二值化超分辨率网络的潜力 (第 4 节), 主要从提高网络表达能力和减少量化训练损失两方面提升二值超分辨率网络的性能, 并进行充分实验证明我们方法的优越性. 随后, 我们总结了当前二值超分辨率重建网络存在的局限性和未来的发展趋势 (第 5 节). 最后我们对全文进行总结归纳 (第 6 节).

2 二值卷积神经网络

2.1 二值卷积神经网络基本描述

在全精度卷积神经网络中, 基本的卷积操作可以表示为

$$Z = W \times A, \quad (1)$$

其中, W 和 A 表示进行卷积运算的权值和激活, $\times$ 表示卷积算子. 全精度卷积神经网络中的运算为浮点运算, 会在网络前向推理过程中消耗大量的计算资源. 因此, 二值神经网络将网络运算所需参数的比特数降低为 1-bit, 同时将模型运算所需的乘法操作替换为简单的位运算, 以此达到减少模型内存耗用, 加快模型推理速度的目的. 模型二值化过程具体如下.

前向过程二值化: 自 2016 年 Hubara 等^[20]提出 BNN 后, 模型二值化成为了有效的模型压缩方法. 对于如何实现网络参数的二值化, 现有方法通常采用 sign 函数获取二值化权值和激活.

$$W^b = \text{sign}(W) = \begin{cases} -1, & \text{if } W < 0, \\ +1, & \text{others,} \end{cases} \quad A^b = \text{sign}(A) = \begin{cases} -1, & \text{if } A < 0, \\ +1, & \text{others.} \end{cases} \quad (2)$$

图 3 (a) Sign 函数; (b) sign 函数的导数; (c) 截断操作; (d) 截断操作对于 sign 函数的导数近似

Figure 3 (a) Sign function; (b) the derivative of the sign function; (c) clip function; (d) the derivative of the clip function for approximating the derivative of the sign function

此时, 卷积的前向运算过程可以转换为

$$Z = W^b \odot A^b, \quad (3)$$

其中, W^b 表示二值化的权值, A^b 表示二值化的激活, $\odot$ 表示 XNOR 操作.

反向过程二值化: 与全精度网络相同, 二值卷积神经网络在网络训练过程中依靠误差反向传播算法 (error back propagation, BP) 更新网络权值. 对于二值网络, 梯度的计算过程为

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial x^b} \frac{\partial x^b}{\partial x} = \frac{\partial L}{\partial x^b} \frac{\partial \text{sign}(x)}{\partial x}, \quad (4)$$

其中, L 表示损失函数, x 和 x^b 表示二值权值或激活. 然而, 由于二值化函数 (sign 函数) 的离散性, 其直接求导得到的导数值 $\frac{\partial \text{sign}(x)}{\partial x}$ 处处为 0 (如图 3(b) 所示), 无法直接进行梯度的反向传播指导权值更新. 因此, 二值神经网络大多采用直通估计器 (straight-through estimator, STE) [21] 近似 sign 函数的梯度. 具体表现为, 对于网络中 $[-1, +1]$ 区间内的参数值, 其梯度值直接定义为 1, 对于 $[-1, +1]$ 区间外的参数值, 其梯度直接赋为 0 (如图 3(d) 所示). STE 具体操作如下:

$$\text{clip}(x, -1, 1) = \max(-1, \min(1, x)). \quad (5)$$

通过 STE 可以直接对网络未量化前的全精度权值进行更新, 实现对二值化神经网络的优化.

2.2 二值卷积神经网络发展概述

近年来, 二值卷积神经网络领域涌现出许多优秀的二值化算法, 在图像分类、图像识别等领域取得了非常优异的成果. 其中, 早期朴素二值化算法直接将网络中的权值和激活进行量化, 与全精度网络相比, 二值神经网络的性能受到严重破坏. 目前基于多角度进行优化改进的二值化算法被广泛研究, 二值神经网络与全精度网络之间的性能差被有效降低 [22]. 本文对现有二值化算法进行了广泛调研, 并将现有基于多角度优化改进的二值化算法分为两大类: 一是在网络前向信息传递过程中, 通过网络结构设计, 引入增益项等手段减少网络的量化误差, 提高二值网络的表达能力. 二是在反向信息传递过程中, 通过精心设计的二值化函数提高网络的精确训练能力. 下面, 本文将从这两方面阐述目前二值化算法技术细节的具体改进.

2.2.1 基于前向过程改进的二值卷积神经网络

作为早期的二值化代表性算法, XNOR-Net [9] 实现了网络权值与激活的全部二值化, 并首次引入尺度因子增加权值和激活的表达范围. 首先, XNOR-Net 确定了一个最佳的量化矩阵:

$$J(B, \alpha) = \|W - \alpha B\|^2, \quad (6)$$

$$\alpha^*, B^* = \underset{\alpha, B}{\operatorname{argmin}} J(B, \alpha), \quad (7)$$

其中, W 表示全精度的权值, $B \in \{-1, +1\}$ 表示量化后的权值矩阵, α 表示引入的权值增益项, 推导可得

$$B^* = \operatorname{sign}(W), \quad \alpha^* = \frac{1}{n} \|W\|_{l_1}. \quad (8)$$

同样的, XNOR-Net 引入激活增益项:

$$X^* = \operatorname{sign}(X), \quad \gamma^* = \frac{1}{n} \|X\|_{l_1}. \quad (9)$$

XNOR-Net^[9] 的二值卷积过程通过引入权值尺度因子 (α) 和激活尺度因子 (γ^*) 近似全精度卷积过程, 有效较少了量化过程中产生的量化误差. DoReFa-Net^[23] 通过可微分的非线性激活函数将激活值约束到固定区间, 随后进行权值、激活和梯度的低 bit 量化. 在此基础上, ABC-Net^[24] 采用多个二值化卷积与激活的线性组合逼近全精度网络的性能. High-Order Residual Quantization^[25] 通过高阶滤波器逐步补偿量化损失. 此外, 为了减少网络前向信息传递过程中的量化损失, Bi-Real Net^[26] 采用短距离跳跃连接增加网络的表征能力, 通过将上一个阶段 BN 输出的实数结果与当前阶段 BN 输出的实数结果相加, 有效扩大了网络中激活的表达范围, 增加了二值神经网络更多样化的表达能力. XNOR-Net++^[27] 则将二值卷积过程中的权值与激活的尺度因子融合为一个可通过网络自适应学习的参数, 解决 XNOR 中由于固定尺度因子造成的激活值被限制在固定区间的问题. 与 XNOR-Net++ 方法类似, RTN^[28] 网络提出了重新参数化量化数值的总体流程, 量化后的激活由一个尺度因子重新调整大小, 同时由一个偏移因子重新调整区间, 以实现更高的网络容量. 此外, 对于 1-bit 网络而言, 性能对于激活值的特征分布十分敏感. 因此, ReActNet^[29] 提出了一个新的激活函数, 可以重塑和移动二值网络的激活分布. 基于传统的 sign 函数和 PReLU 激活函数, 该方法分别学习一组偏移量, 将函数分别泛化为 RSign 和 PReLU 在网络训练过程中自适应地调整激活的特征分布. Real-to-Bin^[30] 则利用全精度网络对二值网络进行指导, 通过对网络不同层特征分布的监督, 二值网络的性能被有效提升.

2.2.2 基于反向过程改进的二值卷积神经网络

由于二值化函数的离散性, 大多方法采用 STE 近似 sign 的梯度指导参数更新. 然而, 利用 STE 对 sign 函数导数的近似会带来较大的近似误差, 导致梯度的不匹配问题. 为此, Bi-Real Net^[26] 首次提出使用一个精心设计的二值化函数近似 sign 的导数缓解梯度不匹配问题. 二值化函数如下:

$$\operatorname{ApproxSign}(x) = \begin{cases} -1, & \text{if } x < -1, \\ 2x + x^2, & \text{if } -1 \leq x < 0, \\ 2x - x^2, & \text{if } 0 \leq x < 1, \\ 1, & \text{otherwise,} \end{cases} \quad \frac{\partial \operatorname{ApproxSign}(x)}{\partial x} = \begin{cases} 2 + 2x, & \text{if } -1 \leq x < 0, \\ 2 - 2x, & \text{if } 0 \leq x < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

同 Bi-Real Net 近似, Self-Binarizing Net^[31] 利用 tanh 函数的变体逐步减小梯度的饱和区间实现对二值函数的精确近似. BNN+^[32] 则引入可变参数调整梯度的饱和区间, 寻求最优的梯度近似. 二值化函数具体为

$$SS_{\beta}(x) = 2\sigma(\beta x)[1 + \beta x \{1 - \sigma(\beta x)\}] - 1. \quad (11)$$

此外, IRNet^[33] 采用了两阶段梯度近似的方法最小化反向过程中信息传递的损失, 首先在第一阶段保留梯度反向传播算法的更新能力, 将梯度估计函数的导数值截断为 1; 之后, 保留参数在 0 附近的精

表 1 不同二值神经网络方法在 ImageNet 数据集上的分类性能 (%)

Table 1 Image classification performance of binary neural networks on ImageNet dataset

		Full-Precision neural networks	BNN	XNOR	ABC-Net	Bi-Real Net
Top-1	ResNet18	69.6	42.2	51.2	65.0	56.4
	ResNet34	73.3	—	—	—	62.2
Top-5	ResNet18	89.2	69.2	73.2	85.9	79.5
	ResNet34	91.3	—	—	—	83.9
		BWN (Full-Precision activation)	XNOR++	BNN+	IRNet	ReActNet
Top-1	ResNet18	60.8	57.1	53.01	58.1	69.4
	ResNet34	—	—	—	62.9	—
Top-5	ResNet18	83.0	79.9	72.6	80.0	—
	ResNet34	—	—	—	84.1	—

度, 并逐步将导数曲线演变成阶梯函数的形状. IRNet 的训练将有效缓解量化前后梯度的不匹配问题, 使得网络训练过程中的参数得到合理的更新.

2.3 实验结果

本小节我们简要介绍现有二值神经网络应用于目标分类任务上的性能. 为方便比较, 我们选用分类任务中常用的数据集: ImageNet 大规模图像数据集^[34] (由 1000 个类别组成, 包含约 120 万张训练图片和 5 万张验证图像) 以及分类任务中常用的基础网络 ResNet18 和 ResNet34^[3], 文中比较的结果均为原始论文中发表的结果, 性能评估如表 1 所示.

从表中我们可以看出, 早期的朴素二值化方法 BNN 对于网络性能损害严重, 二值化网络与全精度网络之间存在较大性能差. XNOR, ABC-Net, Bi-Real Net, XNOR++, ReActNet 等算法的高性能表明尺度因子、跳跃连接、分布调整等方式确实可以提高网络性能. 此外, BNN 与 BNN+, Bi-Real Net, IRNet 的性能对比表明反向传播过程中存在的梯度不匹配问题对网络训练存在较大的不利影响. 梯度近似方法能够缓解反向传播过程中的梯度失配问题.

3 二值超分辨率重建网络

3.1 背景

目前对于二值超分辨率重建网络的研究已有初步进展, Ma 等^[35]的工作首次考虑将二值化思想引入超分辨率重建领域. 作为首个二值超分辨率重建方法, 该方法成功实现了超分辨率网络的权值二值化. 基于 XNOR-Net^[9], 该算法采用 sign 函数二值化网络权值并引入可学习的尺度因子减少量化过程造成的损失. 反向过程同样采用 clip 操作, 将反向梯度保留在 $(-5, 5)$ 区间内指导网络训练. 然而, 该方法仅实现了二值超分辨率网络中的权值二值化, 并未实现网络中激活的二值化. 因此网络并不是完全的二值网络, 虽然在一定程度上可以减少模型的内存耗用, 但模型运算需要的乘法操作无法被替换为简单的位运算, 从而无法加快模型的推理速度. 因此, 基于该方法的二值超分辨率网络效率依然有限, 无法满足计算资源有限的移动设备的需求.

随后, Xin 等^[36]首次探索了一个权值与激活完全二值化的超分辨率重建网络. 在前向过程中, 该方法提出了一个基于比特累加机制 (bit accumulation mechanism, BAM) 的二值化算法, 通过多个 1-bit

数值的累加估计多 bit 的数值, 在前向信息传递过程中可有效减少量化过程产生的信息损失. 该方法采用结构块的形式沿模型推理方向逐步累加 1-bit 的权值与激活, 并在每一次累加过程中引入可学习的尺度因子进一步提高量化精度. 然而, 过多实值权值与实值激活的累计相加会消耗一定的计算资源, 降低 1-bit 模型硬件实施的有效性. 此外, 该方法简单引入 STE 实现对 sign 函数的导数近似, 因此反向传递过程中依然存在大量信息损失, 导致全精度网络与二值网络之间依然存在较大的性能差.

3.2 二值超分辨率重建网络性能研究

考虑到上述方法均是对二值超分辨率领域的初步探索, 其二值网络的性能和效率均存在较大的提升空间. 因此, 本文进一步探索二值化算法在超分辨率领域的实际应用效果, 挖掘二值超分辨率网络更深层次的潜力. 因此本文将从两个方面对二值超分辨率网络进行探索和分析: 首先我们考虑将目前具有代表性的二值神经网络引入超分辨率领域, 实验分析不同方法在超分辨率任务中的性能. 其次, 我们将现有二值化算法应用于具有代表性的超分辨率网络, 充分挖掘其在超分辨率任务上的训练潜力.

3.2.1 二值神经网络在超分辨率任务上的性能研究

实验细节. 首先, 我们选取目前分类任务中具有代表性的二值神经网络来评估其在超分辨率领域内的性能, 网络基于 ResNet18^[3] 网络结构, 但为满足超分辨率任务, 我们将其最后用于分类的全连接层替换为上采样模块. 训练基于图像超分辨率领域广泛使用的数据集 DIV2K^[37], 测试选用标准的测试数据集 Set5^[38], Set14^[39], Urban100^[40] 和 B100^[41]. 此外, 本文使用峰值信噪比 (peak signal-to-noise ratio, PSNR) 和结构相似度 (structure similarity index, SSIM) 来衡量网络性能.

训练细节. 本小节所有实验均为 scale $\times 4$ (4 倍放大) 超分辨率任务. 输入的低分辨率图像大小为 48×48 , 由双三次插值 (bicubic interpolation) 下采样得到. 实验中需要的超参数 mini-batch 被设置为 16, 初始学习率为 1×10^{-4} , 训练次数为 300 次. 所有实验均选用 Adam 优化器, 其内部参数 β_1 和 β_2 被设置为 0.99 和 0.999. 此外, 实验采用 PyTorch 1.2 框架, 硬件设备为 Telsa V100 GPU.

实验结果. 在此, 我们选取了 BNN, XNOR, Bi-Real Net, XNOR++, RTN, IRNet 等方法进行对比, 用以展现当前二值网络在图像超分辨率重建任务上的性能水平. 从图 4 我们可以看出, 相比于全精度网络, 所有二值网络在超分辨率任务上均取得较差的网络性能. 在 Set5 数据集上, 基于 BNN 方法的二值神经网络可取得 31.22 dB 的 PSNR 值, 与全精度网络相比性能下降约 1.5 dB, 其网络的特征映射能力被严重破坏. 而基于 XNOR, Bi-Real Net, XNOR++, RTN 和 IRNet 等方法的二值网络在 Set5 上分别取得 30.36, 29.18, 30.12, 30.25 和 29.11 dB 的 PSNR 值, 其在超分辨率任务上取得的网络性能与其在分类任务上取得的网络性能不相匹配. 相比于 XNOR, XNOR++ 中引入的可学习尺度因子无法带来性能上的增益甚至导致网络性能下降. 而相比于 BNN, Bi-Real Net 网络性能下降严重, 精心设计的二值化函数无法使超分辨率网络得到良好训练.

3.2.2 二值化方法在超分辨率网络上的性能研究

实验细节. 为进一步探索二值化算法在超分辨率领域的实际应用效果, 我们选取代表性的二值化算法应用于超分辨率网络 SRResNet^[42], 该网络为超分辨率领域中广泛使用的残差块级联的网络结构. 为保证实验对比的公平性, 选用的二值化方法均在相同的数据集、相同的网络结构和相同的实验设置下进行训练. 具体设置与 3.2.1 小节相同.

实验结果. 从图 5 可以看出, 相比于 Res18 网络, 基于 SRResNet 的二值网络性能上有较大增益. 例如, 在 Set5 数据集 4 倍放大下, BNN 与全精度 ResNet18 网络的性能差为 1.5 dB, 而基于 BNN 方

图 4 (网络版彩图) 基于 ResNet18 网络的不同二值化方法在 4 个基准测试集上的超分辨率性能, 测试集分别为 (a) Set5, (b) Set14, (c) B100, (d) Urban100

Figure 4 (Color online) Super-resolution performance on different benchmark datasets of different binary methods on ResNet18. The benchmark datasets: (a) Set5; (b) Set14; (c) B100; (d) Urban100

图 5 (网络版彩图) 基于 SRResNet 网络的不同二值化方法超分辨率性能

Figure 5 (Color online) Super-resolution performance of different binary methods on SRResNet. (a) Set5; (b) Set14; (c) B100; (d) Urban100

法的 SRResNet 网络与全精度 SRResNet 网络之间的性能差为 0.74 dB, 将二值化算法嵌入超分辨率网络更适合实现高性能的二值超分辨率网络. 而且其在 BNN, XNOR, Bi-Real Net, XNOR++, RTN,

IRNet 上分别取得 31.22, 31.29, 30.66, 31.17, 31.39, 和 31.32 dB. 网络的性能增益大致与其在分类任务上的性能相匹配, 证明基于前向过程和反向过程改进的策略的确对二值网络的性能有所增益. 然而, 基于 Bi-Real Net 方法的二值超分辨率网络在 PSNR 上下降 1.36 dB, 基于该方法改进的反向优化在超分辨率任务中存在一定优化问题. 而基于 IRNet 方法的二值超分辨率网络可取得一个较好的性能, 证明两阶段的近似在减少反向信息损失上更有帮助. 然而, 基于 BNN, XNOR, XNOR++, RTN, IRNet 等方法的二值超分辨率网络在 PSNR 上性能下降约 0.6~0.8 dB, 二值超分辨率重建网络与全精度网络相比依然存在较大性能差.

3.3 二值超分辨率重建网络性能分析

本小节我们进一步分析二值超分辨率网络性能下降的原因. 基于上述实验结果, 我们将模型性能下降主要归结于以下两方面原因: 一是现有二值化网络对于超分辨率任务的结构不适用问题; 二是二值化本身带来的信息损失较大.

(1) 结构误差. 目前二值化算法多为高层视觉任务 (如图像分类、目标检测等) 设计, 主要基于 ResNet 网络进行实验验证. 而目前直接针对底层视觉任务 (如图像超分辨率重建等) 设计的二值化方法较少. 高层视觉任务依赖语义信息的精准度, 而底层视觉任务则依赖像素级别的精准度. 因此, 由于任务的不同, 网络的结构设计多有不同. 例如, 图像分类网络中大多需要池化层和全连接层提取图像特征, 而图像重建任务中网络完成图像到特征再到图像的映射, 网络中不包含池化等操作. 此外, 现有超分辨率方法多结合全局跳跃连接与短距离跳跃连接实现图像低频信息的传输. 因此, 采用图像分类领域中广泛使用的二值网络无法在超分辨率任务上取得良好的实现结果.

(2) 量化误差. 二值神经网络将网络中的参数约束为 -1 或 $+1$, 严重限制了网络的表达能力, 其影响主要可分为两方面: 一是在网络前向信息传递过程中产生量化误差. 二值权值与激活卷积后的特征值将被约束在固定的范围区间, 降低了网络的多样性表达. 尽管尺度因子或跳跃连接的引入可以补偿部分精度, 但网络的表达能力依然有限, 限制了网络对于超分辨率任务中高频细节信息的恢复. 二是网络反向过程中产生的梯度误差. 由于二值化函数导数处处为 0, 普通的基于梯度下降的反向传播算法不能直接用于更新二值化参数. 尽管目前提出了多种解决策略, 但无论早期反向传播过程中使用的 STE 还是最新用于二值化函数近似的高阶二值化函数, 都不可避免地会带来梯度不匹配问题, 影响网络充分训练.

4 改进的二值超分辨率重建网络

基于第 3 节中对于二值化超分辨率网络的性能分析, 我们将从减少量化误差与结构误差两个方面对现有二值超分辨率重建算法进行改进, 致力于最小化二值网络与全精度网络之间的性能差.

4.1 二值超分辨率网络结构

首先, 基于 SRResNet, 我们重新定义了二值超分辨率网络的基础结构. 网络大致可分为 3 部分: 浅层特征提取、非线性映射、图像恢复. 令 x 表示超分辨率网络的输入, y 表示超分辨率网络的输出, 图像超分辨率重建过程可表示为

$$y = \mathcal{R}(\mathcal{M}(\mathcal{E}(x))), \quad (12)$$

其中, $\mathcal{E}$, $\mathcal{M}$ 和 $\mathcal{R}$ 分别表示上述提及的 3 个过程, 浅层特征主要为一个单独的卷积操作, 完成对输入图片的特征提取. 中间的非线性映射过程由多个残差块构成, 将低分辨率的特征映射为高分辨率特征,

图 6 (网络版彩图) 二值 SRResNet 网络结构

Figure 6 (Color online) System architecture of the binary SRResNet

这部分也决定了网络的复杂程度和内存占用空间. 图像恢复模块根据学习到的高分辨率特征进行图像重建. 值得注意的是, 网络中包含一个长距离的跳跃连接. 沿用现有二值化算法思想, 我们对网络中的非线性映射过程进行二值化操作. 为使二值化算法更适用于超分辨率网络, 我们重新定义了二值残差块的具体结构, 如图 6 所示. 对于二值卷积操作, 我们的设置为

$$I_{n+1} = BN(F_A(C_{B2}(A_B(BN(F_A(C_{B1}(A_B(I_n))))))), \quad (13)$$

其中, I_n 与 I_{n+1} 分别表示残差块的输入与输出, A_B 表示输入激活的二值化过程, C_{B1} , C_{B2} 表示二值化卷积, F_A 和 BN 表示激活函数和正则化.

4.2 二值超分辨率算法

基于此定义, 对于二值化算法的前向过程, 我们提出了一个特定的结构来弥补量化误差, 提升网络的多样性表达. 为此, 我们引入两个可变的尺度因子混合二值网络基础残差块的输入与输出:

$$I'_{n+1} = \lambda \times I_n + \beta \times I_{n+1}, \quad (14)$$

其中, I'_{n+1} 表示混合后的输出. 现有方法证明了增益项和短距离跳跃连接的引入可以在一定程度上减少前向过程中的量化损失. 然而, 固定的尺度因子会将卷积后的参数限制在固定的区间, 网络的多样性表达依然受到限制. 而跳跃连接的引入或二元基的线性组合会增加网络的计算消耗, 影响二值化算法最为突出的压缩性能. 考虑这些限制, 我们的方法可以在不增加过多二值卷积计算复杂度的前提下, 利用前后信息的融合补偿量化损失, 提高了网络的多样性表达. 通过此方法, 我们可以更好地捕获网络中的空间信息, 获取更具特征表达的网络信息.

第 3 节的实验表明, 一个导数更近似于 sign 函数的二值化函数更有利于二值网络反向过程的权值优化. 因此, 我们提出了一个新的二值化函数解决梯度不匹配的问题. 二值化函数具体如下:

$$F(x) = \begin{cases} -1, & \text{if } x < -0.5, \\ -4x^3 + 3x, & -0.5 \leq x \leq 0.5, \\ 1, & \text{otherwise,} \end{cases} \quad \frac{\partial F(x)}{\partial x} = \begin{cases} -12x^2 + 3x, & \text{if } -0.5 < x < 0.5, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

如图 7 所示, 同 Bi-Real Net 方法类似, 但我们减小了其截断区间以实现更准确的对于 sign 函数导数的近似, 在二值化算法反向优化中得以提供更精确的训练过程. 此外, 我们将中间的导数近似曲线由原本的两阶段转换为一阶段, 训练过程更为简单有效.

图 7 改进的二值化函数 (a) 与其导数 (b)

Figure 7 The improved binarization function (a) and the derivative of it (b)

表 2 本文方法与现有二值超分辨率方法的定量比较

Table 2 Quantitative comparison of our method and existing binary super-resolution methods

Method	Scale	W (bit)	A (bit)	Params (K)	OPs (10^9)	PSNR/SSIM			
						Set5	Set14	B100	Urban100
Full-Precision	$\times 2$	32	32	1399	332.16	37.76/0.958	33.25/0.914	31.94/0.893	31.17/0.917
Bicubic	$\times 2$	32	32	—	—	33.66/0.930	30.24/0.869	29.56/0.843	26.88/0.840
Ma et al.	$\times 2$	1	32	1399	332.16	35.66/0.946	31.56/0.897	—/—	28.76/0.882
BAM	$\times 2$	1	1	1399	61.11	37.21/0.956	32.74/0.910	31.60/0.891	30.20/0.906
Ours	$\times 2$	1	1	1399	61.11	37.40/0.957	32.96/0.911	31.75/0.892	30.53/0.910
Full-Precision	$\times 3$	32	32	1584	172.76	34.14/0.923	30.11/0.835	28.96/0.798	27.65/0.839
Bicubic	$\times 3$	32	32	—	—	30.39/0.868	27.55/0.774	27.21/0.739	24.46/0.735
BAM	$\times 3$	1	1	1584	53.90	33.33/0.915	29.63/0.827	28.61/0.790	26.69/0.816
Ours	$\times 3$	1	1	1584	53.90	33.69/0.919	29.80/0.831	28.76/0.793	27.05/0.824
Full-Precision	$\times 4$	32	32	1547	127.75	31.99/0.890	28.43/0.778	27.47/0.730	25.74/0.778
Bicubic	$\times 4$	32	32	—	—	28.42/0.810	26.00/0.703	25.96/0.668	23.14/0.658
Ma et al.	$\times 4$	32	32	1547	127.75	30.34/0.864	27.16/0.756	—/—	24.48/0.728
BAM	$\times 4$	32	32	1547	60.01	31.24/0.878	27.97/0.765	27.15/0.719	24.95/0.745
Ours	$\times 4$	32	32	1547	60.01	31.51/0.883	28.15/0.770	27.27/0.723	25.22/0.753

4.3 实验评估

实验细节. 为保证实验对比的公平性, 本小节所选用的数据集和训练细节与 3.2.1 小节相同. 此外, 本小节实验包括 scale $\times 2$, $\times 3$ 与 $\times 4$ 超分辨率任务.

实验结果. 首先, 表 2 对本文方法和现有二值超分辨率方法进行了定量比较, 在 Set5 数据集 2 倍放大下, 我们的方法可以取得 37.40 dB 的 PSNR 值, 相比于 Ma 等^[35] 的工作高出近 1.7 dB 的 PSNR 值, 相比于 Xin 等^[36] 的 BAM 方法高出近 0.2 dB 的 PSNR 值. 在 Urban100 数据集 4 倍放大下, 我们的方法可以取得 25.22 dB 的 PSNR 值, 相比于 Ma 等的工作高出 0.74 dB 的 PSNR 值, 相比于 Xin 等的工作高出近 0.3 dB 的 PSNR 值. 实验结果证明了我们的方法明显优于目前二值超分辨率方法. 此外, 我们在表格中同样展示了对比方法的细节信息, Params 和 OPs 分别表示方法的模型参数和乘法计算量, 对于 M -bit 网络, $OPs = BOPs/64 \cdot M + FLOPs$, 其中, BOPs 是二值运算次数, FLOPs 是浮

图 8 (网络版彩图) 本文方法与现有方法的定性比较

Figure 8 (Color online) Qualitative comparison between our method and existing methods

点数运算次数. 从表中可以看出, 相比于全精度网络, 二值网络的模型参数数量并没有发生明显变化, 但在硬件存储时, 1-bit 模型享有 32 倍的模型压缩比. 与 Ma 等的工作相比, 我们的方法不仅实现了网络权值的二值化, 同时实现了网络运行过程中激活值的二值化. 因此我们的方法享有更高的速度增益比, 更适用于手机等便携式移动设备. 与 Xin 等的工作相比, 我们的方法可以在模型运行时减少 16 次全精度特征图相加操作, 同时可以取得更优异的网络性能. 此外, 在 Set5 数据集 4 倍放大下, 我们方法可取得 31.51 dB 的 PSNR 值, 同样高于目前基于分类任务的最优二值化算法 (RTN 31.39 dB). 实验结果充分表明我们的二值化方法更适用于图像超分辨率重建任务. 图 8 给出了本文方法与现有二值方法的定性比较, 从图中可以看出, 我们的方法展现出更清晰的细节纹理, 取得与全精度网络相近的视觉效果.

5 问题与挑战

二值化超分辨率网络的研究目前还处于起步阶段, 目前仍未解决二值网络与全精度网络之间巨大

的性能差异, 且现有方法在二值化效率方面显示出了一定的局限性, 有以下待解决的问题与挑战.

(1) 算法性能. 目前二值超分辨率算法相对简单, 网络的前向过程和反向过程中仍然存在较多信息损失, 导致二值化网络与全精度网络之间存在较大的性能差异, 还不足以实现二值网络对全精度网络的替代. 因此, 有必要针对二值超分辨率任务进行深入剖析, 设计特定的网络结构或构建新的二值化函数, 提高量化精度, 较少量化损失. 与此同时, 也应考虑二值网络所具有的高度离散特性, 思考更适合二值网络训练的特殊策略, 以帮助网络实现更稳定和高效的训练过程.

(2) 全精度上采样模块. 目前二值化超分辨率网络仅二值化网络中的非线性映射模块, 模型中的上采样模块依然采用全精度操作. 然而, 这种方法通常需要额外的存储和计算量, 降低了二值网络硬件实施的有效性. 因此, 有必要针对二值超分辨率网络设计特有的上采样模块, 以期实现二值化的采样进程, 提高网络的运行效率.

(3) 全精度 BN 操作. 目前无论是基于分类任务的二值化算法还是基于超分辨率的二值化算法大都采用 BN 层缓解量化过程产生的影响. BN 层可重新归一化卷积后的特征分布, 有效降低量化过程产生的误差. 然而 BN 的存在会在运行过程中引入全精度浮点型操作, 影响模型推理速度. 尽管现有全精度超分辨率网络在移除 BN 后依然可以取得较高性能, 而直接去除二值网络中的 BN 将导致性能严重下降. 因此, 有必要针对二值网络的特征分布进行深度研究, 希望通过更为简单的操作实现 BN 层对特征分布的重新调整, 减少网络中的全精度运算, 实现更快的模型推理速度.

6 结束语

本文提出了一个新的二值化超分辨率重建方法, 首先对现有二值神经网络进行了系统综述, 详细介绍了算法的技术细节和方法性能. 其次, 本文从实验角度分析了基于不同二值化算法的超分辨率网络性能, 深入剖析其方法利弊. 进而, 我们进一步挖掘二值超分辨率网络的潜力, 提出了一个特有的网络结构和二值化函数来提高二值网络前向过程中的表达能力, 减少二值网络反向过程中的信息损失, 并实验分析其算法性能. 最后, 本文介绍了二值超分辨率重建领域面临的问题和未来研究中的挑战.

参考文献

- 1 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105
- 2 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 3 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
- 4 Huang G, Liu Z, van Maaten L D, et al. Densely connected convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 4700–4708
- 5 Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 1135–1143
- 6 Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems, 2016. 2074–2082
- 7 Sindhwani V, Sainath T, Kumar S. Structured transforms for small-footprint deep learning. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 3088–3096
- 8 Cheng Y, Yu F X, Feris R S, et al. An exploration of parameter redundancy in deep networks with circulant projections. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 2857–2865
- 9 Rastegari M, Ordonez V, Redmon J, et al. XNOR-Net: ImageNet classification using binary convolutional neural networks. In: Proceedings of European Conference on Computer Vision, 2016. 525–542

- 10 Courbariaux M, Bengio Y, David J P. Binaryconnect: training deep neural networks with binary weights during propagations. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 3123–3131
- 11 Bucilu C, Caruana R, Niculescu-Mizil A. Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006. 535–541
- 12 Ba J, Caruana R. Do deep nets really need to be deep? In: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014. 2654–2662
- 13 Govindu G, Zhuo L, Choi S, et al. Analysis of high-performance floating-point arithmetic on FPGAs. In: Proceedings of the 18th International Parallel and Distributed Processing Symposium, 2004. 149–157
- 14 Greenspan H. Super-resolution in medical imaging. *Comput J*, 2008, 52: 43–63
- 15 Huang Y, Shao L, Frangi A F. Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6070–6079
- 16 Lin F, Fookes C, Chandran V, et al. Super-resolved faces for improved face recognition from surveillance video. In: Proceedings of the International Conference on Biometrics, 2007. 1–10
- 17 Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution. In: Proceedings of European Conference on Computer Vision, 2014. 184–199
- 18 Jiang X, Wang N, Xin J, et al. Image super-resolution via multi-view information fusion networks. *Neurocomputing*, 2020, 402: 29–37
- 19 Xin J, Li J, Jiang X, et al. Wavelet-based dual recursive network for image super-resolution. *IEEE Trans Neural Netw Learn Syst*, 2020. doi: 10.1109/TNNLS.2020.3028688
- 20 Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems, 2016. 4107–4115
- 21 Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation. 2013. ArXiv:1308.3432
- 22 Qin H, Gong R, Liu X, et al. Binary neural networks: a survey. *Pattern Recogn*, 2020, 105: 107281
- 23 Zhou S, Wu Y, Ni Z, et al. DoReFa-Net: training low bandwidth convolutional neural networks with low bandwidth gradients. 2016. ArXiv:1606.06160
- 24 Lin X, Zhao C, Pan W. Towards accurate binary convolutional neural network. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2017. 345–353
- 25 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, 2017. 2584–2592
- 26 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. In: Proceedings of European Conference on Computer Vision, 2018. 722–737
- 27 Bulat A, Tzimiropoulos G. XNOR-Net++: improved binary neural networks. 2019. ArXiv:1909.13863
- 28 Li Y, Dong X, Zhang S Q, et al. RTN: reparameterized ternary network. In: Proceedings of the Association for the Advance of Artificial Intelligence, 2020. 4780–4787
- 29 Liu Z, Shen Z, Savvides M, et al. ReActNet: towards precise binary neural network with generalized activation functions. In: Proceedings of European Conference on Computer Vision, 2020. 143–159
- 30 Martinez B, Yang J, Bulat A, et al. Training binary neural networks with real-to-binary convolutions. 2020. ArXiv:2003.11535
- 31 Lahoud F, Achanta R, Márquez-Neila P, et al. Self-binarizing networks. 2019. ArXiv:1902.00730
- 32 Darabi S, Belbahri M, Courbariaux M, et al. BNN+: improved binary network training. 2018. ArXiv:1812.11800
- 33 Qin H, Gong R, Liu X, et al. Forward and backward information retention for accurate binary neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 2250–2259
- 34 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
- 35 Ma Y, Xiong H, Hu Z, et al. Efficient super resolution using binarized neural network. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019. 1–10

- 36 Xin J, Wang N, Jiang X, et al. Binarized neural network for single image super resolution. In: Proceedings of European Conference on Computer Vision, 2020. 91–107
- 37 Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 126–135
- 38 Bevilacqua M, Roumy A, Guillemot C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of British Machine Vision Conference, 2012. 135: 1
- 39 Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. In: Proceedings of International Conference on Curves and Surfaces, 2010. 711–730
- 40 Huang J B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 5197–5206
- 41 Martin D, Fowlkes C, Tal D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of IEEE International Conference on Computer Vision, 2001. 2: 416–423
- 42 Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4681–4690

Binary neural networks for image super-resolution

Xinrui JIANG¹, Nannan WANG^{1*}, Jingwei XIN¹, Keyu LI¹, Xi YANG¹ & Xinbo GAO²

1. *State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China;*
 2. *Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*

* Corresponding author. E-mail: nnwang@xidian.edu.cn

Abstract Recently, deep convolutional neural networks (DNNs) have achieved state-of-the-art performance on the super-resolution task. However, bigger and deeper networks often lead to high computational cost and high memory usage, preventing massive applications on resource-limited devices. Therefore, reducing the model storage and computation costs will expand the application range of DNNs. Binary neural networks, which could reduce the model size and allow for efficient inference, are energy-efficient for embedded devices. It motivates us to apply the binary methods into super-resolution field, enabling us to meet the requirements of the hardware platforms in practical applications. Therefore, this paper focuses on the binary image super-resolution network. To this end, we first review the current binary neural networks and introduce the technical details and algorithm characteristics of the existing algorithms in detail. Then, we propose to apply the existing binary training methods to the super-resolution task. On this basis, we propose a new binary super-resolution method, which advances network performance by improving the representation ability of the quantized network in the forward propagation and reducing information loss in the backward propagation. Extensive experiments demonstrate that the proposed method not only outperforms existing binary methods but also outperforms the state-of-the-arts binary super-resolution methods.

Keywords binary neural network, single image super-resolution, binarization, quantization, model compression

Xinrui JIANG received her B.S. degree in electronic and information engineering from Shandong University, Jinan, China, in 2018. She is currently pursuing her Ph.D. degree at the School of Telecommunications Engineering, Xidian University. Her current research interests include computer vision, pattern recognition, and machine learning.

Nannan WANG received his B.S. degree in information and computation science from Xian University of Posts and Telecommunications in 2009. He received his Ph.D. degree in information and telecommunications engineering from Xidian University in 2015. Now, he works with State Key Laboratory of Integrated Services Networks at Xidian University. From September 2011 to September 2013, he has been a visiting Ph.D. student at the University of Technology, Sydney, NSW, Australia. His current research interests include computer vision, pattern recognition, and machine learning.

Jingwei XIN received his B.S. degree in electronic and information engineering from Shaanxi Normal University, Xi'an, China, in 2016. He is currently pursuing his Ph.D. degree in intelligent information processing at the VIPs Laboratory, School of Electronic Engineering, Xidian University, Xi'an, China. His current research interests include computer vision and machine learning.

Xinbo GAO received his B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been a professor at School of Electronic Engineering, Xidian University. He is currently a Cheung Kong professor of Ministry of Education, and the director of the State Key Laboratory of Integrated Services Networks, Xi'an, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.