文章编号：1009 - 2552(2020)10 - 0087 - 05　　DOI：10.13274/j.cnki.hdzj.2020.10.016

基于参数量化的轻量级图像压缩神经网络研究

孙浩然[1]，王　伟[2]，陈海宝[1]

(1. 上海交通大学微电子学院，上海 200240；2. 北京宇航系统工程研究所，北京 100076)

摘　要： 随着深度学习的发展，神经网络模型参数的数量越来越大，消耗了大量的存储与计算资源。而在面向图像压缩应用的自编码神经网络中，其编码器网络和解码器网络往往占用着更大的存储空间。因此，文中提出了一种基于参数量化的轻量级图像压缩神经网络，采用训练中参数量化的方法将模型参数从 32 位浮点型量化到 8 位整型。实验结果表明，相比原始模型，提出的轻量级图像压缩神经网络模型节约了 73% 的存储空间。在图像压缩码率小于 0.16bpp 的条件下，重建图像的多尺度结构相似度指标 MS-SSIM 仅损失 1.68%，依然优于经典压缩标准 JPEG 与 JPEG2000。

关键词： 参数量化；模型压缩；图像压缩；神经网络

中图分类号： TP183　　**文献标识码：** A

Lightweight image compression neural network based on parameter quantization

SUN Hao-ran[1], WANG Wei[2], CHEN Hai-bao[1]

(1. School of Microelectronics, Shanghai Jiaotong University, Shanghai 200240, China;
2. Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China)

Abstract: With the development of Deep Learning, the number of neural network model parameters is getting larger and larger, consuming a large amount of storage and computing resources. In the auto-encoder neural network for image compression applications, the encoder network and decoder network often occupy more storage space. Therefore, a lightweight image compression neural network model is proposed based on parameter quantization. The method of aware training quantization is used to quantize the model parameters from 32-bit floating point to 8-bit integer. The experimental results show that compared to the original model, the proposed lightweight image compression neural network model saves 73% of storage space. When the image compression bit-rate is less than 0.16bpp, MS-SSIM (multi-scale structure similarity) of the reconstructed image just loses 1.68%, and is still higher than the classic compression standards JPEG and JPEG2000.

Key words: parameter quantification; model compression; image compression; neural network

0　引言

随着计算机硬件设备和神经网络的共同发展，人们为了追求更高的网络性能，往往通过加深网络层数的方法提升神经网络的非线性表达能力。但是，网络层数的加深使得网络的占用存储越来越大，消耗了大量的内存和计算资源。因此，利用模型压缩方法降低模型存储成为了热门研究课题。但是，现有模型压缩技术多应用于图像分类、目标检测等神经网络，应用于图像压缩神经网络的模型压缩研究极少，而图像压缩神经网络模型往往占用着更大的存储空间。

收稿日期：2020 - 03 - 24

作者简介：孙浩然(1993 -)，男，硕士研究生，研究方向为基于深度学习的图像压缩算法。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

为了实现图像压缩的功能，图像压缩神经网络[1-3]常采用卷积自编码网络获得图像的压缩表示。在卷积自编码网络中，首先通过编码器网络对图像降采样，得到图像的压缩表示；然后通过解码器网络对压缩表示上采样，恢复信息并重建图像。编解码器的双网络对称结构使得面向图像压缩的神经网络层数更深，占用着更大的存储空间，因此更需要模型压缩。

为了降低图像压缩神经网络的存储并保证生成图像的质量，本文提出了基于参数量化的轻量化图像压缩神经网络，采用了文献[4]和文献[5]提出的参数量化方法对网络进行模型压缩，该方法通过将模型权重从32位浮点型量化到8位整型的方式降低模型存储。本文探讨了模型训练收敛后参数量化与模型训练中参数量化两种量化方法，并通过图像质量评价指标衡量二者对神经网络模型生成图像质量的影响。

1 图像压缩神经网络基础模型

传统图像压缩算法如JPEG[6]、JPEG2000[7]压缩标准由于其分块、线性变换、量化变换后高频分量系数的编码方式，在低码率压缩图像的条件下，其解压图像会产生块状失真与振铃效应[8]。基于神经网络的图像压缩算法可以有效解决传统算法的失真问题，神经网络作为一种非线性模型，可以代替传统算法的编解码方式，获得图像的压缩表示并根据压缩表示重建图像。此外，神经网络作为参数可学习的优化模型，可以针对特定的质量评价指标对模型进行端到端的优化训练。

本文的模型压缩研究基于如图1所示的图像压缩算法，该算法基于全卷积自编码网络，由编码器网络、量化器、熵编解码器、解码器网络五个模块组成。编码器由卷积层组成，通过步长为2的卷积对图像降采样；量化器将特征图由32位浮点型量化为L位整型；熵编解码器采用算术编码对特征图无损压缩为二进制码流，二进制码流所占用比特位数即压缩存储大小；解码器由卷积层组成，通过变换卷积对特征图上采样，还原图像信息。本文基础模型名称为64_64_C4_L6，64_64表示训练时尺寸为512×512的输入图像经过编码器降采样后输出特征图的尺寸，C表示特征图通道数，L表示特征图中每个数值的量化位数。本文图像压缩神经网络主体由卷积层构成，编码器网络与解码器网络都包含34层卷积层，共68层卷积层，因此量化对象为68层卷积层所包含的卷积核与偏置。

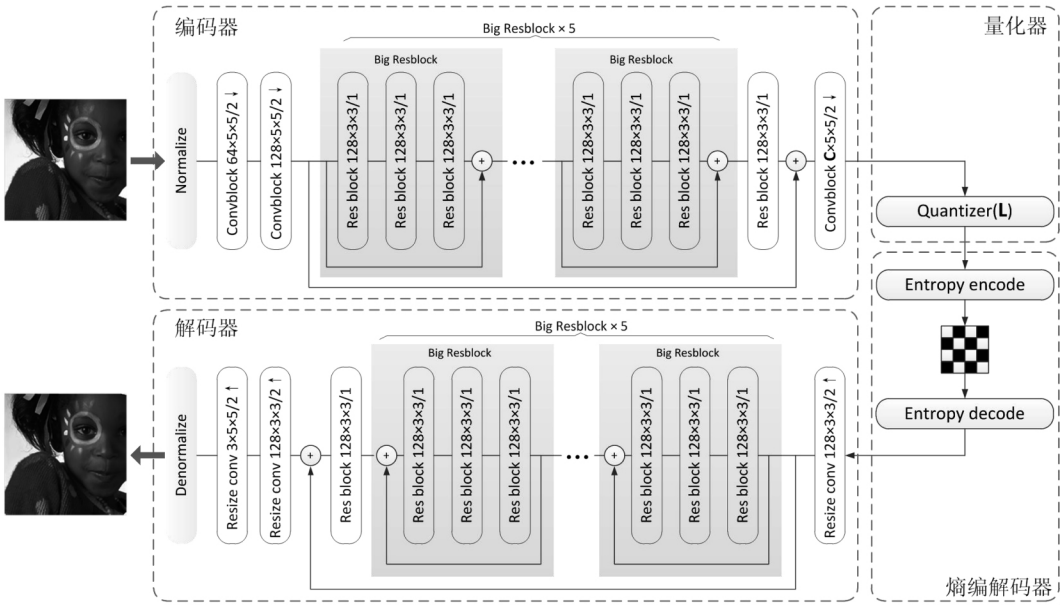


图1 图像压缩神经网络基础模型示意图

2 基于参数量化的轻量化图像压缩模型

2.1 训练后参数量化算法

在上述图像压缩神经网络模型的基础上，本文采用文献[4]中提出的参数量化方法，将卷积层中的卷积核权重与偏置从32位浮点型量化至8位整型。训练后量化是指在模型收敛、参数已固定的状态下对模型中的卷积核、偏置等参数进行量化。在模型存储过程中，通过存储量化后的8位整型参数达到模型压缩的效果；在推理过程中，通过量化、反量化的方式仍以浮点型精度进行运算。

训练后量化包括量化与反量化两个过程。首

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

先，量化过程采用逐层非对称的量化方式，即以卷积层为单位，量化区间范围不以零点作为量化中心。设待量化参数为 x，并统计参数 x 中最小值为 $x_{\min}$，最大值为 $x_{\max}$，设量化比特数为 q，则量化级别 n 由下式计算可得：

$$n = 2^q - 1 \tag{1}$$

得到量化前参数区间为 $[x_{\min}, x_{\max}]$，量化后参数区间为 $[0, n]$，则量化尺度 Δ 由下式计算可得：

$$\Delta = \frac{x_{\max} - x_{\min}}{n - 0} \tag{2}$$

设 x 量化后的整型参数为 x_{int}，则 x_{int} 由下式计算可得：

$$x_{int} = clamp(0, n, round((x - x_{\min})/\Delta)) \tag{3}$$

$$clamp(a, b, x) = \begin{bmatrix} a & x \leqslant a \\ x & a \leqslant x \leqslant b \\ b & x \geqslant b \end{bmatrix} \tag{4}$$

其中，*round* 函数将尺度变换后的浮点型数值量化为数值最近的整型数值。在第一次参数量化过程中，存储浮点型参数最大值 $x_{\max}$、最小值 $x_{\min}$ 和整型量化参数 x_{int} 到量化模型，模型参数格式由 32 位浮点型转为 8 位整型，占用空间约减小为原来的四分之一。

反量化过程是将量化整型参数恢复为浮点型参数，用于模型推理。设反量化后得到的浮点型参数为 x_{float}，则 x_{float} 由下式计算可得：

$$x_{float} = (x - x_{int}) \times \Delta + x_{\min} \tag{5}$$

以上为参数量化、反量化的过程描述，本文量化对象为编解码网络每一层卷积层中的卷积核与偏置参数，设该卷积层中的所有卷积核参数为 w，偏置为 b，量化位数为 q，基于 64_64_C4_L6 模型的训练后量化步骤如下所示。

算法：训练后参数量化

输入 卷积层卷积核 w

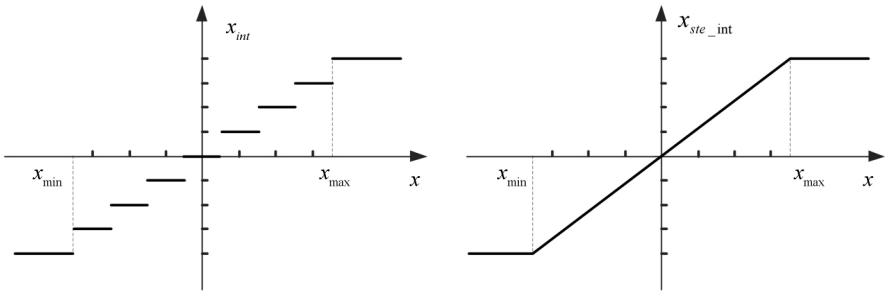
输入 卷积层偏置 b

输入 量化比特数 q

1. **最大最小值：** $w_{\max} = \max(w)$，$w_{\min} = \min(w)$，$b_{\max} = \max(b)$，$b_{\min} = \min(b)$
2. **量化：由式**(1) – (3) **得** w_{int} **和** b_{int}
3. **输出：存储** w_{int}，$w_{\min}$，$w_{\max}$ **和** b_{int}，$b_{\min}$，$b_{\max}$
4. **反量化：由式**(5) **得** w_{float} **和** b_{float}

2.2 训练中参数量化算法

本文采用文献[5]中提出的训练中参数量化算法在模型训练时加入了对卷积核、偏置参数的量化与反量化步骤，量化与反量化参与了训练的前向推理阶段与反向传播阶段。在前向推理阶段，训练中参数量化的原理与训练后量化相同。但是在误差反向传播阶段，如图 2(a) 所示，由于 round 这类硬阈值量化函数的导数几乎处处为零，根据链式求导法则将导致参数的梯度处处为零，不利于模型参数的训练与更新。



(a) *round*量化函数　　(b) 直通估计器函数

图 2 *round* 量化函数与直通估计器函数图像

因此在模型训练阶段，本文采用了文献[5]中提出的直通估计器(straight through estimator)替代 round 量化函数进行误差的反向传播与参数的梯度计算，直通估计器函数表达式如下所示：

$$x_{ste\_int} = clamp(0, n, (x - x_{\min})/\Delta) \tag{6}$$

直通估计器的函数图像如图 2(b) 所示，相比 round 量化方式，直通估计器函数在反向传播过程中保证了梯度不为零，使误差效地反向传播并按照梯度更新参数。本文采用微调的方式获得训练中参数量化模型，首先加载已经收敛的 64_64_C4_L6 基础模型权重，然后加入如量化、反量化、直通估计器函数模拟量化的操作，在此基础上开始训练模型直至再次收敛。

3 结果分析

本文实验采用单个英特尔 Xeon(R) E5-2698 v3 CPU 和单块英伟达 1080Ti GPU 的服务器作为模型训练和测试的硬件平台，在实验中，学习率设置为 0.0001，训练约 116 个周期，每个周期包含 10343 次

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

迭代。经过总共 1199788 次迭代模型训练结束，整个训练流程持续约 100 小时，平均每次迭代约 0.30 秒。实验中采用 CLIC [10]（workshop and challenge on learned image compression）图像压缩竞赛数据集进行训练和验证，训练集包含 1048 张不同尺寸的 24 位真彩色图像，被裁减为 10343 张尺寸为 512×512 的图像进行训练。以 Kodak 数据集 [11] 作为测试集，对模型的压缩性能指标进行评估。

3.1 训练中量化与训练后量化模型对比

本文原始模型大小为 39.86MB，8bit 量化后模型大小为 10.70MB，是原始模型存储的 26.8%，节约了约 3/4 的存储空间。本文以模型生成图像的图像质量评价指标来衡量参数量化后模型的精度损失，评价指标包括图像压缩码率 BPP（bit per pixel）、多尺度结构相似度 MS-SSIM（multi-scale structural similarity index）、结构相似度 SSIM（structural similarity index）、峰值信噪比 PSNR（peak signal to noise ratio）。原始模型、训练后量化模型、训练中量化模型在 Kodak 数据集上 24 张图像的平均测试结果如表 1 所示。

表 1 训练后量化、训练中量化与原始图像压缩网络模型性能比较

method	BPP	MS-SSIM	SSIM	PSNR
64_64_C4_L6	0.1591	0.9285	0.7229	22.56
64_64_C4_L6 训练后量化	0.1577	0.5681	0.4420	14.74
64_64_C4_L6 训练中量化	0.1451	0.9129	0.6704	19.83

与原始模型 64_64_C4_L6 相比，观察计算两种量化方式带来的精度损失。在压缩码率接近的情况下，训练后量化模型的 MS-SSIM 指标由 0.9285 下降为 0.5681，降低 38.8%；SSIM 由 0.7229 下降为 0.4420，降低 38.9%；PSNR 由 22.56 下降为 14.74，降低 34.7%。反观训练中量化模型的性能表现，MS-SSIM 指标由 0.9285 下降为 0.9129，仅降低 1.68%；SSIM 由 0.7229 下降为 0.6704，降低 7.26%；PSNR 由 22.56 降低为 19.83，降低 12.1%。

再将两种量化方式的性能指标对比，相比训练后量化模型，训练中量化模型将 MS-SSIM 指标提高 60.7%，SSIM 提高 51.7%，PSNR 提高 34.5%。

如图 3－4 所示，从视觉角度对比原始模型、训练后量化模型、训练中量化模型三种方法生成的图像质量。相比原始模型，训练后量化模型生成的图像丢失了大量纹理、结构细节，并且出现了严重颜色偏差；但是对于训练中量化模型，肉眼较难分辨出纹理、结构细节的丢失和颜色偏差。

综上，采用逐层非对称的训练后参数量化方法并不适用于本文提出的图像压缩神经网络基础模型，其训练后对卷积核权重和偏置量化累积的误差将导致生成图像质量极差，但是训练中参数量化的方法可以通过端到端的训练微调模型参数，再次优化图像质量指标，因而具有更高的图像质量评价指标和视觉效果。

3.2 训练中量化模型与 JPEG、JPEG2000 对比

为了验证训练中参数量化模型的性能，本文从压缩算法在 Kodak 测试集上的性能指标、生成图像的视觉效果两个角度，将训练中量化模型与传统图像压缩算法 JPEG [12]、JPEG2000 [13] 进行比较。

首先，测试训练中参数量化模型、JPEG 与 JPEG2000 三种压缩算法在 Kodak 数据集上的性能。如表 2 所示，与 JPEG、JPEG2000 相比，本文提出的轻量化神经网络，即 64_64_C4_L6 训练中量化模型，以更低的压缩码率 0.1451bpp 取得了更高的 MS-SSIM 指标 0.9129。

表 2 训练中量化模型与 JPEG、JPEG2000 性能比较

method	BPP	MS-SSIM	SSIM	PSNR
JPEG	0.1715	0.8410	0.6696	24.79
JPEG2000	0.1591	0.9072	0.7225	26.90
64_64_C4_L6 训练中量化	0.1451	0.9129	0.6704	19.83

接下来，从算法生成图像的视觉效果对比训练中量化模型与 JPEG、JPEG 2000。如图 3 所示，以图片 kodim24 作为样张，本文提出的轻量化神经网络

图 3 多种压缩算法生成的 kodim24 图像

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

模型生成图像具有更清晰的视觉显示，JPEG、JPEG2000 则出现了对人眼不友好的块状失真与振铃效应。

为了进一步清晰地对比训练中量化模型与 JPEG、JPEG 2000 算法生成图像的视觉效果，本文放大图像的局部细节。如图 4 所示，观察图像 kodim24 中的房屋墙体上的彩绘，在 JPEG 的压缩图像中，太阳图案的轮廓、纹理细节损失严重，图像出现了多个分界明显的图像块，即块状失真；在 JPEG2000 算法中，太阳图案有模糊的轮廓，在燕子图案的边缘附近，出现了波纹状伪影，如同钟被敲击后震动发声引起的空气波纹，即振铃效应；在训练中量化模型中，太阳图案较为清晰，太阳的光线部分也较为分明，燕子图案的边缘清晰，附近没有波纹状伪影。

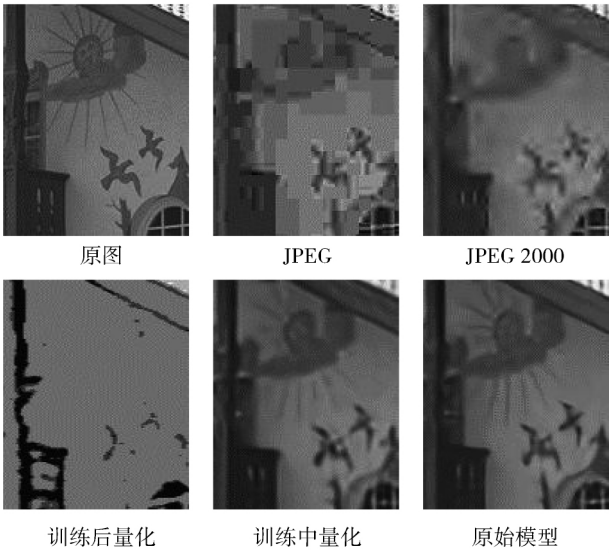


图 4　多种压缩算法生成的 kodim24 图像细节放大

4　结束语

①使用训练后参数量化方法对本文的图像压缩网络模型量化，相比原模型，会造成较大的精度损失，其 MS-SSIM、SSIM、PSNR 等指标均有不少于 30% 的降幅，并且其生成图像丢失大量纹理、结构细节，并呈现明显色差。

②使用训练中参数量化方法对本文的图像压缩网络模型量化，相比原模型，造成精度损失较小，基于人眼视觉感知的 MS-SSIM 指标仅下降 1.68%，其生成图像视觉上较难分辨纹理、结构等细节差别。

③本文提出的轻量化神经网络，即训练中量化模型，只占用原模型约 1/4 的存储空间，有效降低了存储开销，并且比现行压缩标准 JPEG、JPEG2000 呈现出更高的图像质量。

参 考 文 献:

[1] Theis L, Shi W, Cunningham A, et al. Lossy image compression with compressive autoencoders [C]. International Conference on Learning Representations (ICLR), 2017: 1 - 19.

[2] Zhou L, Cai C, Gao Y, et al. Variationalautoencoder for low bit-rate image compression [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2617 - 2620.

[3] Wen S, Zhou J, Nakagawa A, et al. Variational autoencoder based image compression with pyramidal features and context entropy model [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4321 - 4324.

[4] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2704 - 2713.

[5] Krishnamoorthi R. Quantizing deep convolutional networks for efficient inference: a whitepaper [EB/OL]. (2018 - 06 - 21) [2020 - 04 - 09]. https: //arxiv. org/abs/1806. 08342.

[6] Wallace G K. The JPEG still picture compression standard [J]. IEEE Transactions on Consumer Electronics, 1992, 38(1): xviii - xxxiv.

[7] Majid, Rabbani, Rajan, et al. An overview of the JPEG 2000 still image compression standard [J]. Signal Processing: Image Communication, 2002, 17(1): 3 - 48.

[8] Ballé J, Laparra V, Eero P. End-to-end optimized image compression [C]. International Conference on Learning Representations (ICLR), 2017: 1 - 27.

[9] Courbariaux M, Bengio Y, David J P. Binaryconnect: training deep neural networks with binary weights during propagations [C]. Advances in Neural Information Processing Systems, 2015: 3123 - 3131.

[10] The computer vision lab of ETH Zurich. Challenge data set [EB/OL]. (2018 - 6 - 21) [2020 - 4 - 9]. http: //www. compression. cc/challenge/.

[11] Franzen R. PhotoCD PCD0992 [EB/OL]. (2020 - 03 - 18) [2020 - 04 - 09]. http: //r0k. us/graphics/kodak/.

[12] Lane T, the Independent JPEG Group. Libjpeg [EB/OL]. (2009 - 06 - 22) [2020 - 04 - 09]. http: //libjpeg. sourceforge. net/.

[13] Kakadu Software Pty Ltd. Kakadu software [EB/OL]. (2020 - 04 - 09) [2020 - 04 - 09]. http: //kakadusoftware. com/.

责任编辑: 丁晓清

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net