METAGRAD：基于超网络的自适应梯度量化

Kaixin Xu[1,4], *Alina Hui Xiu Lee*[3], *Ziyuan Zhao*[1,2],*Zhe Wang*[1,2,4],*Min Wu*[1,2],*Weisi Lin*[4]

[1]资讯通信研究院，A*STAR，新加坡[2] 人工智能、分析与信息学（AI[3]），
A*STAR，新加坡[3] 新加坡国立大学[4]南洋理工大学计算机科学与工程学院，新加坡

arXiv:2303.02347v2 [cs.CV] 1 Nov 2023

摘要

网络压缩的一种常见方法是量化感知训练（QAT），它能够加速神经网络训练和推理过程中的前向传播。然而，尽管反向传播约占训练时间的一半，先前却鲜有研究尝试对训练期间的反向传播进行量化和加速。这在一定程度上是因为，在反向传播中，低精度梯度所产生的误差无法像 QAT 设置那样通过训练目标来摊销。本文提出一种解决方案：通过超网络，将梯度纳入下一次训练迭代的计算图中。在 CIFAR-10数据集上，使用不同卷积神经网络架构开展的多项实验表明，我们基于超网络的方法能够有效减轻梯度量化噪声的负面影响，并成功将梯度量化至 INT4；对于 CIFAR-10上的 VGG-16，准确率仅下降 0.64。

索引词— 网络压缩，网络量化，量化感知训练，梯度量化，卷积神经网络

1. 引言

模型压缩 [1, 2, 3, 4]是缩减深度模型大小和计算复杂度的重要技术。其中，量化感知训练（QAT）是一项热门研究方向，它在训练过程中模拟神经网络的量化（权重和激活值），以遏制低比特模型在推理时的精度下降（例如 INT8量化）。另一方面，理论和实证分析 [5, 6]表明，训练期间反向传播所需的计算量超过前向传播。通过在反向传播过程中量化梯度，可以利用低精度计算卷积核进一步加速量化感知训练模型的训练。

尽管这一潜在优势巨大，但与量化前向变量相比，量化反向梯度尤其困难，因为反向变量并未包含在计算图中，因此无法通过目标函数进行优化，这可能导致参数更新朝错误的方向偏移。过去已有少数关于梯度量化的研究 [7, 8]，主要通过设计适应梯度特性的梯度量化器，实现了稳定的 INT8 前向和反向量化。然而，由于上述困难，低于 INT8 的量化研究极少，早期唯一的尝试 DoReFa [1] 在 INT6 梯度下效果不佳。这表明，依靠启发式方法设计梯度量化器已达到极限。

为应对低比特梯度量化的挑战并遏制其造成的显著性能下降，本文提出了一种新颖的量化方法，能够适应任意量化噪声，并最大限度地保持稳定的训练方向。该元量化器通过引入超网络 [9], 实现；该超网络与网络一同进行端到端训练，根据给定的全精度梯度预测其量化版本。通过将低精度梯度纳入计算图，它们现在具备损失感知能力，从而实现更稳健的量化感知训练。所提出的元梯度量化器在多种卷积神经网络架构上取得了更高的测试性能。

2. 相关工作

2.1. 量化感知训练（QAT）

DoReFa-Net [1] 提出分别优化权重和激活值均匀量化器的裁剪值与缩放因子。该方法在多个位宽下通过图像分类任务进行了验证，但使用的架构仅为相当简单的 AlexNet。大多数量化感知训练工作通过逐层优化均匀量化参数，采用逐通道混合精度量化 [2, 13]，，或利用对数量化器 [14]等非均匀量化方法，同时量化权重和激活值。最近的大多数量化感知训练工作 [1, 11, 12] 采用“直通估计器”（STE） [15]来估计不可微量化函数的梯度，而另一项工作 [16] 则对线性量化操作进行软化处理

以便使用 STE 与真实梯度相匹配。

2.2. 量化感知训练中的梯度量化

早期的一项尝试 [1] 采用了基于梯度均匀量化器的简单量化器设计（不进行缩放或其他优化）；使用低位梯度训练时，观察到了显著的性能下降。SBM [6]采用定点 8 位梯度量化，但只专注于改进前向传播中的量化方案。WAGEUBN [17] 将梯度量化为 8 位整数，但与其全精度版本相比，性能差距仍然很大。在 UINT8 [7], 中，作者考虑到梯度分布尖锐且范围较宽，提出在量化前根据梯度分布的偏差对梯度进行裁剪，从而取得了与全精度训练相当的结果。为了弥补梯度量化造成的损失，AFP [18] 和 CPT [19] 使用更高精度数据来辅助低精度训练。DAINT8 [8]采用专门设计的 8 位逐通道梯度量化，以抑制训练过程中量化噪声带来的负面影响。低于 INT8 表示精度的梯度量化仍基本未得到探索。FP 4 [20] 成功使用 4 位梯度训练现代卷积神经网络架构，且没有显著的精度损失，但其梯度以浮点数表示。

2.3. 超网络

另一项有趣的研究方向采用了超网络的概念，该概念最早由 [9], 提出，旨在将神经网络中的不可微部分转化为可微操作。MetaPruning [21]训练超网络，以自动生成用于卷积神经网络结构化剪枝的稀疏层权重。MetaQuant [22]在量化感知训练中用超网络替代 STE [15]，以减少 STE 的估计误差。

3. 预备知识

量化是一种将全精度（FP32）数值映射到一组离散值的过程。在常规梯度量化方案中， [1, 7, 8], 反向传播期间实时计算的相对于激活值的误差信号 $\nabla_{\mathbf{x}}L$ 以及权重的梯度 $\nabla_{\mathbf{W}}L$ 均采用低精度，这与前向传播的量化方案类似。为简洁起见，后文省略训练损失 L ，并将梯度表示为 $\nabla_{(\cdot)}$。本文讨论整数型量化。对于典型的梯度量化（例如 [7]），采用对称均匀量化。给定梯度

∇L 以及裁剪值 $c \in (0,\ \max(|\nabla L|)]$, 位宽为 B 的对称均匀量化器可表示为:

$$(\nabla L)_q = Q(\nabla L) = \text{round}\left(\text{clip}(\nabla L, c) \cdot \frac{2^{B-1}-1}{c}\right), \tag{1}$$

其中 $\text{clip}(\nabla L,\ c) = \min(\max(\nabla L,\ -c)\ c)$。量化后的梯度反量化为 $\widetilde{\nabla L} = (\nabla L)_q \cdot \frac{c}{2^{B-1}-1}$。

4. 方法

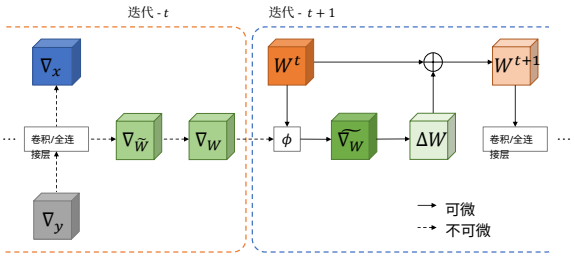


图1。梯度元量化延迟更新策略示意图。

4.1. 梯度元量化器

现在，我们详细介绍如何使用超网络设计梯度元量化器。我们旨在解决的主要挑战是，如何将反向传播中的梯度量化操作纳入前向计算图，使梯度量化器具备损失感知能力。为此，我们首先定义一个轻量、可微的神经网络 f_ϕ ，该网络在主模型 f 的所有可量化层之间共享。 f_ϕ 由 ϕ 参数化，其输入为传播得到的全精度梯度，或梯度与其他变量（例如权重 $\mathbf{W}$）的组合，并输出相应的量化梯度。超网络 f_ϕ 的详细结构将在后续章节中讨论。受 MetaQuant [22], 启发，我们对层权重采用延迟更新策略（见图1）。具体而言，首先，在训练的第 t 次迭代中，当反向传播经过一个卷积层/全连接层时，根据链式法则得到 $\nabla_{\mathbf{W}}$ ，

$$\nabla_{\mathbf{W}} = \left[\frac{\partial \mathbf{y}}{\partial \tilde{\mathbf{W}}}\right]^\top \left[\frac{\partial \tilde{\mathbf{W}}}{\partial \mathbf{W}}\right]^\top \nabla_{\mathbf{y}}, \tag{2}$$

其中，$\mathbf{W}$ 是前向传播中的量化权重。接下来，我们使用元量化器对 $\nabla_{\mathbf{W}}$ 进行量化，然后更新$\mathbf{W}$。具体做法是将全精度 $\nabla_{\mathbf{W}}$ 作为超网络 f_ϕ的输入，由超网络输出低精度梯度$\widetilde{\nabla_{\mathbf{W}}}$，表示为:

$$\widetilde{\nabla_{\mathbf{W}}} = f_\phi(\nabla_{\mathbf{W}}, \mathbf{W}). \tag{3}$$

为了获得迭代 $t+1$的权重更新量 $\Delta\mathbf{W}$ ，随后使用优化策略 π 进一步细化量化梯度，例如带动量的随机梯度下降或 Adam 方法，并按学习率 μ进行缩放，表示为:

$$\Delta\mathbf{W} = -\mu\pi\left(\widetilde{\nabla_{\mathbf{W}}}\right), \tag{4}$$

其中， π 的实现会根据不同的优化选项而变化；对于随机梯度下降和 Adam 等常见优化器，通常是可微的。具体而言，对于

随机梯度下降，$\pi\left(\widetilde{\nabla_{\mathbf{W}}}\right)=\widetilde{\nabla_{\mathbf{W}}}$。从图1可以看出，梯度 $\nabla_{\mathbf{W}}$ 与权重更新$\mathbf{W}^{t+1}$ 之间的所有计算都是可微的，因此我们成功地将元量化器纳入主网络的计算图中。因此，前向传播期间完整的权重更新过程可以写为：

$$\mathbf{W}^{t+1}=\mathbf{W}^{t}-\mu\pi\left(f_{\phi}(\nabla_{\mathbf{W}},\mathbf{W})\right). \tag{5}$$

4.2. 超网络的设计选择

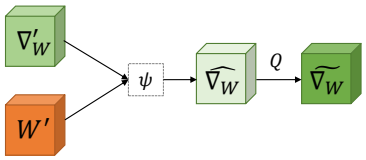


图2。所提出的用于元梯度量化的超网络总体架构。

如第4.1节所述，为了将梯度纳入计算图，超网络 f_{ϕ} 的总体设计应是一个满足以下条件的可微网络

$$f_{\phi}:\mathbb{R}^{k\times 1}\times\mathbb{R}^{k\times 1}\rightarrow\mathbf{S}^{k\times 1}, \tag{6}$$

其中$\mathbf{S}$ 是满足以下条件的有限集

$$\mathbf{S}=\left\{\frac{ck}{2^{B-1}-1}\mid -2^{B-1}\leq k\leq 2^{B-1},k\in\mathbb{Z}\right\}. \tag{7}$$

这可以通过令 $f_{\phi}=f_{\psi}\circ Q$ 来实现，其中 f_{ψ}：$\mathbb{R}^{N\times 1}\times\mathbb{R}^{N\times 1}\rightarrow\mathbb{R}^{N\times 1}$，且 $Q:\mathbb{R}^{N\times 1}\rightarrow\mathbf{S}^{N\times 1}$ 是公式1中定义的量化器。使用 STE 为 Q赋予梯度后，f_{ϕ} 便变得可微。

这样，根据用于全连接层的 [22],，我们可以先将梯度 $\nabla_{\mathbf{W}}\in\mathbb{R}^{n\times c}$ 和权重 $\mathbf{W}$ $\mathbb{R}^{(n)\times c}$ $\mathbb{R}^{(nc)\times 1}$ $\vec{\nabla_{\mathbf{W}}}$ $\vec{}$ $\in$ 分别转置为张量和 $\mathbf{W}$，然后将它们输入 f_{ϕ}；对于卷积层卷积核，则可以将 $\mathbb{R}^{n\times c\times k\times k}$ 转置为 $\mathbb{R}^{(nck^2)\times 1}$。至于 f_{ψ}的设计，我们可以直接采用 [22] 中的两种设计：

$$\texttt{MultiFC}:\quad f_{\psi}=\nabla'_{\mathbf{W}}\cdot FCs(\mathbf{W}'), \tag{8}$$

$$\texttt{LSTMFC}:\quad f_{\psi}=\nabla'_{\mathbf{W}}\cdot FCs(LSTM(\mathbf{W}')). \tag{9}$$

然而，我们发现，这种逐坐标神经网络设计 [23]无法充分利用权重和梯度的层内关系。因此，我们设计了另一种f_{ψ}：

$$\texttt{DualLSTMFC}:\quad f_{\psi}=FCs(LSTM(\mathcal{J})), \tag{10}$$

其中 $\mathcal{J}=[\mathbf{W}';\ \nabla'_{\mathbf{W}}]\in\mathbb{R}^{N\times 2}$ 表示拼接。该设计能够更好地交换权重与梯度张量之间的信息，并有望对梯度进行量化，从而有利于权重更新。在我们的实证评估中，我们还观察到 LSTMFC设计带来了持续的改进。

4.3. 元量化器的训练

根据公式5，通过链式法则可以得到超网络在第 $t+1$ 次迭代时接收到的梯度f_{ϕ}：

$$\frac{\partial L}{\partial\phi^{t+1}}=\frac{\partial L}{\partial\mathbf{W}^{t+1}}\frac{\partial\mathbf{W}^{t+1}}{\partial\phi^{t+1}}=-\mu\pi'\cdot\frac{\partial L}{\partial\mathbf{W}^{t+1}}\frac{\partial f_{\phi}(\nabla_{\mathbf{W}},\mathbf{W})^{\top}}{\partial\phi^{t+1}}. \tag{11}$$

我们还可以根据 $ff\ \circ Q$ 求得关于 ψ 的梯度：

$$\begin{aligned}\frac{\partial L}{\partial\psi^{t+1}}&=\frac{\partial L}{\partial\phi^{t+1}}\frac{\partial\phi^{t+1}}{\partial\psi^{t+1}}\\&\overset{\text{STE}}{\approx}=\frac{\partial L}{\partial\phi^{t+1}}=-\mu\pi'\frac{\partial L}{\partial\mathbf{W}^{t+1}}\frac{\partial f_{\phi}(\nabla_{\mathbf{W}},\mathbf{W})^{\top}}{\partial\phi^{t+1}},\end{aligned} \tag{12}$$

因此（省略上标 $t+1$）：

$$\nabla_{\psi}=\left[\frac{\partial L}{\partial\psi}\right]^{\top}\approx-\mu\pi'\frac{\partial f_{\phi}(\nabla_{\mathbf{W}},\mathbf{W})}{\partial\phi^{t+1}}\nabla_{\mathbf{W}}. \tag{13}$$

由于元梯度量化器的梯度已定义，我们可以以端到端的方式联合训练超网络和主模型。

5. 实验

5.1. 实现细节

我们首先在 CIFAR-10数据集上，对 ResNet-20 和 VGG-16 两种架构进行 4位量化，测试了不同的超网络选择，即 Mul-tiFC、LSTMFC 和 DualLSTMFC。我们在 CIFAR-10数据集上，对多种卷积神经网络架构开展了 INT8 和 INT4 量化感知训练实验。我们将方法与 UINT8 [7]和 DAINT8 [8]进行了比较。为确保公平比较，我们使用 DoReFa [1] 作为量化方法，并选择 DualLSTMFC 作为超网络。所有实验均选择带动量的随机梯度下降作为优化器。对于 INT8 实验，ResNet-20 的初始学习率为 0.01，ResNet-32、MobileNet-V2 和 VGG-16 的初始学习率为 0.005；ResNet-20 的学习率权重衰减为 0.001，ResNet-32、MobileNet-V2 和 VGG-16 的学习率权重衰减为 0.0001。对于 INT4，所有模型的学习率权重衰减均为 0.0001；对于 VGG-16 和 MobileNet-V2，ResNet-20 的初始学习率为 0.01，ResNet-32 为 0.005，ResNet-56 为 0.001，MobileNet-V2 和 VGG-16 为 0.0005，ResNet-110 为 0.0001。

5.2. 实验结果

INT4 结果。 表 1 展示了不同卷积神经网络在 CIFAR-10 INT4 量化下 Top-1 准确率的比较。对于 UINT8 [7], ，我们报告了使用相同的权重和激活值量化器方案 DoReFa 重新实现所得的结果。如表 1 所示，在 4 位量化下，对于所有模型，所提出的量化方案相比 UINT8 的准确率下降都明显更小。特别是对于 VGG-16，我们甚至观察到相较于 UINT8 提升了惊人的 69.38，我们推测这是由于缺少

模型	方法	FP32 /%	准确率 /%	Δ /%
ResNet-20	UINT8	91.36	73.45	-17.91
	Ours	91.36	**91.2**	**-0.16**
ResNet-32	UINT8	91.42	63.46	-27.96
	Ours	91.42	**90.56**	**-0.86**
ResNet-56	UINT8	91.95	65.27	-26.68
	Ours	91.95	**90.24**	**-1.71**
ResNet-110	UINT8	91.67	62.57	-29.10
	Ours	91.67	**89.16**	**-2.51**
MobileNet-V2	UINT8	93.57	88.12	-5.45
	Ours	93.57	**89.00**	**-4.57**
VGG-16	UINT8	92.25	22.23	-70.02
	Ours	92.25	**91.61**	**-0.64**

表 1。INT4 量化下与其他方法的比较。 Δ表示相较于全精度（FP32）结果的性能下降： $\Delta = \text{Acc} - \text{FP}$（越高越好）。

VGG-16 模型中的残差结构会加剧梯度量化误差带来的负面影响。我们的方法始终能够支持稳定的梯度量化训练，在所有模型上，相较于 FP32 的性能下降最多为 4.57。

INT8结果。 表 2 展示了不同卷积神经网络在 CIFAR-10 INT8量化下的 Top-1 准确率比较。我们观察到，在 8 位量化下，与最先进的梯度量化方法 [7, 8] 相比，我们在几乎所有测试模型上都取得了最佳表现，尤其是在 VGG-16 上：基线 UINT8 完全无法收敛，而我们的量化方案成功将相较于 FP 的性能下降控制在仅 -0.05。唯一的例外是 ResNet-20，我们的表现略逊于 DAINT8 [8], ，但这是预期之中的，因为 DAINT8 使用了更细粒度的逐通道量化，而我们和 UINT8 [7]使用的是事实上更简单的逐层量化。尽管如此，在 MobileNet-V2 上，我们的性能下降仍低于 DAINT8 相较于 FP32 的性能下降，这表明我们的方法潜力巨大：仅通过将梯度纳入目标函数便能提升性能。

5.3. 消融研究与其他讨论

我们在 CIFAR-10 上使用两种模型 ResNet-20 和 VGG-16，在 INT4量化条件下，对超网络设计的选择进行了消融研究。表 3 显示，两种模型均呈现出 DualLSTMFC $>$ LSTMFC $>$ MultiFC 的明显趋势，这表明我们新提出的 DualLSTMFC设计能够更有效地生成有利于权重更新的量化梯度，从而使准确率下降幅度最低。然而，它们之间的整体性能差异

模型	方法	FP32 /%	准确率 /%	Δ /%
ResNet-20	UINT8	91.36	91.10	-0.26
	DAINT8	92.35	**92.76**	**0.41**
	Ours	91.36	91.30	-0.06
ResNet-32	UINT8	91.42	90.70	-0.72
	Ours	91.42	**91.40**	**-0.02**
MobileNet-V2	UINT8	93.57	93.38	-0.19
	DAINT8	94.73	**94.37**	-0.36
	Ours	93.57	93.51	**-0.06**
VGG-16	UINT8	92.25	31.32	-60.93
	Ours	92.25	**92.20**	**-0.05**

表 2。与其他 INT8 量化方法的比较。

模型	超网络	FP32	准确率 /%	Δ /%
ResNet-20	MultiFC	91.36	91.06	-0.30
	LSTMFC	91.36	91.15	-0.21
	DualLSTMFC	91.36	**91.20**	**-0.16**
VGG-16	MultiFC	92.25	91.50	-0.75
	LSTMFC	92.25	91.54	-0.71
	DualLSTMFC	92.25	**91.61**	**-0.64**

表3。INT4量化下超网络设计的消融研究。

并不大，这表明超网络的选择对所提方法的有效性影响甚微。

6. 结论

在本研究中，为了提高量化感知训练（QAT）期间梯度量化的稳定性，我们提出了一种新颖的梯度量化元量化器，它能够在整个训练过程中自动优化量化器，因此对反向传播中产生的量化噪声具有更强的鲁棒性。在 CIFAR-10数据集上使用各种卷积神经网络进行的大量实验表明，元量化器有助于量化感知训练在较低梯度比特率下保持稳定，其表现远超基线方法，尤其是在位宽极低的情况下。

致谢

本研究由新加坡科技研究局（A*STAR）通过其基金资助（项目编号 A1892b0026 和 C211118009）。本文所表达的任何观点、研究结果、结论或建议均为作者所有，不代表 A*STAR 的观点。

7. 参考文献

[1] Shuchang Zhou、Yuxin Wu、Zekun Ni、Xinyu Zhou、He Wen 和 Yuheng Zou，《DoReFa-Net：使用低位宽梯度训练低位宽卷积神经网络》，*arXiv* 预印本 *arXiv:1606.06160*，2016年。

[2] Qing Jin、Linjie Yang 和 ZhenyuLiao，“Adabits：采用自适应位宽的神经网络量化”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2020年6月。

[3] Wang Zhe、Jie Lin、Vijay Chandrasekhar 和 Bernd Girod，“优化深度神经网络权重和激活值压缩的比特分配”，载于*2019*年*IEEE*图像处理国际会议（*ICIP*）。IEEE，2019年，第3826–3830页。[4] Kaixin Xu、Zhe Wang、Xue Geng、Min Wu、Xiaoli Li 和 Weisi Lin，“深度神经网络中层自适应权重剪枝的高效联合优化”，载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2023年10月，第17447–17457页。[5] Jcjohnson，“Cnn-benchmarks：热门卷积神经网络模型的基准测试”，2016年。[6]Ron Banner、Itay Hubara、Elad Hoffer 和 Daniel Soudry，“神经网络8位训练的可扩展方法”，神经信息处理系统进展，第31卷，2018年。[7] Feng Zhu、Ruihao Gong、Fengwei Yu、Xianglong Liu、Yanfei Wang、Zhelong Li、Xiuqi Yang 和 Junjie Yan，“面向卷积神经网络的统一INT8训练”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第1969–1979页。[8] Kang Zhao、Sida Huang、Pan Pan、Yinghan Li、Yingya Zhang、Zhenyu Gu 和 Yinghui Xu，“用于卷积神经网络训练的分布自适应INT8量化”，载于人工智能促进协会会议论文集，2021年，第35卷，第3483–3491页。[9] David Ha、Andrew M. Dai 和 Quoc V. Le，“超网络”，载于国际学习表征会议，2017年。

[10] Dongqing Zhang、Jiaolong Yang、Dongqiangzi Ye 和 Gang Hua，“Lq-nets：面向高精度、紧凑型深度神经网络的学习式量化”，载于欧洲计算机视觉会议论文集（*ECCV*），2018年，第365–382页。

[11] Steven K Esser、Jeffrey L McKinstry、Deepika Bablani、Rathinakumar Appuswamy 和 Dharmendra S Modha，“学习步长量化”，载于*ICLR*，2020年。[12]Yash Bhalgat、Jinwon Lee、Markus Nagel、Tijmen Blankevoort 和 Nojun Kwak，“Lsq+：通过可学习偏移量和更优初始化改进低比特量化”，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，2020年，第696–697页。[13] Qian Lou、Feng Guo、Minje Kim、Lantao Liu 和 Lei Jiang，“Autoq：自动逐核神经网络量化”，载于*ICLR*，2020年。[14] Daisuke Miyashita、Edward H Lee 和 Boris Murmann，“采用对数数据表示的卷积神经网络”，*arXiv*预印本 *arXiv:1603.01025*，2016年。[15] Yoshua Bengio、Nicholas L´eonard 和 Aaron Courville，“通过随机神经元估计或传播梯度以进行条件计算”，*arXiv:1308.3432*，2013年。[16] Ruihao Gong、Xianglong Liu、Shenghu Jiang、Tianxiang Li、Peng Hu、Jiazhen Lin、Fengwei Yu 和 Jun- jie Yan，“可微软量化：连接全精度与低比特神经网络”，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第4852–4861页。[17] Yukuan Yang、Lei Deng、Shuang Wu、Tianyi Yan、Yuan Xie 和 Guoqi Li，“使用全8位整数训练高性能、大规模深度神经网络”，《神经网络》，第125卷，第70–82页，2020年。[18]Xishan Zhang、Shaoli Liu、Rui Zhang、Chang Liu、Di Huang、Shiyi Zhou、Jiaming Guo、Qi Guo、Zidong Du、Tian Zhi 等，“定点反向传播训练”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第2330–2338页。[19] Yonggan Fu、Han Guo、Meng Li、Xin Yang、Yining Ding、Vikas Chandra 和 Yingyan Lin，“Cpt：通过循环精度实现高效深度神经网络训练”，*arXiv*预印本 *arXiv:2101.09868*，2021年。[20] X Sun、N Wang、C Chen、J Ni、A Agrawal、X Cui、S Venka、K El、V Srini 和 K Gopa，“深度神经网络的超低精度4位训练”，载于*NeurIPS*，2020年。[21] Zechun Liu、Haoyuan Mu、Xiangyu Zhang、Zichao Guo、Xin Yang、Kwang-Ting Cheng 和 Ji an Sun，“Metapruning：用于自动神经网络通道的元学习

p行内g,” 在会议论文集 *gs of the*

*IEEE/CVF*国际计算机视觉会议，2019年，第3296–3305页。

[22] Shangyu Chen、Wenya Wang 和 Sinno Jialin Pan，“Metaquant：通过学习穿透不可微量化来学习量化”，神经信息处理系统进展，第 32 卷，2019年。

[23] Marcin Andrychowicz、Misha Denil、Sergio Gomez、Matthew W Hoffman、David Pfau、Tom Schaul、Brendan Shillingford 和 Nando De Freitas，“通过梯度下降学习梯度下降”，神经信息处理系统进展，第 29 卷，2016年。