

深度整数训练和推理 神经网络

吴爽¹, 李国琪¹, 陈峰², 石路平¹ 1精密仪器系2自动化系 类脑计算
算研究北京未来芯片创新中心清华大学{lpshi,chenfeng}
@mail.tsinghua.edu.cn

抽象的

对具有离散参数的深度神经网络及其在嵌入式系统中的部署的研究一直是活跃且有前途的课题。尽管之前的工作已成功降低推理精度,但尚未同时证明将训练和推理过程转移到低位宽整数。在这项工作中,我们开发了一种称为“WAGE”的新方法来离散化训练和推理,其中层间的权重(W)、激活(A)、梯度(G)和误差(E)被移动并线性约束为低位宽整数。为了对定点设备执行纯离散数据流,我们进一步用一个常量缩放层代替批量归一化,并简化其他难以实现整数的组件。可以在多个数据集上获得更高的准确性,这表明 WAGE 在某种程度上充当了一种正则化。根据经验,我们展示了在硬件系统(例如基于整数的深度学习加速器和神经形态芯片)中部署训练的潜力,具有可比的精度和更高的能效,这对于未来具有迁移和持续学习需求的可变场景中的 AI 应用至关重要。

1简介

最近,深度神经网络(DNN)被广泛用于众多人工智能应用(Krizhevsky 等人,2012年;Hinton 等人,2012年;Silver 等人,2016年)。取决于大量可调参数,DNN 被认为具有强大的多级特征提取和表示能力。然而,训练 DNN 需要具有高精度(float32)处理单元和丰富内存的 GPU 和 CPU 等能源密集型设备,这极大地挑战了它们在便携式设备上的广泛应用。此外,最先进的网络通常具有更大的权重和有效打碎所有训练样本的能力(Zhang et al., 2016),容易导致过度拟合。

因此,人们对在推理过程中减小网络规模非常感兴趣(Hubara 等人,2016年;Rastegari 等人,2016年;Li 等人,2016年),以及用于商业解决方案的专用硬件(Jouppi 等人)等人,2017年;陈等人,2017年;施等人,2015年)。由于随机梯度下降(SGD)优化的积累,训练的精度要求通常高于推理(Hubara et al., 2016; Li et al., 2017)。因此,大多数现有技术只专注于部署训练有素的压缩网络,同时在训练过程中仍然保持高精度和计算复杂度。在这项工作中,我们将这个问题解决为如何使用低位宽整数处理训练和推理,这对于在专用硬件中实现 DNN 至关重要。为此,解决了离散训练 DNN 的两个基本问题:i) 如何量化所有操作数和操作,以及 ii) SGD 计算和累加需要多少位或状态。

关于这些问题,我们提出了一个称为“WAGE”的框架,它在训练和推理中将所有层中的权重(W)、激活(A)、梯度(G)和错误(E)限制为低位宽整数。首先,对于操作数,应用线性映射和方向保持移位

实现三元权重,8 位整数用于激活和梯度累积。其次,对于操作,批量归一化 (Ioffe & Szegedy, 2015) 被常数比例因子取代。

其他用于微调的技术,例如带有动量的 SGD 优化器和 L2 正则化,在性能下降很小的情况下被简化或放弃。考虑到整体双向传播,我们将推理完全简化为累加比较循环,并分别将训练简化为具有对齐操作的低位宽乘法累加 (MAC) 循环。

我们启发式地探索整数对误差计算和梯度累积的位宽要求,这在以前的工作中很少讨论。实验表明,引导前几层收敛的是误差的相对值 (方向) 而不是绝对值 (数量级)。此外,小值虽然逐层传播,但对先前方向的影响可以忽略不计,可以在量化中部分丢弃。我们利用这些现象并使用保持方向的移位操作来限制错误。

至于梯度累积,虽然在推理中权重被量化为三元值,但相对较高的位宽对于存储和累积梯度更新是必不可少的。

所提出的框架在 MNIST、CIFAR10、SVHN、ImageNet 数据集上进行了评估。与那些只在推理时离散化权重和激活的人相比,它具有相当的准确性,并且可以进一步缓解过度拟合,表明某种类型的正则化。WAGE 为 DNN 生成纯双向低精度整数数据流,可以灵活地应用于专用硬件中的训练和推理。我们在 GitHub¹ 上发布代码。

2 相关工作

我们主要关注降低训练和推理中操作数和操作的精度。用于降低网络压缩、修剪 (Han 等人, 2015 年; Zhou 等人, 2017 年) 和紧凑型架构 (Howard 等人, 2017 年) 等复杂性的正交和互补技术非常高效,但超出了本文的范围。

重量和激活 Courbariaux 等人。 (2015); 胡巴拉等人。 (2016) 提出了连续使用二进制权重 (BC) 和激活 (BNN) 训练 DNN 的方法。他们将噪声作为一种正则化形式添加到权重和激活中,但实值梯度在实值变量中累积,这表明 SGD 优化可能需要高精度累积。

XNOR-Net (Rastegari et al., 2016) 有一个过滤器式的权重比例因子来提高性能。XNOR-Net 中的卷积可以使用 XNOR 逻辑单元和位计数操作有效地实现。然而,这些浮点因子是在训练过程中同时计算的,这通常会加重训练工作量。在 TWN (Li et al., 2016) 和 TTQ (Zhu et al., 2016) 中,引入了两个对称阈值来约束权重为三元值 $\{-1, 0, +1\}$ 。他们声称在模型复杂性和表达能力之间进行权衡。

梯度计算和累积 DoReFa-Net (Zhou et al., 2016) 在反向传递中将梯度量化为具有离散状态的低位宽浮点数。TernGrad (Wen et al., 2017) 将梯度更新量化为三元值,以减少分布式训练中梯度同步的开销。尽管如此,DoReFa-Net 和 TernGrad 中的权重在训练期间像以前的作品一样存储和更新为 float32。此外,批量归一化及其导数的量化被忽略。因此,训练过程的整体计算图仍然以 float32 呈现,并且通过外部量化变得更加复杂。通常,很难直接在基于整数的硬件中应用 DoReFa-Net 训练,但它显示出探索具有离散梯度下降方向的高维离散空间的潜力。

3 工资量化

WAGE 量化的主要思想是将四个操作数限制为低位宽整数:推理中的权重 W 和激活 a ,反向传播训练中的误差 e 和梯度 g ,见图 1。

我们将原来误差的定义扩展到多层:误差 e 是激活 a 对于每个卷积层或全连接层的梯度,而梯度 g 特指权重 W 的梯度累加。考虑到第 i 个前馈网络的层,我们

¹ <https://github.com/boluoweifenda/WAGE>

有：

$$\frac{\partial L_i}{\partial a_i} g_{ai} = \frac{\partial L}{\partial W_i} \tag{1}$$

其中 L 是损失函数。我们将这两个在大多数现有方案中混淆的术语分开。权重 g 的梯度和激活 e 的梯度在每一层中流向不同的路径，在推理和反向训练中都是一个分支，通常作为 MAC 操作的节点。

对于第 i 层的前向传播，假设权重是用 kG 位整数存储和累加的，那么许多工作都在争取更好的量化函数 $QW(\cdot)$ 将更高精度的权重映射到它们的 kW 位反射，对于例如 $[-0.9, 0.1, 0.7]$ 到 $[-1, 0, 1]$ 。尽管权重像 float32 一样以高精度累积，但在训练后在专用硬件中部署反射的内存效率要高得多。使用函数 $QA(\cdot)$ 将激活量化为 kA 位，以对齐由 MAC 引起的增加的位宽。在之前的工作中，权重和激活被离散化为二进制值，然后 MAC 退化为非常高效的逻辑和位计数操作（Rastegari 等人，2016 年）。

对于第 i 层的反向传播，激活和权重的梯度是通过 MAC 的导数计算的，通常被认为至少是 16 位浮点精度。如图 1 所示， kA 位输入和 kW 位权重之间的 MAC 会将输出的位宽增加到带符号整数表示形式的 $[kA + kW - 1]$ ，并且错误 e 也会发生类似的扩展。考虑到仅使用低位宽整数进行训练，我们提出了附加函数 $QE(\cdot)$ 和 $QG(\cdot)$ 来分别将 e 和 g 的位宽限制为 kE 位和 kG 位。一般来说，有 MAC 运算的地方，在推理和反向传播中就有 $QW(\cdot)$ 、 $QA(\cdot)$ 、 $QG(\cdot)$ 、 $QE(\cdot)$ 等量化算子。

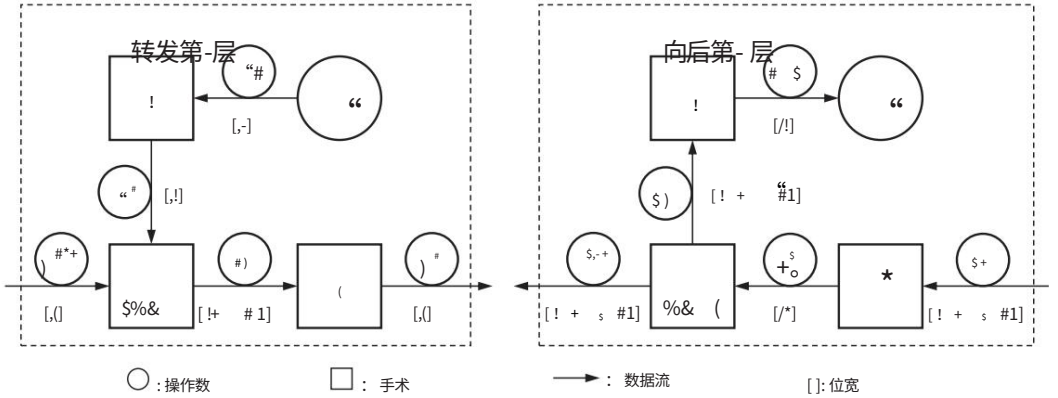


图 1: 在 WAGE 计算数据流中添加了四个运算符 $QW(\cdot)$ 、 $QA(\cdot)$ 、 $QG(\cdot)$ 、 $QE(\cdot)$ 以降低精度，有符号整数的位宽在箭头下方或右侧，激活包含在 MAC 简洁。

3.1 基于移位的线性映射和随机舍入

在 WAGE 量化中，为简单起见，我们采用 k 位整数的线性映射，其中连续和无界值以均匀距离 σ 离散化：

$$\sigma(k) = 2^{1-k} \quad , \quad k \in \mathbb{N}^+ \tag{2}$$

那么将浮点数 x 转换为其 k 位宽带符号整数表示的基本量化函数可以表示为：

$$Q(x, k) = \text{Clip} \left(\frac{x}{s(k)} \right) \cdot \text{round} \tag{3}$$

其中 round 将连续值近似为最接近的离散状态。Clip 是将无界值裁剪为 $[-1 + \sigma, 1 - \sigma]$ 的饱和函数，其中负最大值 -1 为

移除以保持对称。例如, $Q(x, 2)$ 将 $\{-1, 0.2, 0.6\}$ 量化为 $\{-0.5, 0, 0.5\}$ 。

等式 3 仅用于 GPU 等浮点硬件的仿真,而在定点设备中,量化和饱和是自动满足的。

在某些操作数 (例如,错误)中应用线性映射之前,我们引入了一个额外的单片比例因子,用于将值分布移动到适当的数量级,否则值将全部饱和或由等式 3 清除。比例因子由下式计算移位 t 函数,然后在后面的步骤中划分:

$$\text{位移 } t(x) = 2 \text{round}(\log_2 x) \quad (4)$$

最后,我们提出随机舍入来代替小的和实值的更新来代替训练中的梯度累积。3.3.4 节将详细介绍运算符 $QG(\cdot)$ 的实现,其中高位宽梯度由 16 位随机数生成器随机限制为 kG 位整数。图 2 总结了 WAGE 中使用的量化方法。

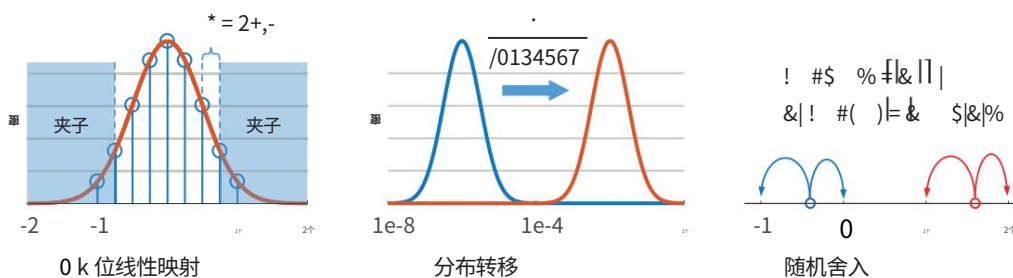


图 2: WAGE 中使用的量化方法。符号 P 、 x 、 \cdot 和 \cdot 分别表示概率、向量、floor 和 ceil。Shift(\cdot)指的是具有特定自变量的等式 4。

3.2 权重初始化

在以前的工作中,权重直接由 sgn 函数二值化或由训练期间计算的阈值参数三值化。然而, BNN 在没有批量归一化的情况下无法收敛,因为权重值 ± 1 对于典型的 DNN 来说相当大。批量归一化不仅有效地避免了梯度爆炸和消失的问题,而且减轻了对正确初始化的需求。然而,在没有浮点单元 (FPU) 的情况下,对每一层的输出进行归一化并计算它们的梯度是相当复杂的。此外,批量输出的移动平均占用外部存储器。BNN 显示了批量归一化的基于移位的变化,但很难将所有元素转换为定点表示。因此,在这项工作中应该谨慎地初始化权重,其中批量归一化被简化为一个常量缩放层。基于 MSRA (He et al., 2015) 的改进初始化方法可以表述为:

$$W \sim U(-L, +L), L = \max\{6/n_{in}, L_{min}\}, L_{min} = \beta\sigma \quad (5)$$

其中 n_{in} 为层扇入数, MSRA 中原始极限 $6/n_{in}$ 计算为理论上保持同一层的输入和输出之间的方差相同。附加限制 L_{min} 是均匀分布 U 应达到的最小值, β 是大于 1 的常数,以在最小步长 σ 和最大值 L 之间创建重叠。在 kW 位线性映射的情况下,如果权重 W 是直接采用原始限制量化,当位宽 kW 足够小 (例如 4) 或扇入 n_{in} 足够宽时,我们将得到全零张量,其中初始化权重可能永远不会达到定点整数表示的最小步长 σ 。所以 L_{min} 保证了权值在随机初始化的时候可以超越 σ 并在 $QW(\cdot)$ 之后量化为非零值。

3.3 量化细节

3.3.1 重量 $QW(\cdot)$

等式 5 中修改后的初始化将整体放大权重并保证其适当分布,然后直接使用等式 3 量化 W :

$$W_q = QW(W) = Q(W, kW) \quad (6)$$

应该注意的是,权重的方差与原始限制相比是按比例缩放的,这将导致网络输出的爆炸。为了减轻放大效应,XNOR-Net 提出了一种以全精度连续计算的过滤器缩放因子。考虑到整数实现,我们引入了一个基于分层移位的缩放因子 α 来减弱放大效果:

$$\alpha = \max\{\text{Shift } t(L_{\min}/L), 1\} \quad (7)$$

其中 α 是由网络结构确定的每一层的预定义常量。修改后的初始化和衰减因子 α 一起将浮点权重近似为其整数表示,除了 α 在激活后生效以保持由 kW 位整数表示的权重精度。

3.3.2 激活 QA(\cdot)

如上所述,操作数的位宽在 MAC 之后增加。然后,典型的 CNN 通常会进行池化、归一化和激活。避免平均池化,因为平均操作会增加精度需求。此外,我们假设每个隐藏层的批量输出近似为零均值,然后批量归一化退化为缩放层,其中可训练和批量计算的缩放参数被公式 7 中提到的 α 替换。

如果激活以 kA 位表示,则激活的整体量化可以表示为:

$$aq = QA(a) = Q(a/\alpha, kA) \quad (8)$$

3.3.3 错误 QE(\cdot)

在训练过程中使用链式法则逐层计算误差 e 。虽然反向传播的计算图类似于推理,但输入是 L 的梯度,与网络的实际输入相比相对较小。更重要的是,错误是无限的,并且可能比激活的范围大得多,例如 $[10^{-9}, 10^{-4}]$ 。DoReFa-Net 首先对 e 应用仿射变换将它们映射到 $[-1, 1]$,然后在量化后反转变换。因此,量化后的 e 仍然呈现为具有离散状态和大部分较小值的 float32 数字。

然而,实验发现,引导前面的层收敛的是方向而不是误差的数量级,那么 DoReFa-Net 中的量化后的逆变换就不再需要了。仅方向保存提示我们彻底传播整数误差,其中误差分布首先通过除以图 2 所示的移位因子按比例缩放为 $[-\sqrt{2}, +\sqrt{2}]$,然后由 $Q(e, kE)$ 量化:

$$eq = QE(e) = Q(e/\text{Shift } t(\max\{|e|\}), kE) \quad (9)$$

其中 $\max\{|e|\}$ 提取误差 e 中所有元素的层级最大绝对值,多通道用于卷积,多样本用于批量训练。误差的量化丢弃了大部分小于 σ 的值,我们将在后面讨论对精度的影响。

3.3.4 梯度 QG(\cdot)

由于我们仅在移动后保留误差的相对值,因此从后向误差 e 和前向激活 a 之间的 MAC 派生的梯度更新 g 会随之移动。我们首先用另一个缩放因子重新缩放梯度 g ,然后引入基于移位的学习率 η :

$$gs = \eta \cdot g / \text{Shift } t(\max\{|g|\}) \quad (10)$$

其中 η 是 2 的整数次幂。偏移梯度 gs 表示最小步数和更新权重的方向。如果权重以 kG 位数存储,则整数的最小修改步长为 ± 1 ,浮点值的最小修改步长为 $\pm \sigma(kG)$ 。这里学习率 η 的实现与基于 float32 的普通 DNN 中的实现有很大不同。在 WAGE 中,只剩下权重变化的方向,步长是最小步长 σ 的整数倍。如果 η 为 2 或更大以在开始时加速训练,则移动梯度 gs 可能会大于 1,或者在训练的后半段通常应用学习率衰减时小于 0.5。如图 2 所示,在后者中替代小梯度的累积

在这种情况下,我们将 g_s 分为整数部分和小数部分,然后使用 16 位随机数生成器将高位宽 g_s 随机约束为 kG 位整数:

$$\Delta W = QG(g) = \sigma(kG) \cdot \text{sgn}(g_s) \cdot (|g_s| + \text{伯努利}(|g_s| - |g_s|)) \quad (11)$$

其中 Bernoulli (Zhou et al., 2016) 将小数部分随机采样为 0 或 1。通过适当设置 kG ,梯度量化将限制最小步长,这可以避免局部最小值和过度拟合。此外,当 η 不大于 1 时,梯度将是三元值,这降低了分布式训练的通信成本 (Wen et al., 2017)。最后,权重 W 在用离散增量 ΔW 更新后可能会超过 kG 位整数呈现的范围 $[-1 + \sigma, 1 - \sigma]$ 。因此,Clip 函数对于饱和和确保只有 $2kG - 1 - 1$ 个状态用于权重累积是必不可少的。在第 t 次迭代的情况下,我们有:

$$W_{t+1} = \text{Clip}\{W_t - \Delta W_t, -1 + \sigma(kG), 1 - \sigma(kG)\} \quad (12)$$

3.4 杂项

从上面,我们已经说明了我们对于权重、激活、梯度和误差的量化方法。详细计算图见算法 1。在仅使用整数的整体训练过程中仍然存在一些问题需要说明。

Momentum、RMSProp 和 Adam 等梯度下降优化器至少包含一个梯度更新 ΔW 或其移动平均值的副本,训练期间权重的内存消耗加倍,这部分相当于使用更大的 kG 。由于权重更新 ΔW 被量化为 σ 的整数倍并按 η 缩放,我们采用没有任何形式的动量或自适应学习率的纯小批量 SGD 来显示减少存储需求的潜力。

尽管 L2 正则化对于许多经常发生过拟合的大规模 DNN 效果很好,但 WAGE 去除了等式 3 中的小值并在等式 11 中引入了随机性,作为某些类型的正则化,并且可以在以后的实验中获得相当的精度。因此,我们仍然将 L2 权重衰减和丢失作为补充正则化方法。

Softmax 层和交叉熵准则在分类任务中被广泛采用,但 e 的计算很难应用于低位宽线性映射场合。对于类别数量较少的任务,我们避免使用 Softmax 层并应用均方误差准则,但省略均值运算以形成误差平方和 (SSE) 准则,因为移位误差将在等式 9 中获得相同的值。

4 个实验

在本节中,我们将 WAGE 位设置为 2-8-8-8 作为 CNN 或 MLP 中所有层的默认值。三元权重的位宽 kW 为 2,这意味着在推理过程中没有乘法运算。常量参数 β 为 1.5,使随机初始化时三元权重的概率相等。激活和错误应该具有相同的位宽,因为反向传播的计算图类似于推理,并且可能应用于硬件或忆阻器阵列的相同分区 (Sheridan 等人, 2017)。尽管 XNOR-Net 实现了 1 位激活,但将错误减少到 4 位或更少会显著降低我们测试中的准确性,因此位宽 kA 和 kE 同时增加到 8。权重在训练期间以 8 位整数存储,并在推理期间由两个恒定对称阈值进行三元化。我们首先为普通网络构建计算图,然后在前向传播中插入量化节点,并在 Tensorflow 上的每一层的反向传播中覆盖梯度 (Abadi 等人, 2016 年)。我们的方法在 MNIST、SVHN、CIFAR10 和 ILSVRC12 (Russakovsky 等人, 2015 年) 上进行了评估,表 1 显示了比较结果。

4.1 实施细节

MNIST: 采用了具有 32C5-MP2-64C5-MP2-512FC-10SSE 的 LeNet-5 (LeCun et al., 1998) 的变体。输入的灰度图像被视为激活并通过等式 8 量化,其中 α 等于 1。WAGE 中的学习率 η 在整个 100 个 epoch 中保持为 1。我们报告了在测试集上运行 10 次的平均准确率。

SVHN & CIFAR10: 我们使用类似 VGG 的网络 (Simonyan & Zisserman, 2014) 和 $2 \times (128C3) - MP2 - 2 \times (256C3) - MP2 - 2 \times (512C3) - MP2 - 1024FC - 10SSE$ 。对于 CIFAR10 数据集, 我们遵循 Lee 等人的数据扩充。(2015) 用于训练: 每边填充 4 个像素, 并从填充图像或其水平翻转中随机裁剪出一个 32×32 的补丁。对于测试, 仅评估原始 32×32 图像的单个视图。该模型以 128 的 mini-batch 大小和总共 300 个 epoch 进行训练。学习率 η 设置为 8, 并在第 200 和第 250 轮除以 8。原始图像被缩放并偏置到 $[-1, +1]$ 的范围内, 用于 8 位整数激活表示。至于 SVHN 数据集, 我们省略了随机翻转增强并将训练时期减少到 40, 因为它是一个相当大的数据集。错误率的评估方式与 MNIST 相同。

ImageNet: WAGE 框架使用 AlexNet (Krizhevsky 等人, 2012 年) 模型在 ILSVRC12 数据集上进行评估, 但删除了 dropout 和局部响应归一化层。图像首先调整为 256×256 , 然后随机裁剪为 224×224 并水平翻转, 然后像 CIFAR10 一样进行偏差减法。为了进行测试, 评估了验证集中的单中心作物。由于 ImageNet 任务比 CIFAR10 困难得多并且有 1000 个类别, 在 WAGE 中应用 SSE 或较链损失准则时很难收敛, 因此我们在最后一层添加 Softmax 并删除量化, 以免精度严重下降 (Tang et al., 2017)。该模型以 256 的 mini-batch 大小和总共 70 个 epoch 进行训练。学习率 η 设置为 4, 并在第 60 和第 65 轮时除以 8。

表 1: 先前作品中的测试或验证错误率 (%) 和多个数据集上的 WAGE。Opt 表示梯度下降优化器, withM 表示带动量的 SGD, BN 表示 batch normalization, 32 位表示 float32, ImageNet top-k 格式: top1/top5。

| 方法Opt BN MINIST SVHN CIFAR10 ImageNet | | | | | | | | |
|---------------------------------------|-----------------|----|----|-------|------|------|-------|-----------|
| 公元前 | 10 ⁸ | 32 | 32 | 32 亚当 | 1.29 | 2.30 | 9.90 | - |
| BNN 1 | 10 ⁸ | 32 | 32 | 32 亚当 | 0.96 | 2.53 | 10.15 | - |
| BWN1 1 | | 32 | 32 | 32与M | - | - | - | 43.2/20.6 |
| 异或非 1 | 10 ⁸ | 32 | 32 | 32 亚当 | - | - | - | 55.8/30.8 |
| 台湾 2 | | 32 | 32 | 32与M | 0.65 | - | 7.44 | 34.7/13.8 |
| TTQ 2 | | 32 | 32 | 32 亚当 | - | - | 6.44 | 42.5/20.3 |
| DoReFa2 8 | 8个 | 32 | 8 | 8 亚当 | - | 2.30 | - | 47.0/- |
| TernGrad3 32 32 | | | 2个 | 32 亚当 | - | - | 14.36 | 42.4/19.5 |
| 工资 2 | 8个 | 8个 | 8 | 8新元 | 0.40 | 1.92 | 6.78 | 51.6/27.8 |

4.2 训练曲线和正则化

我们进一步比较了 CIFAR10 上的 WAGE 变体和普通 CNN。vanilla CNN 具有与上述相同的类似 VGG 的架构, 只是没有应用任何操作数或操作的量化。我们在每一层中添加批量归一化, 在最后一层添加 Softmax, 将 SSE 替换为交叉熵准则, 然后使用 $1e-4$ 的 L2 权重衰减和 0.9 的动量进行训练。学习率设置为 0.1, 并在第 200 和第 250 个时期除以 10。对于 WAGE 的变体, 模式 28ff 在反向传播中没有量化节点。尽管 28ff 模式具有与 vanilla 模式相同的优化器和学习率退火方法, 但我们发现等式 7 中的重新缩放因子 α 减少了权重更新。因此, 放大和调整了 28ff 的学习率, 从而减少了错误税率 3%。图 3 显示了三个对应的训练曲线。可以看出, 2888 模式具有与普通 CNN 相当的收敛速度, 比那些在推理时间内只离散化权重和激活的模式具有更好的准确性, 尽管波动性稍大。反向传播的离散化在某种程度上充当了另一种类型的正则化, 并且在降低学习率 η 时具有显着的错误率下降。

1BWN 是 XNOR 的对应物, 只量化权重
2他们使用浮点表示错误
3仅用于分布式训练中 worker-to-server 通信, 权重用 float32 累加

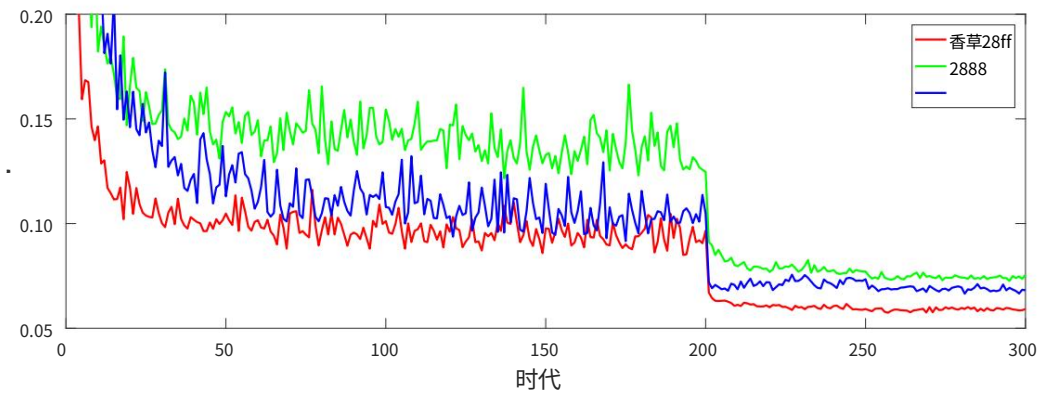


图 3:WAGE 变化的训练曲线和 CIFAR10 上的普通 CNN。

4.3错误位宽

在之前的实验中,位宽 k_E 默认设置为 8。为了进一步探索适当的位宽及其截断边界,我们首先在 100 个训练时期后从 CIFAR10 的普通 CNN 导出错误。128个mini-batch数据中最后一个卷积层的误差直方图如图4所示。很明显,误差大致服从对数正态分布,其中值相对较小,范围明显较大。当用 k_E 位整数量化时,应该选择合适的窗函数来截断分布,同时保留反向传播的近似方向。有关所有 W、A、G、E 操作数的分层直方图的更多详细信息,请参见图 5。

首先,上 (右)边界固定为所有错误元素中的最大绝对值,如等式 9 所述。然后左边界将基于位宽 k_E 。

我们针对 4 到 15 的 k_E 进行了一系列实验。图 4 中的箱线图表明,用整数表示的 4-8 位错误足以完成 CIFAR10 分类任务。默认选择位宽 8 以匹配 8 位图像颜色级别和微控制单元 (MCU) 中的大多数操作数。WAGE-2888同层误差直方图显示,经过逐层移位和量化后,误差分布重塑,大部分聚集成截断窗口。因此,保留了大部分方向信息。此外,较小的误差值对先前方向的影响可以忽略不计,虽然逐层累积,但在量化中被部分丢弃。

由于窗口的宽度已经过优化,我们将窗口左移 γ 以探索其水平位置。右边界可以表示为 $\max\{|e|\}/\gamma$ 。表 2 显示了偏移误差的影响:虽然大值占少数,但它们对反向传播训练起着关键作用,而大多数小值实际上充当噪声。

表 2:当左移上边界与因子 γ 时,CIFAR10 上的测试错误率 (%)。

| C | 1x | 2x | 4x | 8x |
|----|------|------|------|-------|
| 错误 | 6.78 | 7.31 | 8.08 | 16.92 |

4.4梯度位宽

在之前的实验中,位宽 k_G 默认设置为 8。尽管权重在推理中以三元值传播,并实现了比 float32 权重高 16 倍的压缩率,但它们以相对较高的位宽 (8 位)保存和累积,用于反向传播训练。因此,整体压缩率仅为4x。权重更新 k_G 之间的不一致位宽及其对推理 k_W 的影响提供了不可或缺的缓冲空间。否则,可能有太多的权重在每次迭代中改变它们的三元值,使训练非常缓慢和不稳定。到

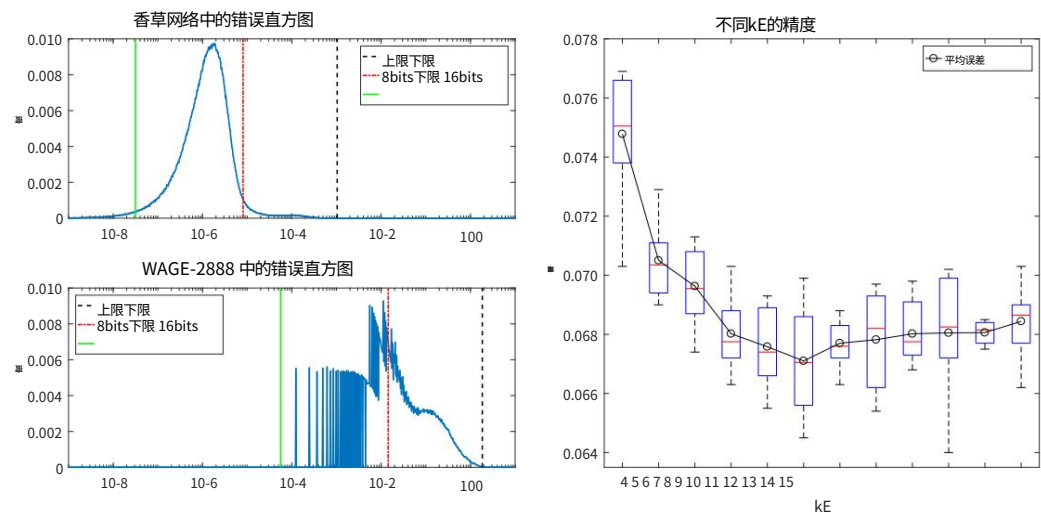


图 4: 左边是 vanilla 网络和 WAGE-2888 网络中同一层的错误 e 直方图。上边界是 $\max\{|e|\}$, 而下边界由位宽 kE 确定。右侧显示了不同 kE 的 10 次运行精度。

进一步探索梯度的合适位宽, 我们使用 CIFAR10 中的 WAGE 2-8-8-8 作为基线, kG 范围从 2 到 12, 每减少 1 位学习率 η 除以 2 以保持近似相等的权重在大量迭代中积累。表 3 的结果显示了 kG 的影响, 并表明了与先前 kE 实验类似的位宽要求。

表 3: CIFAR10 上不同 kG 的测试错误率 (%)。

| 公斤 | 2个 | 3个 | 4个 | 5个 | 6个 | 7 | 8个 | 9 | 10 | 11 | 12 |
|----|-------|-------|-------|-------|-------|------|------|------|------|------|------|
| 误差 | 54.22 | 51.57 | 28.22 | 18.01 | 11.48 | 7.61 | 6.78 | 6.63 | 6.43 | 6.55 | 6.57 |

对于 ImageNet 实现, 我们进行了六种模式来显示位宽要求: 表 1 中的 2888, 288C 用于更准确的错误 (12 位), 28C8 用于更大的缓冲区空间, 28f8 用于梯度的非量化, 28ff 用于错误和 float32 中的梯度无限案例及其 BN 对应物。

原始 AlexNet 复制的准确性被报告为基线。学习率 η 在 28C8 模式中设置为 64 并除以 8, 在 28f8, 28ff 对应物和香草 AlexNet 中设置为 0.01 并除以 10。我们在增加 kG 时观察到过度拟合, 因此分别为 28f8, 28ff 和 28ff-BN 模式添加 $1e-4$, $1e-4$ 和 $5e-4$ 的 L2 权重衰减。在表 4 中, 模式 28C8 和 288C 之间的比较表明, 为梯度积累提供更多缓冲空间 kG 可能比保持高分辨率方向 kE 更重要。此外, 当涉及到 ImageNet 数据集时, 梯度累积, 即梯度的位宽 (kG) 和批量归一化变得更加重要 (Li et al., 2017), 因为训练集中的样本变化很大。

为了避免在训练过程中消耗全精度权重的外部存储器, Deng 等人。 (2018) 在训练和推理中都实现了 1 位权重表示。他们使用更大的 1000 小批量和 float32 反向传播数据流来累积更精确的权重更新, 同样补偿 WAGE 中由 kG 的外部位提供的缓冲区空间。然而, 大批量会显着增加总训练时间, 抵消整数运算单元带来的速度优势。此外, 像特征图这样的中间变量通常比权重消耗更多的内存, 并且与小批量大小线性相关。因此, 我们应用更大的 kG 以获得更好的收敛速度、准确性和更低的内存使用量。

表 4:具有不同kG和kE的 ImageNet 上的 Top-5 错误率 (%)。

| | | | | | | | |
|------|--------------|---------------------|-------|-------|-------|-------|-------|
| 图案香草 | 28ff-BN 28ff | 28f8 28C8 288C 2888 | | | | | |
| 错误 | 19.29 | 20.67 | 24.14 | 23.92 | 26.88 | 28.06 | 27.82 |

5讨论和未来工作

这项工作的目标是展示在 DNN 中使用低位宽整数进行训练和推理的潜力。与 FP16 相比,8 位整数运算不仅可以降低 IC 设计的能量和面积成本 (约 5 倍,见表 5),而且可以将训练期间的内存访问成本和内存大小要求减半,这将大大有利于移动设备具有现场学习能力的设备。有一些工作没有涉及但在未来的算法开发和硬件部署中有待改进或解决。

MAC 操作: WAGE 框架主要测试 2-8-8-8 位宽配置,这意味着虽然在三元权重推理时没有乘法运算,但在训练中仍然需要 MAC 来计算 g。如果我们不考虑 a 和 e 之间的位宽匹配,可能的解决方案是 2-2-8-8 模式。然而,三元 a 会显着降低收敛速度并损害准确性,因为 Q(x, 2) 有两个相对较高的阈值,并且在训练开始时清除每层的大部分输出,这种现象在我们的 BNN 复制中也观察到了。

非线性量化: WAGE中采用距离均匀的线性映射,因为其简单性。然而,非线性量化方法如 \log_2 表示法 (Miyashita 等人,2016 年; Zhou 等人,2017 年)可能更有效,因为经过训练的网络中的权重和激活自然具有对数正态分布,如图 4 所示。此外,与定点表示相比,对数表示中的值具有更大的范围和更少的位数,并且自然地编码在数字硬件中。用对数表示编码的整数来训练 DNN 很有希望。

归一化:在某些 WAGE 演示中,避免或删除了像 Softmax 和批量归一化这样的归一化层。我们认为标准化对于端到端多通道感知至关重要,其中具有不同模式的传感器具有不同的输入分布,以及跨模型特征编码和认知,其中来自不同分支的信息收集以形成更高级别的表示。因此,一种更好的量化归一化的方法在进一步的研究中具有重要意义。

表 5:Size 等人在 45nm 0.9V 中的粗略相对成本。(2017)。

| 手术 | 能量(pJ) | | 面积(μm2) | |
|----------------|--------------|-------------|----------|-----------|
| | 多加 | 多加 | | |
| 8 位 INT 0.2 pJ | 0.03 pJ | 282 16 位 FP | 1.1 pJ | 36 |
| 0.40 pJ 1640 | 1360 32 位 FP | 3.7 pJ | 0.90 pJ | 7700 4184 |

6结论

WAGE 为 DNN 中的训练和推理提供纯低位宽整数数据流。我们引入了一种新的初始化方法和逐层常量缩放因子来代替批量归一化,这是网络量化的痛点。许多其他培训组件也被替代解决方案考虑或简化。此外,还探讨了误差计算和梯度累积的位宽要求。实验表明,我们可以量化梯度的相对值,并在反向传播中丢弃大部分小值及其数量级。虽然权重更新的积累对于稳定收敛和最终精度是必不可少的,但压缩仍然存在,并且可以在训练中进一步减少内存消耗。 WAGE 在多个方面实现了最先进的准确度

具有 2-8-8-8 位宽配置的数据集。它有望通过微调、更高效的映射、批量归一化量化等进行增量工作。总的来说,我们引入了一个没有浮点表示的框架工作,并展示了在基于整数的轻量级上实现离散训练和推理的潜力具有现场学习能力的ASIC或FPGA。

致谢

该工作得到国家自然科学基金项目(61327902)、苏州-清华创新引领计划(2016SZ0102)、国家自然科学基金(61603209)和清华大学自主研究计划(20151080467)的部分支持。我们与 Peng Jiao 和 Lei Deng 进行了很多讨论,感谢他们的深思熟虑的评论。

参考

Martin Abadi,Paul Barham,Jianmin Chen,Zhifeng Chen,Andy Davis,Jeffrey Dean,Matthieu Devin,Sanjay Ghemawat,Geoffrey Irving,Michael Isard 等。Tensorflow:一个用于大规模机器学习的系统。在 OSDI,第 16 卷,第 265-283 页,2016 年。

Yu-Hsin Chen,Tushar Krishna,Joel S Emer 和 Vivienne Sze。Eyeriss:一种用于深度卷积神经网络的节能可重构加速器。IEEE 固态电路杂志,52(1):127-138, 2017。

Matthieu Courbariaux,Yoshua Bengio 和 Jean-Pierre David。Binaryconnect:在传播过程中使用二进制权重训练深度神经网络。在神经信息处理系统的进展中,第 3123-3131 页,2015 年。

邓磊、焦鹏、裴静、吴振智和李国琪。Gxnor-net:在统一的离散化框架下训练具有三元权重和激活的深度神经网络,无需全精度记忆。神经网络,2018 年。

Song Han,Huizi Mao 和 William J Dally。深度压缩:通过修剪、训练量化和霍夫曼编码压缩深度神经网络。arXiv 预印本 arXiv:1510.00149, 2015。

何开明、张翔宇、任少卿和孙健。深入研究整流器:在 imagenet 分类上超越人类水平的表现。在 IEEE 计算机视觉国际会议论文集集中,第 1026-1034 页,2015 年。

Geoffrey Hinton,Li Deng,Dong Yu,George E Dahl,Abdel-rahman Mohamed,Navdeep Jaitly,Andrew Senior,Vincent Vanhoucke,Patrick Nguyen,Tara N Sainath 等。用于语音识别声学建模的深度神经网络:四个研究小组的共同观点。IEEE 信号处理杂志,29(6):82-97, 2012。

Andrew G Howard,Menglong Zhu,Bo Chen,Dmitry Kalenichenko,Weijun Wang,Tobias Weyand,Marco Andreetto 和 Hartwig Adam。Mobilenets:用于移动视觉应用的高效卷积神经网络。arXiv 预印本 arXiv:1704.04861, 2017。

Itay Hubara,Matthieu Courbariaux,Daniel Soudry,Ran El-Yaniv 和 Yoshua Bengio。二值化神经网络。在神经信息处理系统的进展中,第 4107-4115 页,2016 年。

谢尔盖·约夫和克里斯蒂安·塞格迪。批量归一化:通过减少内部协变量偏移来加速深度网络训练。在机器学习国际会议上,第 448-456 页,2015 年。

Norman P Jouppi,Cliff Young,Nishant Patil,David Patterson,Gaurav Agrawal,Raminder Bajwa,Sarah Bates,Suresh Bhatia,Nan Boden,Al Borchers 等。张量处理单元的数据中心内性能分析。在第 44 届年度国际计算机体系结构研讨会论文集集中,第 1-12 页。美国计算机学会,2017 年。

Alex Krizhevsky,Ilya Sutskever 和 Geoffrey E Hinton。具有深度卷积神经网络的 Imagenet 分类。在神经信息处理系统的进展中,第 1097-1105 页,2012 年。

Yann LeCun, Leon Bottou, Yoshua Bengio 和 Patrick Haffner. 基于梯度的学习应用于文件识别。IEEE 会议记录, 86(11):2278–2324, 1998。

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply supervised networks. 在人工智能和统计中, 第 562–570 页, 2015 年。

Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. arXiv preprint arXiv:1605.04711, 2016.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet 和 Tom Goldstein. 训练量化网络:更深入的理解。在神经信息处理系统的进展中, 第 5813–5823 页, 2017 年。

宫下大辅、爱德华·李 (Edward H Lee) 和鲍里斯·穆尔曼 (Boris Murmann). 使用对数数据表示的卷积神经网络。arXiv 预印本 arXiv:1603.01025, 2016。

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon 和 Ali Farhadi. Xnor-net: 使用二进制卷积神经网络的 Imagenet 分类。在欧洲计算机视觉会议上, 第 525–542 页。施普林格, 2016 年。

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein 等。图像大规模视觉识别挑战。国际计算机视觉杂志, 115(3):211–252, 2015。

Patrick M Sheridan, Fuxi Cai, Chao Du, Wen Ma, Zhengya Zhang, and Wei D Lu. Sparse coding with忆阻器网络。自然纳米技术, 12(8):784, 2017。

Luping Shi, Jing Pei, Ning Deng, Dong Wang, Lei Deng, Yu Wang, Youhui Zhang, Feng Chen, Mingguo Zhao, Sen Song, et al. Development of a neuromorphic computing system. In Electron Devices Meeting (IEDM), 2015 IEEE International, pp. 4–3. IEEE, 2015.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot 等。通过深度神经网络和树搜索掌握围棋游戏。自然, 529(7587):484–489, 2016 年。

卡伦·西蒙尼安和安德鲁·齐瑟曼。用于大规模图像的非常深的卷积网络。arXiv 预印本 arXiv:1409.1556, 2014。

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang 和 Joel S Emer. 神经网络的高效处理:教程和调查。IEEE 会刊, 105(12):2295–2329, 2017 年。

韦唐、纲华、梁王。如何训练一个紧凑的二元神经网络。准确性? 在 2017 年第 30 届 AAAI 人工智能会议上。

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen 和 Hai Li. Terngrad: 减少分布式深度学习中通信的三元梯度。在神经信息处理系统的进展中, 第 1508–1518 页, 2017 年。

Chiyan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht 和 Oriol Vinyals. 理解深度学习需要重新思考泛化。arXiv 预印本 arXiv:1611.03530, 2016。

Aojun Zhou, Anbang Yao, Yiwon Guo, Lin Xu 和 Yurong Chen. 增量网络量化: 迈向具有低精度权重的无损 CNN。arXiv 预印本 arXiv:1702.03044, 2017。

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606.06160, 2016.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. arXiv 预印本 arXiv:1612.01064, 2016。

算法_

我们假设使用等式 5 定义和初始化网络结构。伪代码后的注释是在定点数据流中实现的潜在对应操作。

算法 1在基于浮点或基于整数的设备上使用 WAGE 方法训练 l 层网络。权重、激活、梯度和误差根据等式 6 - 12 进行量化。

要求:一个小批量的输入和目标 (每层的基础 α ,学习率调度器 q_i , η) 被量化为kA 位整数,移位

η ,先前的权重W以kG位保存。

确保:更新权重 W_{t+1} 1. 前向传播: 1:对

于 $i = 1$ 到 l do 2: W_i

$\leftarrow QW(W_i)$

我3:一个 $\leftarrow \text{ReLU}(\text{一个 } q^{i-1} \text{ 无线})_q$

我4 $\leftarrow QA(a 5: \text{ })$

end for 2. Back

propagation:

电脑和 $\leftarrow \frac{\partial L}{\partial a_i}$ 知道一个 我 和一个 $*$

6:对于 $i = l$ to 1 do $\leftarrow QE(e$

7: $e_{q_i-1}^i$

8: $e \leftarrow \text{和 }) i q W_i q$

9: $g_{10:}^{iii-1} \leftarrow e a q^T q$

$\Delta W_i \leftarrow QG(g$

11:根据公式 12更新和裁剪 W_i

12:结束

#夹子

#MAC,剪辑

#Shift,剪辑

#基质

#Max,移动,剪辑

#MAC,剪辑

#MAC,剪辑

#Max,Shift、随机、剪辑

#更新,剪辑

13

B层直方图

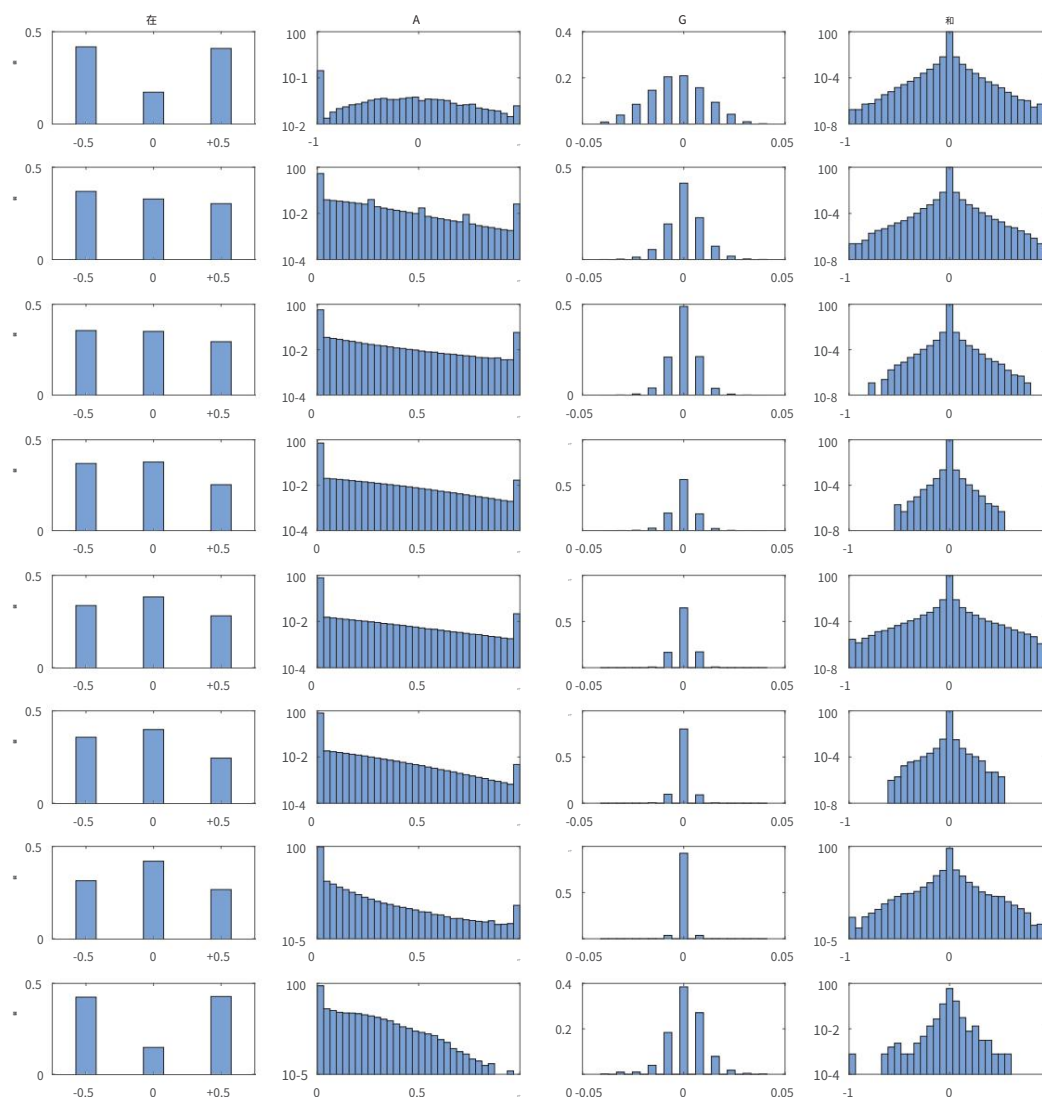


图 5:经过训练的类 VGG 网络的分层直方图,其位宽配置为:2-8-8-8,学习率 η 等于 8。Y 轴代表 W-plots 和 G-plots 中的概率,以及对数概率分别在 A-plots 和 E-plots 中。在 A-plots 中,直方图在前一层,因此第一个图显示了量化的输入数据。