基于权值交互思想的卷积神经网络量化算法

肖国麟,杨春玲,陈　宇
(哈尔滨工业大学 电气工程及自动化学院,黑龙江 哈尔滨 150001)

摘　要：传统的卷积神经网络量化算法广泛使用对称均匀量化操作对模型权值进行量化，没有考虑到相邻权值量化之间的相互关系，即上一个权值的量化操作产生的量化噪声可以通过调整之后权值的量化方向加以弥补。针对上述问题,提出了一种基于权值交互思想的三值卷积神经网络量化算法,达到了16倍的模型压缩比,以ImageNet作为数据集,量化后的AlexNet和ResNet-18网络上模型预测准确率只下降了不到3%。该方法达到了较高的模型压缩比,具有较高的精度,可以用于将卷积神经网络移植到计算资源有限的移动端平台上。

关键词：三值量化;卷积神经网络;权值交互;模型压缩
中图分类号：TN911.73;TP391　　**文献标识码**：A　　DOI:10.16157/j.issn.0258-7998.200263

中文引用格式：肖国麟,杨春玲,陈宇. 基于权值交互思想的卷积神经网络量化算法[J].电子技术应用,2020,46(10):39-41.
英文引用格式：Xiao Guolin,Yang Chunling,Chen Yu. Convolutional neural network quantization algorithm based on weight interaction ideas[J]. Application of Electronic Technique,2020,46(10):39-41.

Convolutional neural network quantization algorithm based on weight interaction ideas

Xiao Guolin,Yang Chunling,Chen Yu
(School of Electrical Engineering and Automation,Harbin Institute of Technology,Harbin 150001,China)

Abstract: Traditional convolutional neural network quantization algorithms widely use symmetric uniform quantization operations to quantize models′ weights, without taking into account the correlation between the quantization of adjacent weights, that is, the quantization noise generated by the quantization operation of the previous weight can be made up after adjusting the quantitative direction of the next weights. Aiming at the above problems, a ternary convolutional neural network quantization algorithm based on the idea of weight interaction is proposed, the model compression ratio is 16 times. On the ImageNet dataset, the model prediction accuracy of ternarized AlexNet and ResNet-18 network only decrease less than 3%. This method achieves a high model compression ratio, has higher accuracy, and can be used to transplant convolutional neural networks to mobile platforms with limited computing resources.

Key words: ternary quantization;convolutional neural network;weight interaction;model compression

0 引言

随着深度学习的飞速发展,卷积神经网络(CNN)被越来越多地应用于各个领域,如图像识别[1-2]和目标检测[3-4]。然而,随着应用要求的提高,CNN的结构越来越深,导致其对于计算力和内存的需求大大提高。同时由于移动端设备的飞速发展,设备小型化的需求和市场不断扩大,将CNN应用到小型移动端设备的需求也随之增加。然而,受限于电源、内存及功耗,移动端平台无法满足高性能CNN对于硬件性能、功耗及内存的要求,因此优化CNN模型从而降低其对于计算资源的要求非常必要[5]。

CNN权值量化是目前一种主流的CNN模型优化方法,而三值量化其中一种能够将权值由32位量化到2位$(0,\pm\alpha)$从而达到16倍压缩比的低位量化方法。自从2016年三值化网络被LI F等人提出以来[6],其方法不断改进,性能不断提高。其中MELLEMPUDI N等人于2017年提出的细粒度三值化网络[7],将激活函数量化到8或4位,将权重量化至2位,将每N个权值分为一个组,分组量化,两组之间相互独立,其中$N=2,4,8,16,\cdots$。该方法主要通过尝试不同的分组和暴力搜索解空间以及二次训练的方式得到最优解，但相比原精度的32位卷积网络仍有较大性能差距。目前的三值量化网络性能不佳的其中一个问题在于都使用对称均匀量化操作,只考虑了单个权值的量化,忽视了相邻权值量化之间的相互关系,量化噪音直接相互累加,导致每一层网络的噪声积累过高,使得量化模型的预测准确率有极大的下降。

针对此问题,本文提出一种新的CNN模型三值量化算法,基于权值交互思想,在同一卷积核内,将先前量化产生的积累噪声作为一个负变量加入到下一个权值的

量化操作中，使得对下一个权值朝着能减小积累噪声的方向进行量化。然后，通过层级贪婪搜索算法逐层搜索局部最优解，得到效果近似最优解的优化解，降低搜索算法复杂度，极大减少搜索所需时间。实验证明，相比于其他使用对称均匀量化操作的算法，本文的算法极大地减小了由于量化操作导致的模型预测准确率的损失。

1 权值交互量化算法

已知 CNN 某卷积层中的一个通道的卷积操作为：

$$x_{out}=\sum_i w_i x_i \tag{1}$$

其中，x_{out} 表示该通道输出，x_i 表示该通道的第 i 个输入，w_i 表示 x_i 对应的权值。量化操作表示将一定区域内的连续分布的数值分成多个区域，分别用一个数值表示该区域的所有数值，其中三值对称均匀量化公式为：

$$Q(w_i)=\begin{cases}\alpha & w_i>\Delta\\ -\alpha & w_i<\Delta\\ 0 & \text{其他}\end{cases} \tag{2}$$

其中，0 和 $\pm\alpha$ 称为目标量化值，Q 表示量化操作，Δ 为阈值，将 w_i 的分布划分为 3 个区域。其量化误差为：

$$E=r_{real}-r_{quan}=\sum_{i=1}^{n}\varepsilon_i x_i \tag{3}$$

其中：

$$\varepsilon_i=Q(w_i)-w_i \tag{4}$$

$$r_{real}=\sum_{i=1}^{n} w_i x_i \tag{5}$$

$$r_{quan}=\sum_{i=1}^{n} Q(w_i)x_i=\sum_{i=1}^{n}(w_i+\varepsilon_i)x_i \tag{6}$$

其中，E 代表当前通道的量化误差，r_{real} 代表输入上一层输入 $X=(x_1,x_2,\cdots,x_n)$ 通过该未量化通道后的真实输出，r_{real} 代表未量化的结果，r_{quan} 代表输入 X 通过量化通道后的结果，w_i 代表第 i 个权值，ε_i 代表第 i 个权值的量化噪声。

现有的量化方法都使用了对称均匀量化操作，不考虑权值量化间的相互影响，独立地量化每个权值，通过分析，发现该方法有其缺陷。例如，设 α 为 1，Δ 为 0.5，输入为 $(1,1,1)$，权值为 $(0.6,0.6,0.6)$，那么权值会被直接量化为 $(1,1,1)$，量化误差即为 -1.2，所有量化操作互不影响，量化误差逐步累加，大小不可控。因此，本文提出一种基于交互性思想的权值交互量化算法，综合考虑权值量化间的相互影响，控制量化误差。该算法如下：

$$Q_{new}(w_i)=Q(w_i+\gamma) \tag{7}$$

$$\gamma=\beta\times\sum_{j=1}^{i-1}\varepsilon_j \tag{8}$$

其中，Q_{new} 代表权值交互量化操作；ε_i 代表第 i 个权重的量化误差；γ 是交互因子，即 i 之前所有权值的量化噪声之和与一个限制因子 β 的乘积；β 是一个正的参数，其作用是限制 γ 影响过大，称之为限制因子。利用权值交互量化算法，输入为 $(1,1,1)$，权值为 $(0.6,0.6,0.6)$，当 β 取 0.5 时，权值被量化为 $(1,0,1)$，量化误差为 -0.2，相比前文结果有了极大减小。

在具体实现过程中发现，在量化过程中，当某个权值满足条件：

$$|w_i-Q(w_i)|\leqslant\delta \tag{9}$$

其中，δ 表示保护阈值，即当对称均匀量化噪声小于一个阈值时，w_i 非常接近目标量化值，那么该权值被称为“临近值”，需要直接被直接量化，避免交互因子对其产生影响。例如，若某个权值为 0.99，目标量化值为 0 和 ±1，当交互因子较大时，0.99 可能会被量化成 0 而非 1，从而产生极大偏离，对实验结果产生负面影响。故在此基础上增添一个保护阈值 δ，使得临近值免受交互因子的影响。

权值交互量化算法的实现流程如下：

输入：限制因子 β、量化阈值 Δ、量化目标值 α、临近值保护阈值 δ、当前通道所有权值集合 list_w、对称均匀量化函数 Q。

输出：量化后的通道权值集合 list_w。

开始：

(1) 积累影响初始化 $\gamma\leftarrow0$

(2) 执行迭代：

(3) 输入权重 $w_i\leftarrow$list_$w[i]$

(4) 检查该权值是否为临近值$|w_i-Q(w_i,\Delta,\alpha)|\leqslant\delta$

(5) 如果是：

(6) 不考虑积累噪声进行量化 $w_q\leftarrow Q(w_i,\Delta,\alpha)$

(7) 如果否：

(8) 考虑积累噪声量化 $w_q\leftarrow Q(w_i+\gamma,\Delta,\alpha)$

(9) 更新交互因子 $\gamma+=(w_i-w_q)\times\beta$

(10) 更新权值 list_$w[i]\leftarrow w_q$

结束。

2 权值交互量化算法的优化技术研究

本课题所提出的权值交互量化算法里有 4 个无法通过训练获取优化解的超参数：目标量化值 α、阈值 Δ、限制因子 β(交互因子 γ 可由其得到)和保护阈值 δ。这 4 个超参数需要通过搜索算法搜寻解空间中的最优解。由于 CNN 中每个卷积层或全连接层的权值分布都不相同，4 个超参数的最优解也各不相同，因此每层网络都需要基于自身权值分布的超参数。设 CNN 深度为 N，那么暴力搜索算法搜寻全部解空间的时间复杂度为 $O(4^N)$，随着 CNN 网络层数增加，算法所花时间呈指数增长，靠暴力搜索寻找最优解非常耗时，在实际应用中无法实现。然而，ZHOU Y 等人于 2018 年发表的文章指出，对于 CNN 量化，每层网络的量化误差对于整个网络预测结果的影响是相互独立的[8]。基于此结论，结合贪婪算法思想，本课题提出层级贪婪搜索算法来替代暴力搜索算法，通过独立地逐层搜寻网络的局部最优解来近似整个网络的全局最优解，将指数时间复杂度 $O(4^N)$ 降低为线

性时间复杂度 $O(N)$，极大地减小了搜索超参数所花时间。

本文搜索局部最优解的方法为在取得 32 位训练模型的基础上，抽取部分训练样本作为搜索测试样本，在模型上进行预测计算，得到标准的预测结果。再以一定步长逐步改变超参数的数值，量化模型，得到结果损失，最后选取损失最小的一组超参数作为该层网络的最优超参数。

逐层贪婪搜索算法的实现流程如下：

输入：已训练的神经网络 NN。

输出：限制因子 β、量化阈值 Δ、量化目标值 α、临近值保护阈值 δ 的每层局部最优解的集合 L。

开始：

(1) 执行迭代：

(2) 初始化解集合 L

(3) 输入单层网络权值 NN[i]

(4) 执行迭代：

(5) 搜索当前层的局部最优解 $r=\{\beta_{best},\Delta_{best},\alpha_{best},\delta_{best}\}$

(6) 更新解集合 $L[i]\leftarrow r$

(7) 返回结果 L

结束。

3 实验结果

本文中，实验平台使用的 CPU 为 Intel® Core™ i9-9900k，GPU 为英伟达 GTX2080TI 11G，内存为 32 GB，系统为 Linux，其版本为 Ubuntu 16.04LTS。实验之前，必须先搭建深度学习运行环境，所需的主要软件有基于 Linux 的 Anaconda 软件包、CUDA 运算平台以及 TensorFlow 机器学习平台。使用 ImageNet 作为数据集，在 AlexNet 和 ResNet-18 网络上的实验结果表 1 所示。

表 1 不同三值化网络预测准确率比较 (%)

网络结构	算法	量化前	量化后	准确率损失
AlexNet	本文算法	56.73	53.91	2.82
	细粒度三值化[7]	56.83	49.04	7.79
ResNet-18	本文算法	65.8	63.02	2.78
	三值化网络[6]	65.4	61.8	3.6

由上述实验结果可以看出，相比于其他三值量化算法，本文提出的权值交互量化算法在 AlexNet 和 ResNet-18 网络上均取得了最小的准确率损失，尤其在 AlexNet 上准确率损失相比细粒度三值化算法减小了一半以上，在同等的模型压缩比之下，性能有了极大提升。

4 结论

本文提出了一种基于权值交互思想的卷积神经网络三值量化算法，通过引入积累量化噪声，使得下一次量化操作朝着抵消噪声的方向进行，并提出层级贪婪搜索算法找寻每一层网络的局部最优解近似全局最优解，极大地减小超参数的搜索时间。在实验阶段，通过与其他两种使用对称均匀量化操作的三值化算法进行对比试验，证明本文所提出的算法有效地减小了由量化导致的模型精度损失。但本文不足之处在于没有通过理论证明本文提出的方法减小了单层网络的总量化误差，下一步将尝试对其进行数学论证。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems, 2012: 1097-1105.

[3] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]. European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. Advances in Neural Information Processing Systems, 2015: 91-99.

[5] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding[J]. arXiv Preprint arXiv: 1510.00149, 2015.

[6] LI F, ZHANG B, LIU B. Ternary weight networks[J]. arXiv Preprint arXiv: 1605.04711, 2016.

[7] MELLEMPUDI N, KUNDU A, MUDIGERE D, et al. Ternary neural networks with fine-grained quantization[J]. arXiv Preprint arXiv: 1705.01462, 2017.

[8] ZHOU Y, MOOSAVI-DEZFOOLI S M, CHEUNG N M, et al. Adaptive quantization for deep neural network[C]. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

（收稿日期：2020-04-03）

作者简介：

肖国麟(1995-)，男，硕士研究生，主要研究方向：深度学习模型压缩。

杨春玲(1965-)，通信作者，女，博士，教授，主要研究方向：电子设计自动化 EDA 技术、深度学习及硬件加速，E-mail: yangcl1@hit.edu.cn。

陈宇(1990-)，男，博士研究生，主要研究方向：深度学习及硬件加速、迁移学习。

版权声明