

融合贝叶斯的压缩 CNN 植物叶片识别模型

晏 铭, 朱良宽, 景维鹏

(东北林业大学 机电工程学院, 黑龙江 哈尔滨 150040)

摘 要: 针对植物叶片识别过程中参数较多且易产生过拟合的问题, 为减少其存储及计算的代价, 本文提出一种基于贝叶斯融合压缩卷积神经网络植物叶片识别模型。首先, 利用递归贝叶斯算法进行网络剪枝, 自适应地去除网络冗余; 然后, 引入 K-Means 聚类量化网络中的卷积层和完全连接层, 用于压缩整个卷积神经网络; 最后, 结合经典卷积神经网络模型 AlexNet, 进行植物叶片识别实验。实验结果表明, 在相同的数据集中, 压缩后的卷积神经网络相比于原网络及其它压缩方法, 占用的存储空间更少, 压缩率达到 36 倍, 并取得了略好的性能, 准确率达到 88.58%。

关键词: 卷积神经网络; 模型压缩; 剪枝; 参数量化; 贝叶斯算法

DOI: 10.15938/j.jhust.2021.03.012

中图分类号: TP183 文献标志码: A 文章编号: 1007-2683(2021)03-0083-08

Compressed CNN Plant Leaf Recognition Model Fused with Bayesian

YAN Ming, ZHU Liang-kuan, JING Wei-peng

(College of Mechanical and Electronic Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: Aiming at the problem that there are many parameters in the process of plant leaf recognition and it is easy to produce over-fitting, in order to reduce the cost of storage and calculation, this paper proposes a plant leaf recognition convolutional neural network model based on Bayesian fusion. Firstly, the recursive Bayesian algorithm is used for network pruning to adaptively remove network redundancy. Then, the convolutional layer and the fully connected layer in the K-Means cluster quantization network are introduced to compress the entire convolutional neural network. Finally, combined with the classical convolutional neural network model AlexNet, plant leaf recognition experiments were carried out. The experimental results show that in the same data set, the compressed convolutional neural network takes less storage space, 36 times compression and achieves a slightly better performance with 88.58% accuracy than the original network and other compression methods.

Keywords: convolutional neural network; model compression; pruning; parameter quantization; Bayesian algorithm

0 引 言

植物是地球生态系统的一个重要组成部分, 目前已知的植物种类数量在 40 万左右, 还存在大量未

被分类或未知的物种。植物的分类识别是生物和环境科学中非常重要的一个研究课题, 对生物多样性保护、生态农业、生物安全等有着重要的意义。作为植物的六大器官之一的植物叶片, 因其方便采集、状态相对稳定, 且不同种类植物一般在其叶片的视觉

收稿日期: 2019-09-24

基金项目: 国家自然科学基金(31370710); 黑龙江省博士后启动基金(LBH-Q13007)。

作者简介: 朱良宽(1978—), 男, 教授;

景维鹏(1979—), 男, 副教授, 硕士研究生导师。

通信作者: 晏 铭(1994—), 女, 硕士研究生, E-mail: 18645970237@163.com.

特征上存在着差异,为植物的分类识别提供了非常有价值的线索,是植物分类的重要依据。传统的植物叶片分类方法准确率不高的原因在于其需要对数据进行预处理,特征提取,特征分类等繁琐步骤,预处理效果差异大,人工设计的提取器难以获得最接近目标自然属性的特征表达。随着科技的进步,植物叶片识别分类任务已经从传统的由植物学专家进行人工识别到近年来刚刚成熟的使用卷积神经网络进行分类。卷积神经网络由于其对几何、形变、光照具有一定程度的不变形,因此被广泛应用于图像领域。其主要特点有:①数据集不需要进行复杂的处理;②特征提取和识别可以同时进行;③权值共享,大大减少了需要训练的参数量^[1]。然而,卷积神经网络常常为了换取高准确度而对模型加深拓宽,这就导致了模型参数数量太大,计算过于复杂,内存占用量偏高的问题^[2]。

大多数 CNN(convolutional neural network) 的模型压缩解决方案有两个方面,一是解决卷积层计算量过于复杂的问题,二是解决完全连接层占据过多网络参数的问题。由于卷积层和完全连接层的结构明显不同,大多数现有的方法中只注重于其中一种解决方案。在文 [3-8] 中采用低秩分解将卷积层分解为多个较小的层。Lebedev 和 Lempitsky^[3] 以及 Anwar 等^[4] 将稀疏性引入卷积层以降低计算复杂度。在文 [5] 和文 [1] 中探索了矩阵和张量分解以压缩完全连接的层。其他方法包括固定点表示^[12-14],二进制值网络^[15-18],基于码本的量化^[19-21],以及紧凑网络设计^[22-25]。

针对模型压缩过程中出现的计算量过大以及参数过多的问题,本文将网络剪枝和量化融合在一个学习框架中,具体包括:

1) 使用扩展卡尔曼滤波器训练方法。EKF(extended kalman filter,扩展卡尔曼滤波) 是递归贝叶斯算法,基于后验概率的灵敏度定义权重重要性度量。利用这种自适应特征,设计了一种基于贝叶斯的深度网络模型剪枝算法;

2) 在学习框架中结合网络剪枝和量化,共同解决神经网络模型压缩问题;

3) 针对整个网络进行参数微调。将以上融合方法用于 AlexNet 卷积神经网络的模型压缩,并在瑞典植物叶片数据集上进行分类实验,验证了所提方法的可行性及有效性。

本研究尝试提出一种新型压缩卷积神经网络的

植物叶片识别模型。充分利用了剪枝和量化算法的优势,并将以上融合方法用于 AlexNet 卷积神经网络的模型压缩,在瑞典植物叶片数据集上进行分类实验,验证了所提方法的可行性及有效性。

1 卷积神经网络的权重剪枝

1.1 传统的权重剪枝方法

剪枝是用于降低网络复杂度的有效方法,可以在几乎不损失模型精度的前提下移除网络中的冗余参数。基于模型剪枝的方法有很多,其主要思想都是挑选出模型中不重要的权重或卷积核将其移除,通过再训练来恢复模型的性能,这样就可以在保证模型性能的前提下,最大程度的压缩模型参数。

权重剪枝包括 3 个步骤:

1) 通过将从正常训练中获得的权重分为相对重要与相对不重要的两个部分,以此来评估权重的重要性;

2) 设置阈值,将权重小于阈值的部分设置为零,然后网络成为稀疏连接的网络;

3) 重新训练新的稀疏网络得到最终的结果。

在剪枝算法中,LMP(layer-wise magnitude-based pruning,根据重要性逐层切割连接) 是一种重要的传统排序算法。LMP 算法常用于深度神经网络的压缩过程,但调整 LMP 算法中每层的阈值是一项艰巨的任务,因为对于每个阈值组合,评估其优缺点需要高计算成本。在传统的方法中,阈值的选择主要由具有专业知识的实验室技术人员手动选择,难以保证阈值选取的合理性。

1.2 基于贝叶斯算法的权重剪枝方法

本文采取 EKF 方法作为训练方法,该方法可以跟踪时变参数,且可为权重向量和误差协方差矩阵提供修剪信息。给定训练数据样本集 $Y^t = \{x(i), y(i)\}_{i=1}^t$,神经网络的剪枝可描述为一个滤波问题,即把网络的所有参数向量 $\theta(t)$ 看作是系统的状态。此时可以取得 $\theta(t)$ 的估计为

$$\hat{\theta}(t) = \arg \max_{\theta} P(\theta(t) | Y^t)$$

后验概率的估计遵循递归贝叶斯方法:

$$P(\theta(t) | Y^t) = \frac{\int P(\theta(t), x(t) | \theta(t)) P(\theta(t) | \theta(t-1))}{\int \int P(\theta(t), x(t) | \theta(t)) P(\theta(t) | \theta(t-1))} \times \frac{P(\theta(t-1) | Y^{t-1}) d\theta(t-1)}{P(\theta(t-1) | Y^{t-1}) d\theta(t-1) d\theta(t)}$$

其中, $P(\theta(t) | Y)$ 和 $P(y(t), x(t) | \theta(t))$ 是服从高斯分布的, 通过非线性函数 $f(y(t), \theta(t-1))$ 来实现局部线性化。利用上式中扩展卡尔曼训练的贝叶斯优化, 可以定义一个衡量权重的重要性标准。由于 $P(\theta(t) | Y)$ 是高斯分布, 近似于给定测量数据 Y^t 的实际后验概率, 我们可以写出方程式表达式:

$$P(\hat{\theta}(t) | Y^t) = c_0 \exp\left\{-\frac{1}{2}(\theta(t) - \hat{\theta}(t))^T \times P^{-1}(t)(\theta(t) - \hat{\theta}(t))\right\}$$

其中 c_0 是一个归一化常数。令 $\hat{\theta}_k(t)$ 为参数向量, 除了第 k 个元素为零之外, 所有元素都等于 $\hat{\theta}(t)$ 。于是得出:

$$P(\hat{\theta}_k(t) | Y^t) = c_0 \exp\left\{-\frac{1}{2}\hat{\theta}_k^T(P^{-1}(t))_{kk}\hat{\theta}_k\right\}$$

其中 $(P^{-1}(t))_{kk}$ 是 P 的逆的第 k 个对角元素。显然, 当 $\hat{\theta}_k^T(P^{-1}(t))_{kk}\hat{\theta}_k$ 值越小, 后验概率 $P(\hat{\theta}_k(t) | Y^t)$ 就越大。因此, 得出了一次修剪一个权重的重要性度量。对于一次修剪多个权重, 将 $\{\pi_1, \dots, \pi_{n_\theta}\}$ 设定为顺序排列表, $\hat{\theta}_{\pi_k}$ 是元素从 π_1 到 π_k 为零的向量, 其余元素与 $\hat{\theta}_{\pi_{k+1}}$ 到 $\hat{\theta}_{\pi_{n_\theta}}$ 相同。由可得从 π_1 到 π_k 索引的权重的重要性如下:

$$P(\hat{\theta}_{\pi_k}(t) | Y^t) = c_0 \exp\left\{-\frac{1}{2}\hat{\theta}_{\pi_k}^T P^{-1}(t) \hat{\theta}_{\pi_k}\right\}$$

上述等式定义了自适应剪枝过程的基本部分:

1) 使用递归方程来获得 $\hat{\theta}(t)$ 和 $P(t)$ 。

2) 通过 $t^{-1} \sum_{i=1}^t (y(i) - \hat{\theta}(i))^2$ 估计训练误差 $E_{tr}(t)$ 。

3) 如果 $E_{tr}(t) < E_{tr}(0)$:

① 估计 $P^{-1}(t)$ 并得出 $\theta_k^2(P^{-1}(t))_{kk}$ 从 1 到 n_θ 的所有 k ;

② 根据 $\theta_k^2(P^{-1}(t))_{kk}$ 的升序重新排序索引 $\{\pi_k\}$;

③ 对于从 1 到 n_θ 的 π_k , 假设删除了从 θ_{π_1} 到 θ_{π_k} , 估计 $P(\hat{\theta}_{\pi_k}(t) | Y^{(t)})$;

④ 若 $\log P(\hat{\theta}_{\pi_k}(t) | Y^{(t)}) - \log c_0 < E_{tr}(0)$, 删除 θ_{π_1} 到 θ_{π_k} 。

2 卷积神经网络的参数量化

通常, 存储网络参数的基本单位是 32 位宽的单精度浮点数据类型。但是, 在大多数情况下, 使用低位数据类型的存储不会导致网络准确性降低。具有低精度数据类型的参数表示和存储也是网络压缩的

重要思想之一。

在本文中, 采用 K-Means 聚类量化卷积神经网络中的卷积层和完全连接层。K 均值聚类对象分为向量和单个值, 本文主要关注单数值聚类。通过聚类, 可将该层的所有参数权重映射到有限小码本中, 并且码本中的码字是通过 K 均值聚类获得的中心点。最接近该中心点的每个参数共享该权重。在进行网络参数存储的时候, 仅存储该层的码本(不同层的码本不同)以及该层参数在码本中对应的索引。例如, 有一个 16 位大小的码本, 则每个参数的索引存储仅需 4 个比特即可, 进而达到网络压缩的效果。以下描述了参数量化的实现过程。

算法: 卷积神经网络的权重聚类量化

1) 预训练网络模型;

2) 对各层参数进行 K-Means 聚类, 得到一个大小为 K 的码本, K 值决定了网络的量化精度以及网络压缩的比重;

3) 各层参数取当前码本中欧式距离最近的码字为值, 进行网络前向推导, 得到损失值;

4) 在进行误差反向回传计算参数梯度值的时候, 各层参数取其真实值进行计算, 将共享同一个码字的参数梯度值相加, 通过随机梯度下降法对该类码字进行更新;

5) 交替迭代步骤 3 和步骤 4, 直到收敛。收敛条件为, 码本的变化趋于零或者达到预先设置好的训练次数, 或者已达到预期的网络精度;

6) 将训练好的模型进行应用时, 各个参数使用的都是量化后的值。

量化之后网络的压缩率 Q 为

$$Q = \frac{N \log_2(K) + 32K}{32N}$$

其中: N 为参数个数; 32 为一个单精度浮点类型的比特数。

只对网络模型进行剪枝操作虽然可以减少参数冗余, 提高网络的稀疏性, 但是它对实际压缩效率的影响是有限的。因此, 本文结合量化操作来解决单一方法的局限性。

所提出的算法流程图如图 1 所示。首先, 对数据集进行预处理, 包括对数据集的扩充以及划分; 其次, 应用自适应贝叶斯算法对网络中的每一层进行超参数预测, 直到确定了所有层的权重剪枝阈值并进行权重修剪; 然后, 对权重参数进行聚类量化; 最后, 测试网络性能, 微调网络参数直到其压缩率及准确率达到满意的效果。

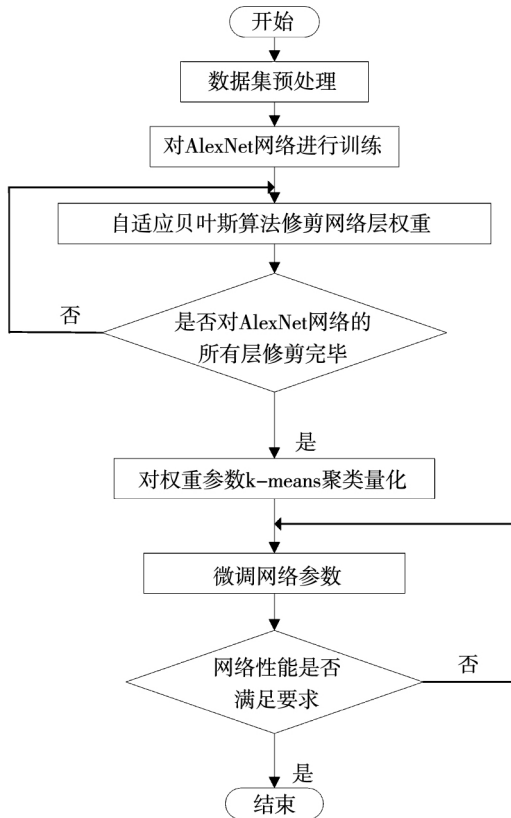


图1 算法流程图

Fig.1 Algorithm flowchart

3 实验设计

3.1 数据集预处理

本文实验所采用的树叶数据集来源于 CVL (computer vision laboratory, 计算机视觉实验室), 该

数据集包括 15 种不同类型的树叶共 1125 张。通过传统方法中的垂直翻转、水平翻转、向右旋转 90° 、向左旋转 90° 等操作将数据集扩充至 112500 张, 然后将所有图像大小调整为 256×256 并将整个数据集按照 6:2:2 的比例划分给训练集、验证集以及测试集。如图 2 所示, 给出了 3 种树叶的预处理效果图。

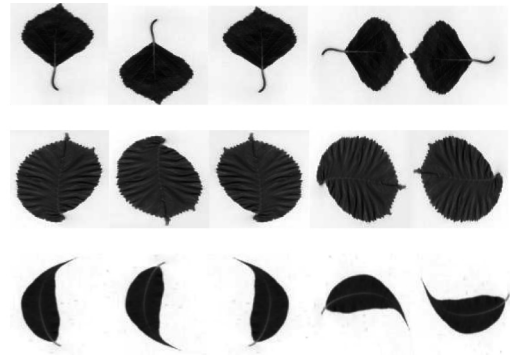


图2 三种树叶图像的预处理效果图

Fig.2 The preprocessing renderings of the three leaf images

3.2 压缩 AlexNet 网络模型

在图像识别领域中, 使用卷积神经网络已经成为了一种主流方法。2012 年, Krizhevsky 等^[24]提出了一个大型的深度卷积神经网络——AlexNet, 作为一种经典的网络模型, AlexNet 由 5 个卷积层和 3 个全连接层组成, 其中集成了各种技术来提高网络的性能以及网络的训练效率, 包括: 非饱和非线性神经元 ReLu、LRN (local response normalization, 局部响应归一化)、dropout 等, 其网络模型结构如图 3 所示。

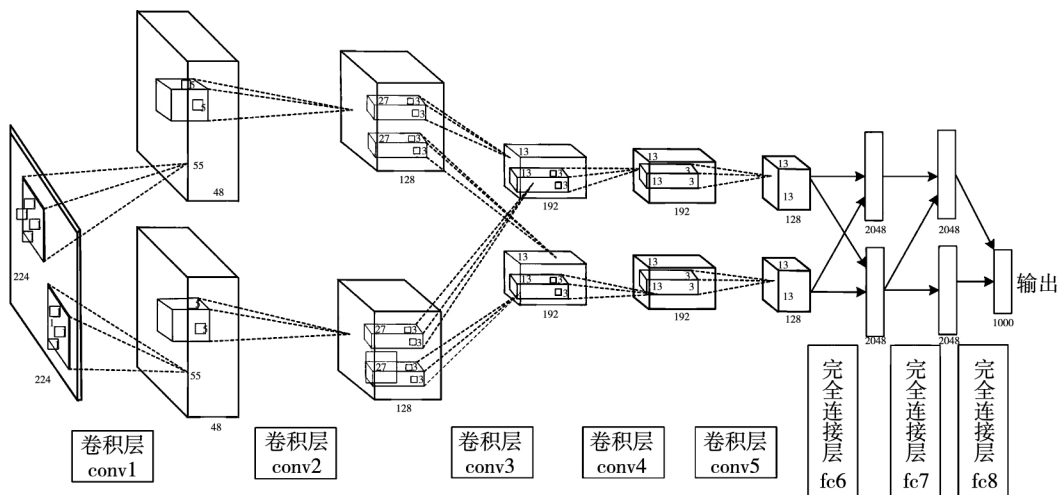


图3 AlexNet 网络模型结构

Fig.3 AlexNet network model structure

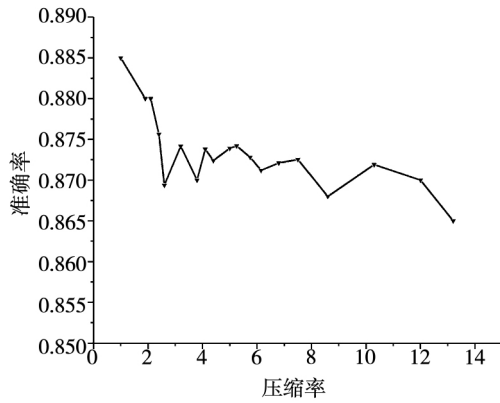
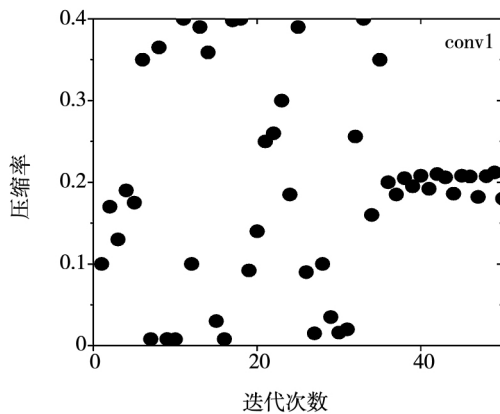


图4 应用传统的剪枝方法得到的压缩率与准确度

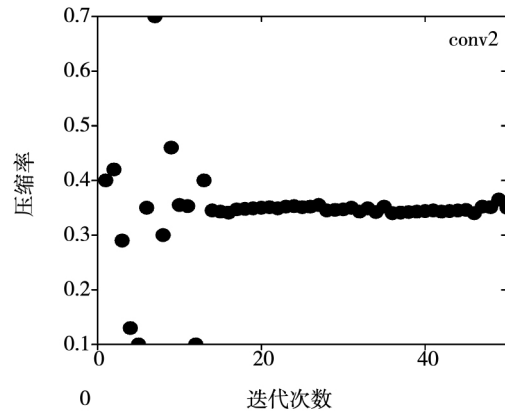
Fig. 4 Compression rate and accuracy obtained by using traditional pruning methods

2016年, Anwar^[14]等人修剪了一个基于 VGG16 模型的猫狗分类器。本文将以 AlexNet 作为深度卷积神经网络模型, 进行网络模型压缩及植物叶片识别实验。首先, 将神经网络中冗余的参数进行修剪, 尝试一些阈值设置来观察使用传统方法对于准确度和压缩率的变化。然后, 使用 EKF 方法作为 AlexNet 上的训练方法来证明基于贝叶斯方法进行剪枝的有效性。使用传统方法进行剪枝的结果如图 4 所示, 按照阈值增加的顺序连接了 19 个测试的结果。可以看出, 较大的阈值将带来更高的压缩率和更低的准确性。当阈值增加时, 压缩率增加并且精度缓慢降低。

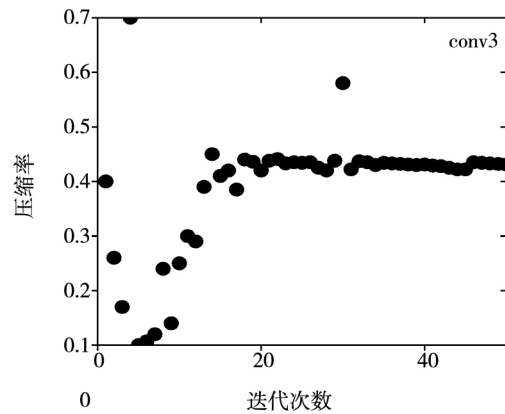
基于与上述实验相同的网络模型 AlexNet, 使用 EKF 方法进行参数训练, 在超参数预测实验中, 遵循 SGD (stochastic gradient descent, 随机梯度下降) 方法的默认实验设置。最大迭代次数设置为 50。在图 5 中, 当迭代次数达到约 40 时, AlexNet 每层的压缩率趋于稳定。因此, 在后续实验中将迭代次数设置为 40。图 4 和图 5 的两个实验证明了贝叶斯方法在卷积神经网络剪枝中的优越性。它可以更快地找到每层的最佳阈值, 既可以保证精度, 又可以最大程度压缩网络。



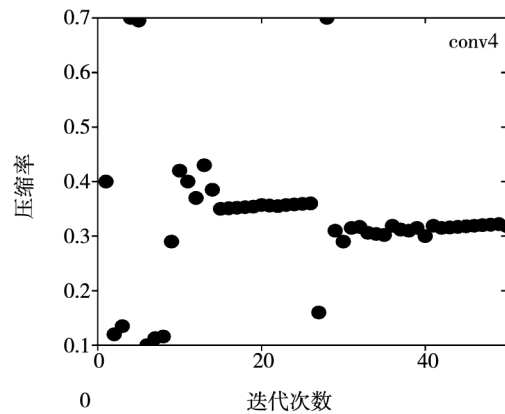
(a) conv1层阈值预测



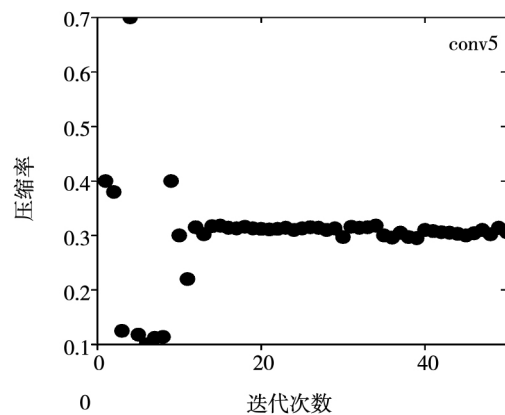
(b) conv2层阈值预测



(c) conv3层阈值预测



(d) conv4层阈值预测



(e) conv5层阈值预测

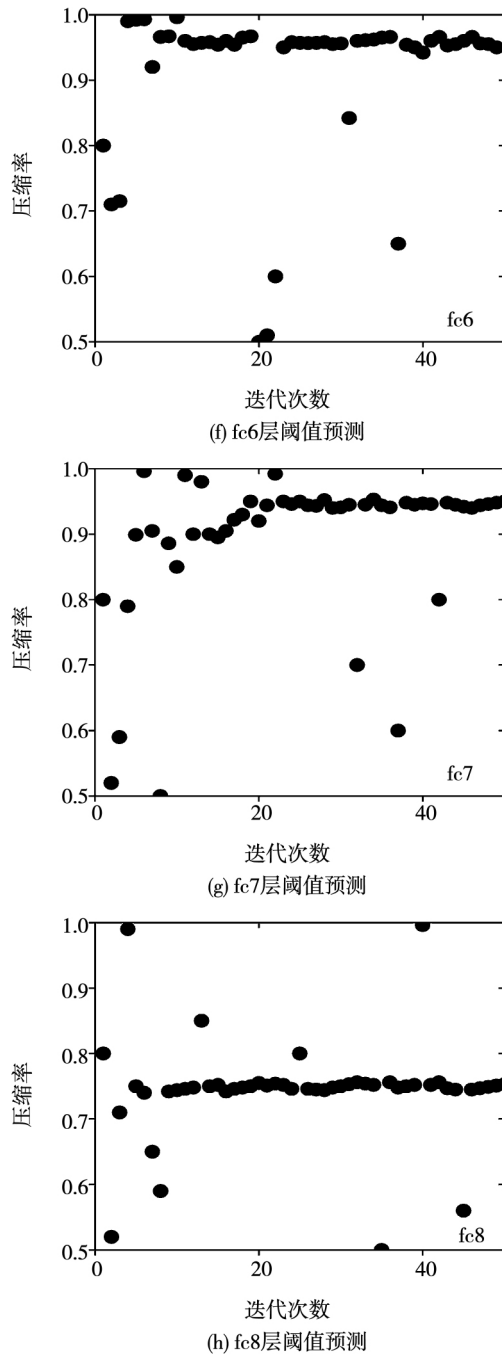


图5 AlexNet 的剪枝阈值预测

Fig. 5 AlexNet's pruning threshold prediction

在整个的实验设计中,首先,采用基于贝叶斯算法的自适应阈值修剪算法,以解决传统的手工设置并进行多次尝试选取最佳值容易造成阈值设置不准确的问题;其次,对权重参数进行聚类量化,使整个网络的参数可以大幅度减少,从而达到高倍压缩的目的;最后,对网络进行微调,使得网络的表达能力维持稳定。

图6显示了融合算法和单独进行剪枝、单独进

行量化时的不同压缩率下的精度损失。当单独工作时,如右侧两条曲线所示,当压缩率低于其原始尺寸的8%时,剪枝网络的准确性开始显著下降;当压缩率低于其原始大小的8%时,量化网络的准确性也开始显著下降。但是当融合剪枝与量化时,如左侧曲线所示,网络可以压缩到原始大小的4%而不会损失准确性。

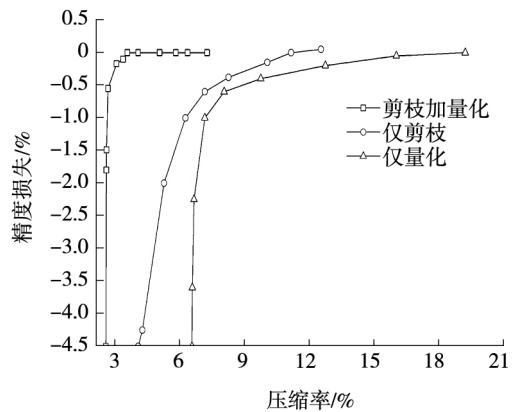


图6 融合剪枝与量化算法对网络性能的影响

Fig. 6 The influence of fusion pruning and quantization algorithm on network performance

由此可知,在基于贝叶斯的优化剪枝算法中融合量化算法可以使网络压缩更具有有效性。AlexNet 网络具体压缩情况见表1,可以看出改进的剪枝过程对于第一层卷积层 conv1 以及最后一层完全连接层 fc8 具有明显的压缩效果,融合了剪枝与量化算法也使得网络得到了较高的压缩率。

表1 AlexNet 网络压缩参数

Tab. 1 AlexNet network compression parameters

网络层	网络参数	剪枝后每层 参数占比/%	压缩率/%
conv1	35K	84	25.8
conv2	305K	38	11.9
conv3	880K	38	11.3
conv4	667K	35	11.1
conv5	441K	39	11.7
Fc6	41M	9	3.4
Fc7	17M	9	2.8
Fc8	4M	24	7.7

3.3 AlexNet 神经网络压缩模型识别植物叶片

本节将在 AlexNet 神经网络压缩模型上进行植物叶片的识别实验,并与在 AlexNet 上的其它压缩方法进行对比实验。如图7神经网络在压缩前后的

准确率所示, AlexNet 神经网络在压缩后没有明显的准确率损失, 并且从迭代次数中可以看出, 压缩后的网络可以很快的达到较高的准确度。表 2 给出了本文方法与其他 4 种主流压缩方法比较结果。可以看出, 本文方法以准确度损失仅仅 0.91% 为代价, 获得了 36 倍的压缩率, 而其他 4 种方法中, 若要获得高倍的压缩率, 将会有较高的精度损失。

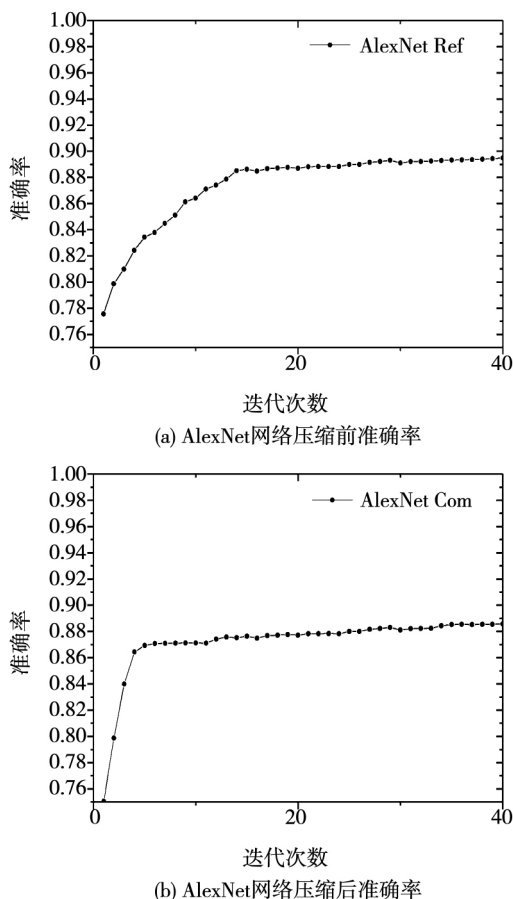


图7 神经网络在压缩前后的准确率

Fig.7 The accuracy of the neural network before and after compression

表2 不同压缩方法对于叶片识别的效果

Tab.2 The effect of different compression methods on leaf recognition

方法	准确度/%	参数大小/MB	压缩率
未压缩	89.49	240	-
本文方法	88.58	6.7	36 ×
Fastfood-32-AD [27]	87.66	131	2 ×
Fastfood-16-AD [27]	87.61	64	3.7 ×
Collins&Kohli [28]	81.31	61	4 ×
SVD [29]	82.97	47.6	5 ×

4 结 论

本文提出了融合剪枝与量化进行植物叶片识别网络的自动压缩方法。通过创建一个集成框架来消除权重冗余并量化表征参数信息的存储元素。该方法采用递归贝叶斯算法预测阈值的超参数, 通过使用 K-Means 聚类方法来共享参数的中心值以达到量化的目的, 并在植物叶片数据集上进行植物叶片识别实验。结果表明, 本文提出的融合方法可以将植物叶片识别网络 AlexNet 压缩 36 倍, 有效的降低了计算复杂度与内存占用量, 同时达到较为满意的识别精度, 提升了植物叶片的识别效率。在未来的研究中, 将考虑叶片图像的形状特征与纹理特征相结合, 以提高识别率; 亦可将此压缩模型部署到移动终端, 为更方便、有效的进行植物识别提供新的工程解决方案。

参 考 文 献:

- [1] 许振雷, 杨瑞, 王鑫春, 应文豪. 基于卷积神经网络的树叶识别的算法的研究 [J]. 电脑知识与技术, 2016, 12(10): 194.
- XU Zhenlei, YANG Rui, WANG Xinchun, et al. Based on Leaves Convolutional Neural Network Recognition Algorithm [J]. Computer Knowledge and Technology, 2016, 12(10): 194.
- [2] 于舒春, 佟小雨. 基于 CNN 特征提取的粒子滤波视频跟踪算法研究 [J]. 哈尔滨理工大学学报, 2020, 25(4): 78.
- YU Shuchun, TONG Xiaoyu. Research on Particle Filter Video Tracking Algorithms Based on CNN Feature Extraction [J]. Journal of Harbin University of Science and Technology, 2020, 25(4): 78.
- [3] JADERBERG M, VEDALDI A, ZISSERMAN A. Speeding Up Convolutional Neural Networks with Low Rank Expansions [J]. arXiv Preprint arXiv: 1405.3866, 2014.
- [4] DENTON E L, ZAREMBA W, BRUNA J, et al. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation [C] // Advances in Neural Information Processing Systems. 2014: 1269.
- [5] LEBEDEV V, GANIN Y, RAKHUBA M, et al. Speeding-up Convolutional Neural Networks Using Fine-tuned cp-decomposition [J]. arXiv Preprint arXiv: 1412.6553, 2014.
- [6] ZHANG X, ZOU J, MING X, et al. Efficient and Accurate Approximations of Nonlinear Convolutional Networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1984.
- [7] ZHANG X, ZOU J, HE K, et al. Accelerating Very Deep Convolutional Networks for Classification and Detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 1943.
- [8] WANG P, CHENG J. Accelerating Convolutional Neural Networks

- for Mobile Applications [C] // Proceedings of the 24th ACM International Conference on Multimedia, 2016: 541.
- [9] LEBEDEV V, LEMPITSKY V. Fast Convnets Using Group-wise Brain Damage [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2554.
- [10] ANWAR S, HWANG K, SUNG W. Structured Pruning of Deep Convolutional Neural Networks [J]. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2017, 13(3): 1.
- [11] NOVIKOV A, PODOPRIKHIN D, OSOKIN A, et al. Tensorizing Neural Networks [C] // Advances in Neural Information Processing Systems, 2015: 442.
- [12] GUPTA S, AGRAWAL A, GOPALAKRISHNAN K, et al. Deep Learning with Limited Numerical Precision [C] // International Conference on Machine Learning, 2015: 1737.
- [13] COURBARIAUX M, BENGIO Y, DAVID J P. Training Deep Neural Networks with Low Precision Multiplications [J]. arXiv Preprint arXiv: 1412.7024, 2014.
- [14] MIYASHITA D, LEE E H, MURMANN B. Convolutional Neural Networks Using Logarithmic Data Representation [J]. arXiv Preprint arXiv: 1603.01025, 2016.
- [15] 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.
- [16] LIN Z, COURBARIAUX M, MEMISEVIC R, et al. Neural Networks with Few Multiplications [J]. arXiv Preprint arXiv: 1510.03009, 2015.
- [17] COURBARIAUX M, HUBARA I, SOUDRY D, et al. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or-1 [J]. arXiv Preprint arXiv: 1602.02830, 2016.
- [18] RASTEGARI M, ORDONEZ V, REDMON J, et al. Xnor-net: Imagenet Classification Using Binary Convolutional Neural Networks [C] // European Conference on Computer Vision. Springer, Cham, 2016: 525.
- [19] GONG Y, LIU L, YANG M, et al. Compressing deep Convolutional Networks Using Vector Quantization [J]. arXiv Preprint arXiv: 1412.6115, 2014.
- [20] CHEN W, WILSON J, TYRES S, et al. Compressing Neural Networks with the Hashing Trick [C] // International Conference on Machine Learning, 2015: 2285.
- [21] 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.
- [22] BA J, CARUANA R. Do Deep Nets Really Need to Be deep? [C] // Advances in Neural Information Processing Systems, 2014: 2654.
- [23] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network [J]. arXiv Preprint arXiv: 1503.02531, 2015.
- [24] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for Thin Deep Nets [J]. arXiv Preprint arXiv: 1412.6550, 2014.
- [25] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet Classification with Deep Convolutional Neural Networks [J]. Communications of the ACM, 2017, 60(6): 84.
- [26] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet Classification with Deep Convolutional Neural Networks [C] // Neural Information Processing Systems, 2012: 1097.
- [27] YANG Z, MOCZULSKI M, DENIL M, et al. Deep Fried Convnets [C] // Proceedings of the IEEE International Conference on Computer Vision, 2015: 1476.
- [28] COLLINS M D, KOHLI P. Memory Bounded Deep Convolutional Networks [J]. arXiv Preprint arXiv: 1412.1442, 2014.
- [29] DENTON E L, ZAREMBA W, BRUNA J, et al. Exploiting Linear Structure within Convolutional Networks for Efficient Evaluation [C] // Advances in Neural Information Processing Systems, 2014: 1269.

(编辑: 王 萍)