JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
01
VOL.26

ISSN1006-8961
CN11-3758/TB

“自动驾驶技术与应用”专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第1期（总第297期）
2021年1月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

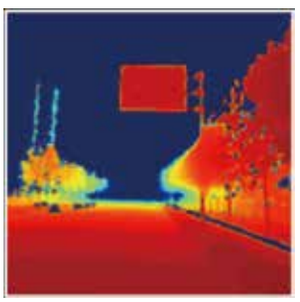
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年1月 第26卷 第1期（总第297期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

目次

面向智能驾驶的平行视觉感知：基本概念、框架与应用(第0067页)



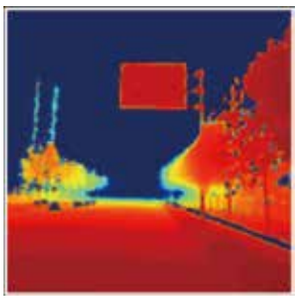
Edge-guided GAN：边界信息引导的深度图像修复(第0186页)

利用边缘计算的多车协同激光雷达SLAM(第0218页)

Volume 26, Number 1
Published January 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Parallel visual perception for intelligent driving: basic concept, framework and application (P0067)



Edge-guided GAN: a depth image inpainting approach guided by edge information (P0186)

Cooperative LiDAR SLAM for multi-vehicles based on edge computing (P0218)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2021)01-0198-10

论文引用格式:Hu J, Huang Q P, Liu J X, Liu W, Yuan H and Zhao H. 2021. Greedy pruning of deep neural networks fused with probability distribution. Journal of Image and Graphics,26(01):0198-0207(胡骏,黄启鹏,刘嘉昕,刘威,袁淮,赵宏. 2021. 引入概率分布的深度神经网络贪婪剪枝. 中国图象图形学报,26(01):0198-0207)[DOI:10.11834/jig.200438]

引入概率分布的深度神经网络贪婪剪枝

胡骏[1,2],黄启鹏[1],刘嘉昕[1,2],刘威[1,3],袁淮[1],赵宏[1]

1. 东北大学计算机科学与工程学院, 沈阳　110169; 2. 东软睿驰汽车技术(沈阳)有限公司, 沈阳　110179;
3. 东软睿驰汽车技术(上海)有限公司,上海　201804

摘　要:**目的**　深度学习在自动驾驶环境感知中的应用,将极大提升感知系统的精度和可靠性,但是现有的深度学习神经网络模型因其计算量和存储资源的需求难以部署在计算资源有限的自动驾驶嵌入式平台上。因此为解决应用深度神经网络所需的庞大计算量与嵌入式平台有限的计算能力之间的矛盾,提出了一种基于权重的概率分布的贪婪网络剪枝方法,旨在减少网络模型中的冗余连接,提高模型的计算效率。**方法**　引入权重的概率分布,在训练过程中记录权重参数中较小值出现的概率。在剪枝阶段,依据训练过程中统计的权重概率分布进行增量剪枝和网络修复,改善了目前仅以权重大小为依据的剪枝策略。**结果**　经实验验证,在 Cifar10 数据集上,在各个剪枝率下本文方法相比动态网络剪枝策略的准确率更高。在 ImageNet 数据集上,此方法在较小精度损失的情况下,有效地将 AlexNet、VGG(visual geometry group)16 的参数数量分别压缩了 5.9 倍和 11.4 倍,且所需的训练迭代次数相对于动态网络剪枝策略更少。另外对于残差类型网络 ResNet34 和 ResNet50 也可以进行有效的压缩,其中对于 ResNet50 网络,在精度损失增加较小的情况下,相比目前最优的方法 HRank 实现了更大的压缩率(2.1 倍)。**结论**　基于概率分布的贪婪剪枝策略解决了深度神经网络剪枝的不确定性问题,进一步提高了模型压缩后网络的稳定性,在实现压缩网络模型参数数量的同时保证了模型的准确率。

关键词:深度学习;神经网络;模型压缩;概率分布;网络剪枝

Greedy pruning of deep neural networks fused with probability distribution

Hu Jun[1,2], Huang Qipeng[1], Liu Jiaxin[1,2], Liu Wei[1,3], Yuan Huai[1], Zhao Hong[1]

1. School of Computer Science and Engineering, Northeastern University, Shenyang　110169, China;
2. Neusoft Reachauto Corporation (Shenyang), Shenyang　110179, China;
3. Neusoft Reachauto Corporation (Shanghai), Shanghai　201804, China

Abstract: **Objective** In recent years, deep learning neural network has continued to develop, and excellent results have been achieved in the fields of computer vision, natural language processing, and speech recognition. In autonomous driving technology, the environment perception is an important application. The environment perception mainly processes the collected image information about the surrounding environment. Thus, deep learning is an important section in this link. However, the number of layers of existing neural network models continues to increase with the continuous increase in the com-

收稿日期:2020-07-31;**修回日期**:2020-10-20;**预印本日期**:2020-10-27

基金项目:上海市经信委工业强基项目(GYQJ-2017-1-04);辽宁省科技创新重大专项(2019JHI/10100026)

Supported by: Shanghai Committee of Economic and Information Technology Industrial Base Project(GYQJ-2017-1-04); Liaoning Province Scientific and Technological Innovation Project(2019JHI/10100026)

plexity of processing problems. Thus, the number of overall parameters of the network and the required computing power are increasing. These models run well on platforms with sufficient computing power, such as server platforms with sufficient computing power. However, many deep neural network models are difficult to be deployed on embedded platforms with limited computing and storage resources, such as autonomous driving platforms. Compressing the existing deep neural network models is necessary to solve the contradiction between the huge amount of calculation required for the application of deep neural networks and the limited computing power of embedded platforms. This process can reduce the number of model parameters and computing power. This paper proposes a greedy network pruning method based on the existing model compression method. The propose method incorporates the probability distribution of weights to reduce redundant connections in the network model and improve the computational efficiency and parameters of the model. **Method** The current pruning method mainly uses the property of weight parameter as a criterion for parameter importance evaluation. The 1 norm of the convolution kernel weight parameter is used as the basis for determining the importance. However, this method ignores the variation of weight during training. In the pruning process, many methods use the trained model to perform one-time pruning. Thus, the accuracy of the model after pruning is difficult to maintain. the proposed is inspired by the study of uncertain graphs to solve the above problem. The probability distribution of weights is introduced, and the importance of the connection is jointly judged in accordance with the probability distribution of the weight parameter value and the size of the current weight in the training. The importance of the network connection and the effect of cutting the connection on the loss function are jointly used. The degree of influence collectively represents the contribution rate of this network connection to the result, thereby serving as the basis for pruning the network connection. In the stage of greedy pruning of the model, the proposed method uses incremental pruning to control the scale and speed of pruning. Iterative pruning and restoration are performed for a small proportion of connections until the state of the current sparse connections no longer changes. The pruning scale is gradually expanded until the expected model compression effect is achieved. Therefore, the incremental pruning and recovery strategy can avoid the weight gradient explosion problem caused by excessive pruning, improve the pruning efficiency and model stability, and realize dynamic pruning compared with the one-time pruning process based on the weight parameters. The proposed pruning method guarantees the maximum compression of the model's volume while maintaining its accuracy. **Result** The experiment uses networks of different depths for experiments, including CifarSmall, AlexNet, and visual geometry group (VGG)16, and nets with residual connections, including ResNet34 and ResNet50 networks, to verify the applicability of the proposed method to different depth networks. The experimental dataset uses the commonly used classification datasets, including CIFAR-10 and ImageNet ILSVRC (ImageNet Large Scale Visual Recognition Challenge)-2012, making it convenient for comparison with other methods. The main comparison content of the experiment includes the proposed method and the dynamic network pruning strategy on CIFAR-10. The pruning effect of the proposed method and the current state-of-the-art (SOTA) pruning algorithm HRank is compared on the Imagenet dataset in ResNet50. Experimental results prove that the accuracy of the proposed method is higher than that of the dynamic network pruning strategy at various pruning rates on the Cifar10 dataset. On the ImageNet data set, the proposed method effectively compresses the number of parameters of AlexNet and VGG16 by 5.9 and 11.4 times, respectively, with a small loss of accuracy. The number of training iterations required is more than that of the dynamic network pruning strategy. Effective compression can be performed for the residual type networks ResNet34 and ResNet50. For the ResNet50 network, a larger compression rate is achieved with a small increase in accuracy loss compared with the current SOTA method HRank. **Conclusion** The greedy pruning strategy fused with probability distribution solves the uncertainty problem of deep neural network pruning, improves the stability of the network after model compression, and realizes the compression of the number of network model parameters while ensuring the accuracy of the model. Experimental results prove that the proposed method has a good compression effect for many types. The probability distribution of the weight parameters introduced in this research can be used as an important basis for the subsequent parameter importance criterion in the pruning research. The incremental pruning and the connection recovery in the pruning process used in this article are important for accuracy maintenance, However, optimizing and accelerating the reasoning of the sparse model obtained after pruning needs further research.

Key words: deep learning; neural network; model compression; probability distribution; network pruning

0 引 言

受生物的大脑皮层启发并发展而来的深度学习在自然语言处理、语音识别、图像处理和计算机视觉等研究领域中取得了显著的效果,从而使得深度学习在众多研究领域中脱颖而出。从最初浅层神经网络到如今的深度神经网络(deep neural network, DNN),从早期的手写字体识别到当前的图像分类、物体定位和检测、自然语言处理、医学影像检测和无人驾驶汽车等领域,深度神经网络已经取得了传统机器学习方法无法达到的高精确度和高识别率。

深度神经网络取得了巨大的成功,深度网络模型的性能和识别率也不断提高,但是也造成了网络规模不断扩大、网络参数和模型复杂度不断增加等问题。例如,Krizhevsky 等人(2012)提出了经典的 AlexNet 网络模型,该模型在 2012 年的大规模视觉识别挑战赛(ImageNet Large Scale Visual Recognition Challenge, ILSVRC)中的图像分类竞赛取得冠军,但其参数量和乘加运算量分别是 6 100 万和 72 000 万。而随后 Simonyan 和 Zisserman(2015)提出的 VGGNet(visual geometry group)和 Szegedy 等人(2015)提出的 GoogLeNet 等复杂网络模型的准确率不断提高的同时,模型规模和深度也愈加庞大,如表 1 所示。对深度神经网络模型而言,庞大的参数和复杂的网络结构具有更强大的特征学习和表达能力,同时也意味着更多的存储空间和计算量。因此使得深度神经网络的相关成果不能大规模应用在计算能力有限的嵌入式和移动设备上。

表 1 常用网络模型参数统计表

Table 1 Common network model parameter statistics table

模型	前 5 准确率/%	参数量/M	乘加计算量/M
AlexNet	83.6	60	720
VGG-16	93.7	138	15 300
GoogLeNet	93.3	68	1 550
Inception-V3	96.4	23.2	5 000

虽然深度神经网络需要依靠其庞大的网络参数来保证模型的高准确率,Cheng 等人(2017)研究表明现有的网络模型往往过度参数化,存在显著的参数冗余,Denil 等人(2013)采用合理的策略,在不明显影响模型的准确率的前提下对模型进行适当的压缩,从而有效地平衡模型计算量和准确度,为深度神经网络部署在嵌入式等设备上提供可能。

在现有的众多模型压缩方法中,网络剪枝策略不仅可以将原始密集型深度神经网络转化成稀疏型网络,从而得到显著的压缩效果,而且可以很好地保证模型预测准确率。例如,Han 等人(2015)提出了通过删除“不重要”的参数并重新训练和微调剩余的参数来进行“无损”DNN 压缩的方法。然而,由于互联神经元之间的相互影响和相互激活,参数重要性可能会在剪枝过程中发生巨大变化。因此,该方法存在两个问题。一方面,它不可避免地会造成无法挽回的网络损伤。由于修剪后的连接无法修复,不正确的修剪可能会导致严重的精度损失。另一方面,它的学习效率非常低,因为要得到预期的网络模型需要不断交替地进行网络剪枝和重训练,这是非常耗时的。随后 Guo 等人(2016)提出了一种剪枝与拼接操作结合的动态网络剪枝策略。它通过剪枝操作减去当前“不重要”的连接,同时利用拼接操作恢复不正确的修剪,从而可以进行实时剪枝,使用很少的迭代次数就可以达到预期的压缩效果。有效地解决了上述剪枝策略带来的不可恢复的网络损害和学习效率低下的问题。但是,该方法依然存在两个问题:1)此方法同样通过设置阈值的方式来衡量该权重的重要性,并以逐渐递减的概率对满足要求的权重进行实时剪枝。由于模型压缩本身存在经验性和不确定性,仅仅利用预先训练好的模型的权重大小并不能有效地衡量网络连接的重要程度。2)利用固定的阈值进行实时剪枝的方法,很容易造成反向传播之后梯度爆炸的问题。由于该方法在实现过程中无法实时控制剪枝数量,它在迭代过程中仅仅通过权重的不断更新进行剪枝或是拼接,这将无法保证网络剪枝的稳定性。本文试图改善现有的剪枝策略,引入基于概率分布的网络连接的重要程度衡量方法以及增量剪枝方法。具体地,采用动态网络手术的实时剪枝方法,并基于网络连接中小权重的概率分布和贡献率等先验知识进行增量贪婪剪枝,其剪枝步骤如图 1 所示。首先,在模型预训练阶段统计所有可学习层的每个连接上小权重出现的概率(图 1(a))。接着,通过每层权重的均值和方差来定义权重的剪枝边界,结合剪枝边界和权重的概率

分布来衡量该连接对损失函数的贡献率。然后，通过动态网络手术的实时修剪和拼接操作进行增量贪婪剪枝(图 1(b))，既能剪掉网络中贡献率较小的连接(图 1(b)虚线)，还能对不正确的剪枝操作进行实时修复(图 1(b)绿色实线)。先对小规模的网络连接进行实时剪枝，待其稳定之后再逐步扩大剪枝规模，这在一定程度上保证了模型的收敛效果和剪枝效率，最终得到预期的稀疏型网络模型(图 1(c))。因此，基于概率分布和贡献率的权重重要性度量方法和动态剪枝策略的结合来实现实时贪婪剪枝是一种行之有效的剪枝策略。

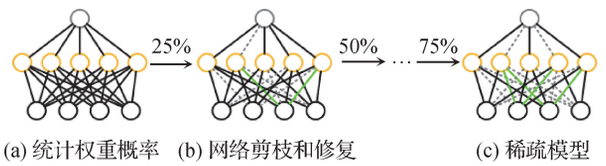


图 1　贪婪剪枝策略示意图

Fig. 1　Schematic diagram of greedy pruning strategy

((a) statistics of weight probability; (b) network pruning and repair; (c) sparse model)

1　相关工作

为了减少模型参数冗余，提高计算和存储资源的利用率，如今已有各种研究提议进行模型压缩和加速。Mathieu 等人(2013)探索图形处理器(graphics processing unit，GPU)上的快速傅里叶变换(fast Fourier transform，FFT)，并通过在频域中执行卷积计算来提高卷积神经网络(convolutional neural networks，CNN)的速度。另一种模型压缩方法是采用矩阵(或张量)分解。Denil 等人(2013)通过用适当构造的低秩分解来近似参数矩阵。此方法在卷积层上达到了 1.6 倍的加速效果，并将精度保持在原始模型的 1% 以内。虽然矩阵(或张量)分解可能有利于深度网络的压缩和加速，但这些方法通常会在模型的高度压缩需求下导致严重的精度损失。针对卷积核进行剪枝的方法也是模型压缩的一种常见方式，Yu 等人(2018)提出了基于特征图重构的方法，关注的是在网络最后一层之前所有层的输出特征图对结果重要性的贡献，这种方式的主要问题是在进行剪枝通道选择时，需要反复的实验来进行参数调整，所需要的时间较多。He 等人(2019)提出一种基于几何中位数的卷积核修剪方法，该指标旨在找出各层中卷积核之间的共同信息，然后将冗余可替代的卷积核修剪掉。Lin 等人(2019)提出了基于生成对抗思路的剪枝方式，将剪枝之后模型设置为生成器，预训练模型输出设置为 baseline，同时使得生成器输出分别与预训练模型和真值对齐。Ding 等人(2019)提出了一种新颖的优化器方法，目的在于在训练时使得一部分卷积核参数接近，使其折叠成参数超空间中的单个点，然后在剪枝时移除接近的卷积核。Lin 等人(2020)提出脱离卷积核本身的性质来分析其重要性，根据特征图矩阵的秩的大小作为该特征图包含信息量大小的衡量标准，然后确定不重要的特征图通道，之后再减去产生该通道的对应的卷积核。Meng 等人(2020)提出与传统的剪枝方式不同的方法，不是去除无效的卷积核，而是通过卷积核嫁接的方式将无效的卷积核重新进行激活。通道剪枝在模型压缩率上表现较好，但是精度损失通常比较严重。此外，矢量量化也是一种可行的模型压缩方式。Zhou 等人(2017)使用渐进式网络量化的方法来迭代训练网络直至所有权重均被量化为低精度权重，本文借鉴了此方法的增量压缩思想。Courbariaux 等人(2016)提出的二值量化神经网络以及 Rastegari 等人(2016)提出的 XNOR-Net 能够得到显著的压缩效果，但不可避免地会带来明显的精度损失。

本文采用网络修剪的压缩策略，其最受人关注的是 Han 等人(2015)结合剪枝和重训练操作进行网络修剪的方法，在图像数据集 ImageNet 上得到了 9 倍的压缩结果并且没有精度损失，同时还验证了稀疏矩阵向量乘法可以通过某些硬件设计进一步加速。但是 Han 等人(2015)的方法具有无法挽回的网络损害和学习效率低下的缺点。

由于优化深度神经网络的过程存在经验性和不确定性，现有剪枝方法仅仅根据权值大小来判断是否对网络连接进行剪枝操作。例如 Han 等人(2015)通过设置阈值的方式减掉权值较小的连接，而 Guo 等人(2016)采取了融合剪枝和拼接的实时动态剪枝方法来修复剪枝策略带来的精度损失，并且提高了网络模型的学习效率。该方法在 ImageNet 数据集上得到了 17.7 倍的压缩结果并且没有精度损失。该方法也是依据模型预先训练好的权重大小衡量参数重要性进行剪枝。然而，贪婪剪枝方法预

期得到的是基于原始网络模型压缩而来的稀疏型网络模型,最终的目标为最优的网络连接,而非参数。预先训练好的权重大小不应成为衡量连接重要性的唯一标准。因此本文试图找到更好的网络连接的重要性评估方法。

本文在探索连接重要性的衡量标准中受到了不确定图研究的启发。由张伟(2017)的研究可知,一个不确定图可用四元组 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E},\boldsymbol{P},\boldsymbol{W})$ 表示,其中 $\boldsymbol{V}$ 和 $\boldsymbol{E}$ 分别表示 $\boldsymbol{G}$ 含有的节点集和边集,$\boldsymbol{P}$ 表示边上的概率集合,即 $\boldsymbol{P}(e)$ 表示边 $e \in \boldsymbol{E}$ 的存在概率,$\boldsymbol{W}$ 表示边上的权值集合,即 $\boldsymbol{W}(e)$ 表示边 $e \in \boldsymbol{E}$ 的权值。在利用不确定图求解最短路径的问题中,最短路径的求解不仅要考虑权值 w(节点间的连接的权值或距离长度),节点连接的概率 p 也是影响求解两个不同节点间的最短路径的一大因素,例如高原(2013)在利用不确定图理论求解任意两个节点间的最短路径问题中针对存在多种路径长度相等的情况,利用概率集合 $\boldsymbol{P}$ 来判断哪条最短路径才是最优的。另外,不确定图在解决道路交通网等实际问题中,权值 W 表示实际的距离长度,而概率 p 表示道路堵车的概率。由此看来,概率 p 在求解出发点到目的地的最短路径时变得尤为重要。

因此,本文尝试在求解密集型卷积神经网络的最优剪枝问题中考虑预训练阶段所有可学习层的网络连接上小权重出现的概率 p,并结合原始网络模型在预训练阶段的权重的概率分布和当前的权重大小对网络连接进行贡献率评估,然后采用依据贡献率的增量贪婪剪枝策略,得到更稳定高效的稀疏型深度神经网络。

2 基于概率分布的贪婪剪枝策略

2.1 符号表示

$\{\boldsymbol{W}_k: 0 \leqslant k \leqslant C\}$ 表示原始深度神经网络模型参数,其中 $\boldsymbol{W}_k$ 表示第 k 层的权重矩阵,C 是网络的总层数。另外,$\{\boldsymbol{P}_k: 0 \leqslant k \leqslant C\}$ 表示第 k 层各网络连接上小权重出现的概率矩阵。为了便于表示模型压缩的剪枝和恢复操作,使用 $\{\boldsymbol{W}_k, \boldsymbol{M}_k: 0 \leqslant k \leqslant C\}$ 表示剪枝过程中的稀疏矩阵,其中 $\boldsymbol{M}_k$ 是一个表示第 k 层的网络连接状态的二值矩阵,取值 0 和 1 分别表示连接被剪枝、连接被保留或是被恢复。

2.2 参数的概率分布

为了保留原始模型中的重要连接,并减掉不重要的连接,首要工作是要判断哪些是重要连接,哪些是不重要连接。因此首先在模型训练阶段统计权重的概率分布,具体地,统计所有可学习层的固定连接上数值较小的权重出现的概率。如果概率值较大,意味着该连接上小权重出现的频率较高,它更有可能被剪枝。然后将网络连接上小权重出现的概率分布映射为网络连接剪枝发生的概率分布。

以第 k 层为例,期望得到以下权重的概率

$$p_k^{(i,j)} = \frac{\sum_1^N I(|W_k^{(i,j)}| < \eta \times Z_k)}{N}, \eta \in (0,1) \tag{1}$$

式中,$p_k^{(i,j)}$ 表示第 k 层中权重 $W_k^{(i,j)}$ 所在的网络连接上小权重出现的概率,N 表示统计次数,η 代表概率统计时的系数,而 Z_k 表示关于第 k 层权重的均值和方差的函数,具体表示为

$$Z_k = \max(mu + a \times std, 0) \tag{2}$$

式中,mu 和 std 分别表示第 k 层权重的均值和方差,a 是均值和方差相加时的比例系数。

2.3 贪婪剪枝策略

因为本文的优化目标是对训练好的模型进行剪枝,因此通过网络剪枝实现的稀疏型神经网络是由训练好的原始模型得到的。基于上述训练过程中统计的概率分布,执行贪婪剪枝策略。其中贪婪剪枝策略主要分为 3 个方面(连接贡献率度量、增量剪枝、连接修复)。首先,本文采用基于贡献率的权重重要性度量方法代替现有的基于权重大小的重要性度量方法。具体地,先通过训练阶段统计的权重的概率分布和当前权重的大小来判断该连接的重要程度,再利用网络连接的重要程度配合剪掉该连接对损失函数的影响程度来表示此网络连接的贡献率,以此来定义网络连接的剪枝依据。然后,通过增量剪枝来控制剪枝的规模和速度,首先针对小比例的连接进行迭代剪枝和恢复,待当前稀疏连接的状态不再变化之后,逐步扩大剪枝规模直到达到预期的模型压缩效果。因此,增量剪枝和恢复策略不仅可以避免过度剪枝带来的权重梯度爆炸问题,还能提高剪枝效率和模型稳定性。

以第 k 层为例,原始的优化函数变更为优化以下损失函数

$$\min_{\boldsymbol{W}_k,\boldsymbol{T}_k} L(\boldsymbol{W}_k \odot \boldsymbol{M}_k) \tag{3}$$

$$\text{s. t. } M_k^{(i,j)} = h_k(w_k^{(i,j)}, p_k^{(i,j)}), \forall (i,j) \in \boldsymbol{I} \tag{4}$$

式中，$L(\cdot)$ 为模型损失函数，$\odot$ 表示哈达玛(Hadamard)积乘积运算，$\boldsymbol{I}$ 是由该层所有的权重集合，$h_k(\cdot)$ 是一个用于表示连接重要性的判别函数。

权重更新方案为

$$W_k^{(i,j)} \leftarrow T_k^{(i,j)} - \alpha \frac{\partial L(\boldsymbol{W}_k \odot \boldsymbol{M}_k)}{\partial (T_k^{(i,j)})}, \forall (i,j) \in \boldsymbol{I} \tag{5}$$

$$T_k^{(i,j)} = W_k^{(i,j)} \times M_k^{(i,j)} \tag{6}$$

上述基于概率分布的贪婪剪枝策略可用算法语言形式描述如下：

输入：训练集 $\boldsymbol{X}$，基础学习率 α，总训练次数 i_m，初始训练阶段 i_i，权重概率统计阶段 i_c，贪婪剪枝阶段 i_p，继续训练阶段 i_r。

输出：$\{\boldsymbol{W}_k, \boldsymbol{M}_k : 0 \leqslant k \leqslant C\}$：更新的参数矩阵 $\boldsymbol{W}_k$ 和二值掩码矩阵 $\boldsymbol{M}_k$。

初始化：随机初始化 $\boldsymbol{W}_k$ 为服从均值为 0、方差为 1 的高斯分布，$\boldsymbol{M}_k$ 初始化为值全为 1 的矩阵，i_i 初始化为 0，设置权重统计次数 n_c 和剪枝次数 n_p。

/* 当前迭代次数小于最大迭代次数时，执行此循环 */

While ($i_i < i_m$) do

begin

1) 从训练集 $\boldsymbol{X}$ 中选择最小批处理的数据作为网络输入

2) 前向传播 $(\boldsymbol{W}_0 \odot \boldsymbol{M}_0), \cdots, (\boldsymbol{W}_C \odot \boldsymbol{M}_C)$ 并计算当前损失 L：

```
if ( i_i == i_c )
  /* 统计较小权重值出现的概率 */
  count_probability(weights);
else if ( i_i == i_p )
  /* 依据贡献率度量函数 h_k(·) 和当前的 W_k 更新 M_k */
  update_mask(weights, &mask);
/* 合并 weights 和 mask */
for (int i = 0; i < num_weights; i++)
  merge[i] = weights[i] * mask[i];
/* 将 merge 前向传递 */
Forward(merge);
```

3) 依据模型输出和损失函数梯度 ∇L 进行反向传播

```
Backward(merge);
```

4) 权重更新：依据式(2)和当前的损失函数梯度 ∇L 更新 $\boldsymbol{W}_k$

```
i_i := i_i + 1  /* i 递增 */
end; /* end of while */
```

由于模型压缩过程中的剪枝和修复策略是基于网络连接的重要程度进行的，所以判别函数 $h_k(\cdot)$ 尤为重要。具体衡量网络连接重要程度的计算式为

$$h_k(w_k^{(i,j)}, p_k^{(i,j)}) = \begin{cases} 0 & I_k^{(i,j)} < a_k \\ M_k^{(i,j)} & \text{其他} \\ 1 & I_k^{(i,j)} \geqslant b_k \end{cases} \tag{7}$$

$$I_k^{(i,j)} = (1 - p_k^{(i,j)}) \times |w_k^{(i,j)}| \tag{8}$$

式中，$p_k^{(i,j)}$ 是在训练过程中统计小权重出现的先验概率，$w_k^{(i,j)}$ 为当前状态的权重。而阈值 a_k 和 b_k 是基于先验知识给出的。首先，如果在该连接上小权重出现的概率较大并且当前权重较小，则表示该连接暂时不重要，可对其执行剪枝操作。如果剪枝该连接对损失函数的影响较大即更新的梯度值 ∇L 较大，这意味着该连接对损失函数的贡献率较大，则对其执行恢复操作。如果均不满足以上条件，则该连接保持原状即不剪枝也不修复。如此迭代执行剪枝和修复策略直至得到预期的稀疏型深度神经网络。

3 实 验

本节将对贪婪剪枝策略实验评估的方法和结果进行说明，并将其应用于主流的网络模型。实验基于深度学习框架 DarkNet 实现。其中所有的模型文件全部由此框架提供，可以直接进行训练。此外，所有的可学习参数均为默认设置，包括训练批量、基础学习率、激活函数以及损失函数选择，只需要修改其最大迭代次数以及预先训练次数、剪枝次数、再训练次数以及卷积层函数和参数，以便于压缩模型得到稀疏型网络。

首先，实验针对包含 10 种常见物体的数据集 CIFAR-10 进行剪枝测试。CIFAR-10 数据集由 10 类尺寸为 32 × 32 像素的彩色图像组成，共包含 60 000 幅图像，每一类包含 6 000 幅图像，其中 50 000 幅图像作为训练集，10 000 幅图像作为验证集。本文使用小型网络 CifarSmall 作为原始参考模型，该网络包含 7 层卷积层和 2 层池化层，卷积层的

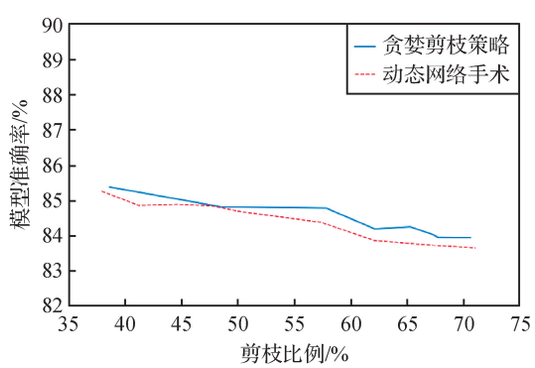
参数总量为 58 K。实验中将 128 幅图像作为最小批量输入。总迭代训练次数为 50 000 次,其中预训练次数为 20 000 次(后 15 000 次进行概率统计),迭代剪枝次数为 10 000 次,再训练次数为 10 000 次。表 2 列出了剪枝率为 70% 的剪枝后结果。

表 2　针对 CifarSmall 网络压缩结果

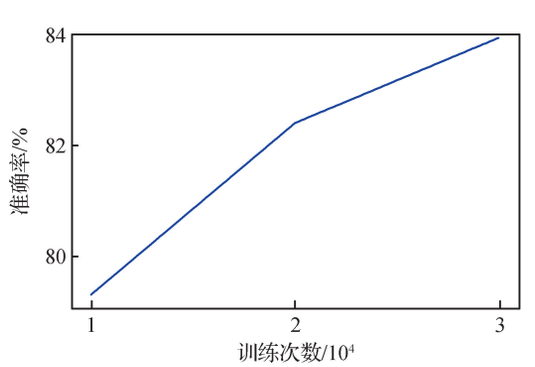
Table 2　CifarSmall experimental results

模型	前 1 错误率/%	前 5 错误率/%	参数量/K	压缩率/倍
CifarSmall Ref	14.22	0.67	58	1
CifarSmall Pruned	15.99	1.33	17.4	3.3

最终得到的压缩效果和模型准确率如图 2 所示,由实验数据看,贪婪剪枝策略相比 Guo 等人(2016)动态网络手术(dynamic network surgery, DNS)方法的压缩率和准确率均有提升。



(a) 模型准确率与剪枝比例关系



(b) 原始模型准确率与训练次数关系

图 2　贪婪剪枝策略与动态网络手术方法(DNS)在 CifarSmall 网络上的实验结果对比

Fig. 2　Comparison results of greedy pruning strategy and dynamic network surgery method (DNS) on the CifarSmall network((a) relationship between model accurecy and pruning ratio; (b) relationship between accuracy of the original model and the number of training iterations)

本文进一步研究了针对 ImageNet ILSVRC-2012 数据集的剪枝性能,其中训练集和验证集的样本数目分别为 1.2 M 和 50 K。使用基于 DarkNet 框架的 AlexNet、VGG16,以及残差网络(ResNet),本文采用 ResNet34 和 ResNet50 两种残差网络模型作为参考对比模型。

首先分别验证本文贪婪剪枝策略在 AlexNet 和 VGG-16 两个经典模型上的压缩效果。如表 3 所示,其中压缩率代表的是原有模型参数量与压缩后模型参数量的比值,本文方法将 AlexNet 模型参数数量压缩了 5.9 倍。同时在 TOP-1 准确率上只下降了 4 个百分点。针对 VGG-16,本文方法将其参数数量压缩了 11.4 倍,同时 TOP1 准确率只有两个百分点的下降。另外本实验的剪枝和重训练需要的迭代次数相对于动态网络手术方法也更少,本文采用在 ImageNet 数据集上预训练好的 AlexNet 模型进行剪枝实验,剪枝以及重训练需要 600 K 次迭代,而动态网络手术方法需要 700 K 次。本文方法在收敛速度上具有优势。

表 3　针对 AlexNet 和 VGG-16 网络压缩结果

Table 3　AlexNet and VGG-16 experimental results

模型	前 1 错误率/%	前 5 错误率/%	参数量/M	迭代次数/K	压缩率/倍
AlexNet Ref	42.78	19.70	61	450	1
AlexNet Pruned	46.78	22.36	10.34	600	5.9
VGG-16 Ref	31.50	11.32	138	224	1
VGG-16 Pruned	33.5	13.22	12.16	500	11.3

针对经典的 AlexNet 结构,表 4 列出了本文方法相对于 Han 等人(2015)的剪枝方法和 Guo 等人(2016)提出的动态剪枝方法在 AlexNet 上每一个网络层在剪枝之后参数保留的比例。可以看出本文方法针对卷积层和全连接层的参数剪枝效果更加均衡。

针对具有 Residual 连接的 ResNet 网络,本文方法剪枝结果如表 5 所示,本文采用了常用的两种特征提取网络 ResNet34 和 ResNet50,由于网络本身包含的冗余参数较少,所以剪枝难度更大,但是仍然在 ResNet50 模型上实现了 2.1 倍的模型压缩率,同时 TOP-5 准确率只下降了一个百分点左右。同样,在更紧凑的网络ResNet34上的压缩效果也较好。显示

表 4　针对 AlexNet 网络不同层剪枝结果表

Table 4　Pruning results for different layers of the AlexNet network

层	参数量	参数保留比例/%		
		Han 等人(2015)	Guo 等人(2016)	本文
conv1	35 K	84	53.8	100
conv2	307 K	38	40.6	22.7
conv3	885 K	35	29.0	30.9
conv4	664 K	37	32.3	38.7
conv5	443 K	37	32.5	38.0
fc1	38 M	9	3.7	34.7
fc2	17 M	9	6.6	16.6
fc3	4 M	25	4.6	15.1
总计	61 M	11	5.7	11.3

表 5　针对 ResNet 压缩结果

Table 5　ResNet experimental results

模型	前 1 错误率/%	前 5 错误率/%	参数量/M	压缩率/倍
ResNet34 Ref	27.6	8.9	21.6	1
ResNet34 Pruned	28.73	9.93	15.42	1.4
ResNet50 Ref	24.2	7.1	25.56	1
ResNet50 Pruned	26.5	8.2	12.38	2.1

了本文方法跨模型的稳定性。

本文与同类方法分别基于 CifarSmall、AlexNet 和 ResNet50 模型进行了横向对比。表 6 列出了不同方法针对 AlexNet 网络在剪枝之后的模型压缩率和精度保持情况，可以看出本文方法在精度损失较低的情况下，实现的模型压缩率相对于 Yang 等人(2014)方法和 Han 等人(2015)的 Native Cut 方法都要更大，虽然与目前表现最好的动态网络手术方法有些差距，但是由表 7 可以看出在 CifarSmall 网络上本文方法在相同压缩率的情况之下表现更好，且如前文所述，本文方法具有更快的收敛特性。

表 8 中列出了本文方法针对 ResNet50 的剪枝结果与目前针对 ResNet50 进行剪枝的最优方法(state of the art，SOTA) HRank 的对比情况，由实验结果可以看出，本文方法在精度损失增加不到两个百分点的情况下，实现了更大的压缩率。

表 6　针对 AlexNet 网络不同方法压缩结果对比

Table 6　Comparison table of compression results for different methods of AlexNet

网络	前 1 错误率/%	前 5 错误率/%	参数量/M	压缩率
Baseline (Krizhevsky 等,2012)	42.78	19.73	61.0	1
Fastfood32(AD) (Yang 等,2014)	**41.93**	–	32.8	2
Fastfood 16(AD) (Yang 等,2014)	42.90	–	16.4	3.7
Naive Cut (Han 等,2015)	57.18	23.23	13.8	4.4
Han 等人(2015)	42.77	**19.67**	6.7	9
DNS(Guo 等,2016)	43.09	19.99	**3.45**	**17.7**
本文	46.78	22.35	10.34	5.9

注：加粗字体表示每列最优结果，“–”表示原论文作者未给出数据。

表 7　针对 CifarSmall 网络不同方法压缩结果对比

Table 7　Comparison table of compression results for different methods of CifarSmall network

模型	前 1 错误率/%	前 5 错误率/%	参数量/K	压缩率
Baseline	15.78	0.67	58	1
DNS(Guo 等人,2016)	16.23	2.13	**17.4**	**3.3**
本文	**16.19**	**1.33**	**17.4**	**3.3**

注：加粗字体表示每列最优结果。

表 8　针对 ResNet50 网络不同方法压缩结果对比

Table 8　Comparison table of compression results of different methods for ResNet50 network

模型	前 1 错误率/%	前 5 错误率/%	参数量/M	压缩率
ResNet50 Ref	14.2	7.1	25.56	1
HRank (Lin 等,2020)	**25.02**	**7.67**	16.15	1.58
本文	27.5	9.2	**12.38**	**2.1**

注：加粗字体表示每列最优结果。

4　结　论

本文提出了一种基于概率分布和贡献率的增量剪枝策略。与先前的剪枝方法不同，该方法考虑了

原始模型的权重的概率分布,降低了剪枝对损失函数的影响,从而提高了剪枝的精度和效率。实验结果表明,本文方法在更少的迭代次数且精度损失较小的情况下,取得了较好的网络压缩效果。然而,模型压缩这一研究方向仍然有许多工作需要进一步研究:

1)对于目前过于依赖经验性的模型剪枝策略,探索更优的网络连接重要性的评估方法是一个亟待解决的问题。

2)对于模型压缩得到的稀疏型深度神经网络模型,如何进一步优化和加速推理也是需要研究者们继续探讨的问题。

虽然模型压缩已经得到了很显著的压缩效果,但是应用于计算资源有限的嵌入式设备还需要进一步解决模型优化和加速等问题,其在未来一段时间内仍将是研究的一大热点。

致　谢　感谢东软睿驰汽车技术有限公司对本论文实验所使用的设备和数据的支持。

参考文献(References)

Cheng Y, Wang D, Zhou P and Zhang T. 2017. A survey of model compression and acceleration for deep neural networks [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1710.09282.pdf

Courbariaux M, Hubara I, Soudry D, El-Yaniv R and Bengio Y. 2016. Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or -1 [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1602.02830v3.pdf

Denil M, Shakibi B, Dinh L, Ranzato M and de Freitas N. 2013. Predicting parameters in deep learning//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 2148-2156

Ding X H, Ding G G, Guo Y C and Han J G. 2019. Centripetal SGD for pruning very deep convolutional networks with complicated structure//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4938-4948 [DOI: 10.1109/CVPR.2019.00508]

Gao Y. 2013. Uncertain Graph and Uncertain Network. Beijing: Tsinghua University (高原. 2013. 不确定图与不确定网络. 北京: 清华大学)

Guo Y W, Yao A B and Chen Y R. 2016. Dynamic network surgery for efficient DNNs [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1608.04493.pdf

Han S, Pool J, Tran J and Dally W J. 2015. Learning both weights and connections for efficient neural networks [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1506.02626.pdf

He Y, Liu P, Wang Z W, Hu Z L and Yang Y. 2019. Filter pruning via geometric median for deep convolutional neural networks acceleration//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4335-4344 [DOI: 10.1109/CVPR.2019.00447]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 1097-1105

Lin M B, Ji R R, Wang Y, Zhang Y C, Zhang B C, Tian Y H and Shao L. 2020. HRank: filter pruning using high-rank feature map [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/2002.10179.pdf

Lin S H, Ji R R, Yan C Q, Zhang B C, Cao L J, Ye Q X, Huang F Y and Doermann D. 2019. Towards optimal structured CNN pruning via generative adversarial learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2785-2794 [DOI: 10.1109/CVPR.2019.00290]

Mathieu M, Henaff M and LeCun Y. 2013. Fast training of convolutional networks through FFTs [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1312.5851.pdf

Meng F X, Cheng H, Li K, Xu Z X, Ji R R, Sun X and Lu G M. 2020. Filter grafting for deep neural networks [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/2001.05868.pdf

Rastegari M, Ordonez V, Redmon J and Farhadi A. 2016. XNOR-Net: ImageNet classification using binary convolutional neural networks//European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 525-542 [DOI: 10.1007/978-3-319-46493-0_32]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1409.1556.pdf

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Yang Z C, Moczulski M, Denil M, de Freitas N, Smola A, Song L and Wang Z Y. 2014. Deep fried convnets [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1412.7149.pdf

Yu R C, Li A, Chen C F, Lai J H, Morariu V I, Han X T, Gao M F, Lin C Y and Davis L S. 2018. NISP: pruning networks using neuron importance score propagation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9194-9203 [DOI: 10.1109/cvpr.2018.00958]

Zhang W. 2017. K-nearest neighbor search algorithm for uncertain graphs based on sampling. Computer Applications and Software, 34(6): 180-186 (张伟. 2017. 基于抽样的不确定图k最近邻搜

索算法. 计算机应用与软件，34(6)：180-186)［DOI：10.3969/j. issn. 1000-386x. 2017. 06. 033］

Zhou A J, Yao A B, Guo Y W, Xu L and Chen Y R. 2017. Incremental network quantization: towards lossless CNNs with low-precision weights［EB/OL］.［2020-06-30］. https://arxiv. org/pdf/1702.03044. pdf

作者简介

胡骏，1985 年生，男，博士研究生，主要研究方向为计算机视觉、自动驾驶。
E-mail：hu. jun@ reachauto. com

刘威，通信作者，男，教授，主要研究方向为计算机视觉、深度学习、多传感器融合及路径规划控制。
E-mail：lwei@ neusoft. com

黄启鹏，男，硕士研究生，主要研究方向为深度学习、模型压缩。E-mail：huangqp@ neusoft. com

刘嘉昕，男，博士研究生，主要研究方向为模式识别、深度学习、模型压缩。E-mail：liu. jiaxin@ reachauto. com

袁淮，男，副教授，主要研究方向为计算机视觉。
E-mail：yuanh@ reachauto. com

赵宏，男，教授，主要研究方向为计算机视觉、高性能计算。
E-mail：zhaoh@ reachauto. com