·人工智能与模式识别·　　文章编号：1000-3428(2023)05-0097-08　　文献标志码：A　　中图分类号：TP18

基于彩票假设的软剪枝算法

马嘉翔，宋晓宁

（江南大学 人工智能与计算机学院，江苏 无锡 214122）

摘　要：神经网络层数的不断增加使网络复杂度也呈指数级上升，导致应用场景受到限制。提出一种基于彩票假设的软剪枝算法实现网络加速。通过使用前一阶段的剪枝网络对其进行知识蒸馏来补偿的方法恢复错误参数，并在知识蒸馏的损失函数中加入稀疏约束来保持稀疏性。在此基础上，将当前阶段得到的剪枝网络与知识蒸馏得到的学生网络进行融合。在进行网络融合时，计算剪枝网络与学生网络的相似性，并通过设计特定的融合公式来突出相近的网络参数和抑制相离的网络参数，使得网络在剪枝率提高后仍然表现良好。在CIFAR-10/100数据集上对VGG16、ResNet-18和ResNet-56模型进行实验，结果显示：剪枝率为80%时，VGG16在CIFAR-10数据集上的分类精度下降0.07个百分点；剪枝率为60%时，ResNet-56在CIFAR-10数据集上的分类精度提升0.06个百分点；剪枝率为85%、95%和99%时，ResNet-18在CIFAR-100数据集上的分类精度仅下降1.03、1.51和2.04个百分点。实验结果表明，所提算法在提高网络剪枝率的同时仍能使其保持较高的精度，验证了算法的有效性。

关键词：网络加速；彩票假设；全局剪枝；稀疏蒸馏；模型融合

开放科学（资源服务）标志码（OSID）：

中文引用格式：马嘉翔，宋晓宁.基于彩票假设的软剪枝算法[J].计算机工程，2023，49(5)：97-104.

英文引用格式：MA J X，SONG X N.Soft pruning algorithm based on lottery ticket hypothesis[J].Computer Engineering，2023，49(5)：97-104.

Soft Pruning Algorithm Based on Lottery Ticket Hypothesis

MA Jiaxiang，SONG Xiaoning

（School of Artificial Intelligence and Computer Science，Jiangnan University，Wuxi 214122，Jiangsu，China）

【Abstract】The increasing number of neural network layers exponentially increases the network complexity and limits its application scenarios.To solve this problem，this study proposes a soft pruning algorithm based on lottery ticket hypothesis. The pruning network of the previous stage is used to compensate for knowledge distillation.To maintain the sparsity in knowledge distillation，the wrongly-pruned parameters are recovered and sparse constraints are added to its loss function. Subsequently，the pruning network obtained at the current stage is integrated with the student network obtained through knowledge distillation.The similarity between the pruning and student networks during the network fusion is then calculated and a specific fusion formula is designed to highlight similar network parameters and inhibit discrete network parameters. Consequently，the network continues to perform well after the pruning rate is increased.The experimental results of VGG16，ResNet-18，and ResNet-56 models on CiFAR-10/100 dataset indicate the following：when the pruning rate is 80%，the classification accuracy of VGG16 in CIFAR-10 dataset decreases by 0.07 percentage points；when the pruning rate is 60%，the classification accuracy of ResNet-56 in CIFAR-10 dataset is improved by 0.06 percentage points；and when the pruning rates are 85%，95%，and 99%，the accuracy of ResNet-18 on CIFAR-100 dataset only decreased by 1.03，1.51，and 2.04 percentage points，respectively.This shows that the proposed algorithm can improve the pruning rate of the network while maintaining high accuracy，thus，proving the effectiveness of the proposed algorithm.

【Key words】network acceleration；lottery ticket hypothesis；global pruning；sparse distillation；model fusion

DOI：10. 19678/j. issn. 1000-3428. 0064139

0　概述

图像分类是机器学习领域的热门研究方向[1]。卷积神经网络算法早在二十世纪就被提出，但直到二十一世纪随着深度学习算法的出现以及数值计算设备性能的大幅提升，神经网络才逐渐取代了传统的机器学习方法，在各类图像分类任务中被广泛应用。

基金项目：国家自然科学基金（61876072）；国家社会科学基金（21&ZD166）；江苏省自然科学基金（BK20221535）。

作者简介：马嘉翔（1996—），男，硕士研究生，主研方向为网络模型加速；宋晓宁（通信作者），教授、博士、博士生导师。

收稿日期：2022-03-09　　**修回日期：**2022-05-01　　**E-mail：**x.song@jiangnan.edu.cn

AlexNet[2]是一个较传统方法有极大突破的神经网络算法，其在ImageNet数据集上超越了人类的分类精度。随后网络的深度被不断地提升：VGG[3]在AlexNet的基础上增加了11层网络，达到了19层；ResNet[4]将网络最高增加到了152层，并加入了重复的残差模块。随着模型大小和数据量的激增，神经网络算法已在图像、语音、视频等领域取得了远超传统算法的突破性效果[1]，并逐渐被应用于日常生活、工业生产和科技研究中[5]。但随之而来的是日益高昂的存储和计算成本，这使得网络的部署变得越来越难。以AlexNet为例，它不但计算量高达105 MFlops，大小也达到200 MB，这远远超过了边缘设备和移动设备的承载能力，使得神经网络的适用范围受到了极大的限制。

为了解决这一问题，大量的网络加速方法被提出：HOWARD等[6]提出了深度可分离卷积，将$n\times n$的二维卷积分解为$n\times n$的一维卷积和1×1的二维卷积的组合，这大大降低了卷积的计算量；之后SANDLER[7]提出的改进版则先使用1×1的二维卷积对特征进行升维，再使用深度可分离卷积提取特征，最后用1×1的卷积对特征进行降维，在降低计算量的同时也兼顾了性能的提升；WU等[8]提出将浮点型参数转化为定点型来降低内积计算量及神经网络在一些特殊设备上的存储压力；LUO等[9]通过比较剪去不同参数后网络的可重建性来找到网络中冗余的参数；ROMERO等[10]使用教师网络的中间表示作为学生网络的辅助损失来提升性能表现。

剪枝和蒸馏是网络加速领域中的常用方法。对网络进行剪枝能够得到剪枝率更高的网络，使得网络的体积与计算消耗变小，从而可以运行在存储和计算能力受限的设备上。剪枝率高的网络在前向传播时花费的时间也较少，可应用于一些对延迟要求较高的任务，如实时目标检测等。

大型网络较小型网络拥有更强的表征能力及泛化性，因此比起学习数据集的特征分布，通过知识蒸馏帮助小型网络直接学习大型网络从数据集中提取的特征信息能够使小型网络拥有更好的精度表现。

在复杂数据集和高剪枝率条件下保持网络的精度是网络加速领域的一个热门问题。为了解决这一问题，本文提出一种基于彩票假设的软剪枝算法。分别迭代地对网络进行非结构化剪枝和稀疏蒸馏，并通过计算其结果的相关性来获得原始网络的最优子网络。为了使网络能够在通用设备上得到加速，对得到的稀疏网络进行结构化剪枝。最后，在不同的神经网络及数据集上进行实验并与其他剪枝算法进行比较，验证本文算法的有效性。

1 相关工作

剪枝可以帮助冗余的网络变得紧凑，从而降低其存储及计算成本。与对每层网络执行相同剪枝比例的分层剪枝方法相比，全局剪枝因为可以对各层网络灵活分配剪枝比例，往往能够达到更好的剪枝效果。WANG等[11]通过计算滤波器间的皮尔逊相关系数来排序各个分类器的可替代性。LIU等[12]在BN层参数中加入可学习的缩放因子，并剪去较小学习因子对应的通道。CHIN等[13]学习不同层滤波器的全局排名，并通过修剪排名靠后的滤波器来获得一组具有不同精度/延迟权衡的结构。

彩票假设是一类经典的全局剪枝算法。FRANKLE等[14]的研究表明，在某些情况下，稀疏子网络可以成功地重新训练并产生较其原始密集网络更好的性能，而且通常在重新训练的计算预算较小的情况下也是如此。这一观察导致了彩票假说，它推测对于一个合理大小的网络f和随机初始化的参数$w^{(0)}\in\mathbb{R}^d$，存在一个稀疏子网络f'，规定约束$m\in\{0,1\}^d$，$\|m\|_0<<d$，可以令网络从$w'^{(0)}=m\otimes w^{(0)}$进行训练，使其性能与原始模型$f$的训练版本相当。迭代强度剪枝（Iterative Magnitude Pruning，IMP）[15]的提出支持了这一假设。IMP能够在为图像分类训练的卷积网络中找到这样的子网络，称为中奖彩票。它运行多轮，以离散的时间间隔稀疏化网络，并在运行过程中生成剪枝率增加的子网络。SAVARESE等[16]提出了IMP的一般形式，其将网络的权重倒回到迭代值$w^{(k)}$，而不是原始初始化值$w^{(0)}$。

MALACH等[17]在数学上证明了对于每个有界分布和每个具有有界权重的目标网络，具有随机权重的充分过参数化的神经网络都包含一个子网络，其大致为与目标网络相同的精度，不需要任何进一步的训练。ZHOU等[18]随后解释了为什么使用掩码对初始化网络进行剪枝以及将参数置为0在彩票猜想中是重要且有效的。但FRANKLE等[19]的研究在处理深层网络时表现不佳，为了解决这一问题，他们提出不在初始化网络上进行剪枝，而是在迭代的过程中对网络进行剪枝，从而在深层网络上获得较好的效果。

目前，已有研究指出了剪枝过程中恢复被错误剪去的参数对网络精度的重要性。PRAKASH等[20]令被剪去的参数重新初始化以与完整网络正交来去除网络的冗余。HE等[21]提出了软剪枝方法，它容许被剪去的参数在微调阶段继续更新，并不断迭代来获得更好的效果。尽管此类方法在精度上较其他算法表现良好，但前者仅将剪枝作为搜索冗余参数的方法，并未增加网络的压缩率；后者通过训练恢复重要参数的方法则缺乏定量的数学描述，因此存在优化的空间。

2 本文算法

2.1 非结构化剪枝

按照剪枝的粒度可将剪枝分为结构化剪枝与非结构化剪枝，结构化剪枝剪去网络中的结构，如通道、滤波器等，而非结构化剪枝则将网络中不重要的参数置为0来增加网络的压缩率。结构化剪枝因为剪枝

的粒度较大，剪枝的自由度较差，所以精度表现相对较差，但可以被直接应用于各类设备来加速网络的推理。非结构化剪枝的剪枝粒度为网络中最小的单位参数，剪枝的自由度最高，因此有着最好的精度表现，但它需要特定的软硬件架构才能取得较好的加速效果，如NVIDIA于2020年推出的安培（Ampere）架构提供了对稀疏度50%的网络的加速支持[22]。随着显卡技术的不断发展，对非结构化剪枝的加速支持的加深会越来越体现其精度表现较好的优势。

在进行非结构化剪枝时，需要对网络参数的重要性进行评估并剪去相对不重要的参数。本文选择L_1范数作为重要性的评估标准，如式（1）所示：

$$I_i=\left\|W_i\right\|_1 \tag{1}$$

其中：W_i是网络中第i个参数；I_i为其重要性参数。

通过式（1）得到网络中每个参数的重要性参数集合$I=\{I_1,I_2,\cdots,I_n\}$后，根据当前阶段网络的剪枝率τ及式（2）得到剪枝的阈值thres_u：

$$\text{thres}_u=\text{sort}_\tau(I) \tag{2}$$

thres_u为剪枝率τ对应的剪枝阈值。$\text{sort}_\tau(I)$将网络中所有参数W的重要性按照升序进行排序，并返回第$n\times\tau$位置的参数值作为当前迭代轮次的剪枝阈值（n为网络中的参数总量），它同时对网络中所有参数进行排序及剪枝，因此又称为全局剪枝。

得到阈值后生成掩码m，它的形状与网络中所有参数一致，约束为$m\in\{0,1\}^d$，$\left\|m\right\|_0<<d$。它将在网络中所有重要性不及阈值的参数所对应位置的值置为0，其余位置的值置为1。在使用m对网络参数进行点乘（如式（3）所示）后，第k轮剪枝结束。

$$w'^{(k)}=m\cdot w^{(k)} \tag{3}$$

2.2　稀疏蒸馏

文献[20-21]已经证明恢复网络被错误剪去的参数能够使网络保留原先的表达能力，并更好收敛，但所提方法仅从当前阶段的剪枝网络获得信息。本文通过使用前一阶段剪枝网络对当前阶段剪枝网络进行蒸馏的方法来帮助网络获得关于剪枝操作的启发式信息，从而在提高网络剪枝率的同时获得更好的训练稳定性以及网络精度。

通过增加模型间的相关性可以帮助被训练的学生网络的概率分布更接近帮助其训练的教师网络[23]，因而得到更好的精度表现。常用的方法是将教师网络与学生网络的KL散度（如式（4）、式（5）所示）作为蒸馏损失加入学生网络的损失函数[24]。这是因为使用KL距离能够很好地衡量模型间的区别，且其作为Loss时优化过程是凸的，更容易使网络收敛到最优解。

$$L_D(t,s)=-\sum_i p(t_i)\frac{\ln p(s_i)}{\ln p(t_i)} \tag{4}$$

$$p(x_i)=\frac{\exp(x_i/T)}{\sum_j\exp(x_j/T)} \tag{5}$$

式（4）为网络的蒸馏损失，又称软标签，它体现了教师网络与学生网络的相似程度：它们越是相似，蒸馏损失就越小，学生网络的分类效果就越好。其中：t为教师模型；s为学生模型；x_i为模型x每个类别i在全连接层的输出；$p(x_i)$为模型x属于第i类的概率；T是控制教师网络对学生网络影响程度的超参数，一般来说，T越大，教师网络对学生网络的影响就越大，但当T趋向于无穷时，教师网络对学生网络的影响就趋向于无。

考虑到训练大型网络的成本，许多蒸馏方法的思路从使用大型网络来蒸馏小型网络转向使用小型网络自身作为教师网络进行知识蒸馏。XU等[25]提出在训练BERT模型时将前k个时间步的参数取平均作为教师网络来对当前时间步模型进行知识蒸馏。KIM等[26]将这一概念引入CNN领域，令上个epoch的模型作为教师网络来对当前epoch的模型进行知识蒸馏。实验证明，这类做法能够很好地提高网络的训练稳定性及精度表现。本文将这一思路与剪枝方法相结合，在迭代时使用上一迭代生成的网络模型$f_{F_{i-1}}$作为教师网络对剪枝网络f_{F_i}进行蒸馏：

$$L_D(f_{F_{i-2}},f_{F_{i-1}})=-\sum_i p(f_{F_{(i-2)i}})\frac{\ln p(f_{F_{(i-1)i}})}{\ln p(f_{F_{(i-2)i}})} \tag{6}$$

知识蒸馏的损失由衡量教师网络与学生网络区别的蒸馏损失L_D与学生网络的分类损失函数L_C构成，一般选择交叉熵作为分类损失函数：

$$L_C=-[y\ln\hat{y}+(1-y)\ln(1-\hat{y})] \tag{7}$$

其中：y为数据集标签；$\hat{y}$为网络的预测值。

L_1正则化已被证明是增加网络稀疏度的优秀方法[12,27-28]，它能在保证训练后网络精度的同时有效地帮助网络适应剪枝操作，进而在剪枝后保证网络的精度。本文在知识蒸馏的损失函数中加入L_1正则化（如式（8）所示）来帮助学生网络在蒸馏的过程中保持一定的稀疏度，使模型在后续的融合及剪枝步骤后精度能够尽量保持不下降，因此称之为稀疏蒸馏。

$$L=\alpha L_D+(1-\alpha)L_C+\lambda\left\|W\right\|_1 \tag{8}$$

其中：α为平衡蒸馏损失与分类损失的超参数，实验表明选择0.6效果最好；λ控制蒸馏过程中对网络稀疏性的约束，实验表明选择10^{-4}时效果最好。

2.3　模型融合

在得到了剪枝率更高的剪枝网络及使用稀疏蒸馏得到的蒸馏网络后，本文通过式（9）将两者融合得到融合网络：

$$W_F=\frac{1}{2}\times\frac{W_P+W_D}{\sqrt{\left|W_D-W_P\right|}} \tag{9}$$

其中：W_P为剪枝网络的参数；W_D为蒸馏网络的参数；W_F为融合网络的参数。当W_P与W_D的值接近（两者差值的绝对值小于1）时，得到的W_F会是一个较大的值，如图1融合模型卷积核的第1行第2列，这就加强了模型中相似的部分；当W_P与W_D的值较为背离（两者差值的绝对值小于1）时，得到的W_F会是一个

较小的值，如图1融合模型卷积核的第2行第3列，这就抑制了模型中不相似的部分。

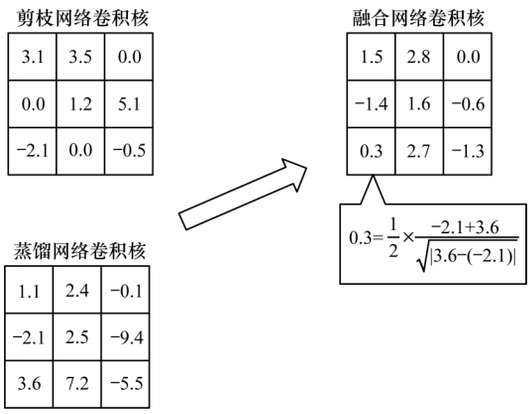

图1 模型融合示意图

Fig.1 Schematic diagram of model fusion

当卷积网络的参数被剪去并置为0时，融合模型卷积核值的计算公式就由式(9)退化为式(10)。若被剪去的参数并不重要，则蒸馏网络对其弥补的效果也非常有限，如图1融合模型卷积核的第1行第3列所示；若被剪去的参数较为重要，需要对其进行弥补，将会得到一个来自蒸馏网络的被抑制的弥补值，如图1融合模型卷积核的第3行第2列所示，这就控制了蒸馏网络对剪枝网络的影响，避免剪枝操作受到过大的影响。

$$W_F=\frac{1}{2}\times\sqrt{|W_D|} \tag{10}$$

在得到融合模型后，本文对其进行微调，得到的融合模型与融合前的模型相比，两者中更相似的值将更偏离零，而更不相似的值更贴近零，这使相似参数的重要性得到了提升，不相似的参数的重要性得到了抑制，方便算法在后续的剪枝操作中剪去不相似的参数而保留相似的参数，从而提升网络的剪枝效果。

以图1中的融合模型为例，若下一阶段的剪枝率为50%，则其剪枝阈值为1.4，所有绝对值小于它的参数的重要性都低于它，因此将在下一阶段被剪去，而绝对值大于它的将被保留。

笔者认为，彩票假设所要寻找的最优子网络同时存在于剪枝网络与蒸馏网络中，因此模型融合的目的是通过加强融合模型中相似的参数，抑制融合模型中不相似的参数来找到这个子网络，如前文所述，利用式(9)能够很好地达到这一效果。

同时，本文使用稀疏蒸馏来借助前一阶段网络的信息弥补剪枝网络中被错误剪去的参数，从而使网络在剪枝后能够获得较高的精度。与剪枝后参数值不再进行更新的硬剪枝相比，本文算法过程中被剪去的参数在后续迭代过程中仍能进行更新，因此称之为软剪枝。

2.4 滤波器剪枝

当结束对网络的全局剪枝后，网络中90%的参数已被置为0，本文定义L_0范数比例为网络中每层值为0的参数占总体数量的比例，如图2所示。可以看出，在网络的第1层及最后4层，网络中值为0的参数所占的比例是非常高的，如果对此时得到的稀疏网络进行滤波器剪枝，网络精度所受到的影响并不大。因此，本文尝试在全局剪枝的基础上，对得到的网络进行滤波器剪枝，这样无论是在适用于非结构化剪枝的特殊架构的软硬件环境，还是适用结构化剪枝的通用环境下，本文算法都能够对网络进行压缩与加速，从而大幅拓宽本文算法及神经网络的应用场景。

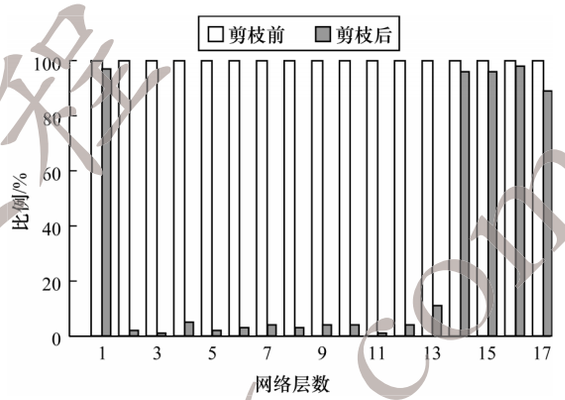

图2 卷积层L_0范数比例

Fig.2 L_0 norm ratios for different convolutional layers

经过全局剪枝后，网络中许多参数的值非常接近0，本文利用式(11)计算网络中各个滤波器的L_2范数，将其作为衡量滤波器重要性的指标。

$$E_i=\sum_j^k\left\|W_j\right\|_2 \tag{11}$$

其中：E_i为第i个滤波器的重要性参数；k为滤波器i中的卷积核数；W_j为滤波器i中第j个卷积核的参数。

通过式(11)得到网络中每个滤波器的重要性参数集合$E=\{E_1,E_2,\cdots,E_m\}$后(m为网络中滤波器的总数)，根据事先设定的剪枝率θ_s及式(12)得到滤波器剪枝的阈值thres_s：

$$\text{thres}_s=\text{sort}_{\theta_s}(E) \tag{12}$$

其中：$\text{sort}_{\theta_s}(E)$将网络中所有滤波器的重要性参数按照升序进行排序，并返回第$m\times\theta_s$位置滤波器的重要性参数作为滤波器剪枝的阈值。得到阈值后将网络中所有重要性参数不及它的滤波器从网络中剪去，并在经过微调后得到最终的输出网络。这个网络剪去了冗余的滤波器，因此可以在通用设备上对其推理过程进行加速。

2.5 算法总体流程

在每个迭代过程中，网络信息的流向如图3所示。输入网络的信息一部分流向上方的剪枝路来获得剪枝率更高的网络，另一部分流向下方的蒸馏路，经上一阶段剪枝网络对其进行的稀疏蒸馏获得来自之前阶段网络的信息，从而保持网络的稳定性。随后经过模型融合来增强两者中相似的部分，而抑制不相似的部分，进而得到更接近最优子网络的网络。

这时的输出网络因为得到了稀疏蒸馏的弥补，其L_0范数会小于等于融合前的剪枝网络，因此网络的剪枝率会有所下降。当迭代结束后需要输出网络时，须对其进行一次额外的剪枝与微调来得到最终剪枝率达到要求的输出网络。

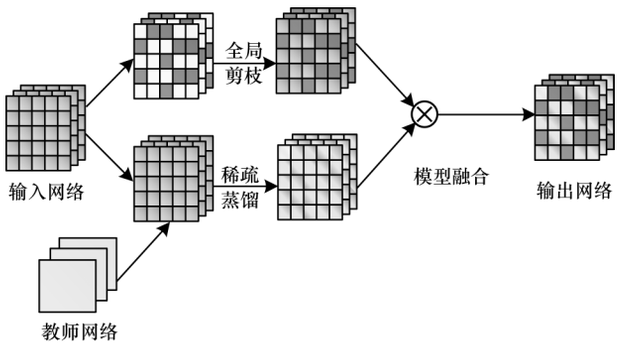

图3　网络信息流向图

Fig.3　Network information flow diagram

本文算法流程如算法1所示。

算法1　基于彩票假设的软剪枝算法

输入　网络模型f，输入数据x，全局剪枝率θ_u，滤波器剪枝率θ_s，总迭代轮次T，当前迭代轮次t

输出　稀疏网络F

1）令t=1，对网络f进行剪枝率为$\frac{\theta_u}{T}$的剪枝得到网络f_{P_1}，进入步骤3）。

2）令$\tau=t\times\frac{\theta_u}{T}$，对网络$f_{P_{t-1}}$进行剪枝率为$\tau$的全局剪枝得到网络$f_{P_t}$。

3）若t=1，教师网络与学生网络均选择网络f；若t=2，教师网络选择f_{P_1}，否则教师网络选择$f_{P_{t-2}}$，学生网络选择$f_{P_{t-1}}$。使用教师网络对学生网络进行带稀疏约束的蒸馏得到蒸馏网络f_{D_t}。

4）计算网络f_{P_t}和网络f_{D_t}的相似性并进行融合得到网络f_{P_t}，对其进行微调。

5）若$\tau=\theta_u$，对网络f_{P_t}进行剪枝率为θ_u的剪枝，并进入步骤6）；否则令$t=t+1$，回到步骤2）。

6）对微调后的网络进行剪枝率为θ_s的滤波器剪枝，微调后输出网络F。

3　实验与结果分析

3.1　数据集及实验设置

CIFAR-10和CIFAR-100是常用的剪枝算法有效性验证数据集。为了验证本文算法的有效性，在这2个数据集上进行实验。2个数据集的图片大小均为32×32像素，类别数分别为10和100，训练集和测试集的图片数量分别为50 000张和10 000张。在VGG16、ResNet-18和ResNet-56上测试本文算法的压缩效果。采取SGD作为网络的优化方法，批量训练样本数量（batch size）为64，训练轮次为240个epoch。学习率设置为0.1，在120和180个epoch时将其衰减为先前的10%，权重衰减（weight decay）设置为5×10^{-4}，动量（momentum）为0.9。

在进行非结构化剪枝时，初始剪枝率为0.1，每轮迭代增加0.1，共迭代9次，最多达到0.9，这意味着网络中90%的参数都将被置为0；在进行结构化剪枝时，网络的剪枝率为0.4，这意味着网络中40%的滤波器被剪去，从而对网络的推理起到加速作用。

3.2　算法结果分析

本文使用VGG16和ResNet-56网络在CIFAR-10上测试算法性能。为了更好地进行效果对比，以剪枝前的模型作为基线，将文献[29-39]方法的剪枝率及分类精度的下降比例与本文算法进行对比，结果如表1所示。可以看出，本文算法在剪枝率远高于其他算法的同时，精度损失与其他算法仍相近，甚至优于一些算法。对VGG16网络，除了剪枝率为33.77%、精度较剪枝前下降0.06个百分点的SCP算法，以及剪枝率为30.55%、精度较剪枝前上升0.10个百分点的GDP算法，本文算法无论是在剪枝率指标还是在精度变化指标上均优于其他算法。尤其是对Hinge，本文在剪枝率高出19.17个百分点的情况下，精度下降指标还高出0.36个百分点。对ResNet-56网络，本文算法无论是在剪枝率指标还是在精度变化指标上均比其他算法的表现更好，是唯一剪枝后精度上升的算法。

表1　在CIFAR-10数据集上的测试结果

Table 1　Test results on CIFAR-100 dataset

模型	算法	基线精度/%	精度/%	精度变化/百分点	剪枝率/%
VGG16	SCP[29]	93.85	93.79	−0.06	33.77
	Hinge[30]	94.02	93.59	−0.43	**60.93**
	HRank[31]	93.96	93.43	−0.53	46.40
	VCNNP[32]	93.25	93.18	−0.07	39.10
	GDP[33]	93.89	93.99	**0.10**	30.55
	本文算法	93.90	93.83	−0.07	80.00
ResNet-56	PFS[34]	93.23	93.05	−0.18	**50.00**
	ABCPrune[35]	93.26	93.23	**−0.03**	45.87
	HRank	93.26	93.17	−0.09	**50.00**
	SCOP[36]	93.70	93.64	−0.06	44.00
	LeGR[37]	93.90	93.70	−0.20	47.00
	本文算法	94.06	94.12	0.06	60.00

本文使用ResNet-18网络在CIFAR-100上测试算法性能。为了更好地进行效果对比，以剪枝前的模型作为基线，将文献[38-39]方法的剪枝率和精度变化与本文算法进行对比，结果如表2所示。可以看出，本文算法在剪枝率与其他算法相同的情况下，精度下降的情况远远好于其他算法，且随着网络剪枝率的提升，本文算法的效果也越来越好。当剪枝率为85%时，本文算法只是略好于其他算法，而当剪枝率为99%时，本文算法的精度变化指标则优于其

他算法十多个百分点，这是因为本文算法在迭代的过程中能够从剪枝率较低的网络中获得关于剪枝的启发式信息，从而在较高剪枝率时相较其他算法也能够保持网络的稳定。

表2 在CIFAR-100数据集上的测试结果

Table 2 Test results on CIFAR-100 dataset

算法	剪枝率为85%		剪枝率为95%		剪枝率为99%	
	精度/%	精度变化/百分点	精度/%	精度变化/百分点	时的精度/%	精度变化/百分点
基线	74.53	—	74.53	—	74.53	—
SNIP[38]	71.61	−2.92	68.39	−6.14	56.69	−17.84
GraSP[39]	71.16	−3.37	68.40	−6.13	58.67	−15.86
本文算法	73.50	−1.03	73.02	−1.51	72.49	−2.04

本文算法CIFAR-10上的表现只是略胜于其他算法，但在CIFAR-100上的表现则远强于其他算法，这说明相比简单数据集，本文算法更适用于复杂数据集的网络加速与压缩任务。

3.3 消融实验

本文使用VGG16及ResNet-18分别在CIFAR-10和CIFAR-100数据集上进行消融实验，结果如表3所示。

表3 消融实验效果

Table 3 Ablation experimental results %

方法	精度	
	VGG16	ResNet-18
本文算法	92.83	72.49
仅进行剪枝	91.04	70.01
对剪枝网络进行蒸馏	92.46	71.03
使用完整网络作为教师网络	92.54	70.89
蒸馏时未加入稀疏约束	92.31	71.21
仅进行剪枝时加入稀疏约束	91.47	70.54
将剪枝与蒸馏模型线性相加	92.35	71.70
网络每层剪枝率相同	92.33	71.28

对消融实验数据进行分析：

1）仅对网络进行迭代剪枝时分类精度不及本文算法，这证明了网络在迭代剪枝的过程中会剪去一些对剪枝率更高的网络来说重要的参数，因此在进一步剪枝后精度表现不佳，而本文通过稀疏蒸馏与模型融合来对错误剪去的参数进行弥补的方法改善了这一问题。

2）与将网络分为剪枝与蒸馏并进行模型融合的方法不同，本文还尝试了使用剪枝前的网络对剪枝后的网络进行蒸馏的方法。这种做法能够在一定程度上弥补较高剪枝率的网络中那些被错误剪去的参数，但最终的结果不如本文算法，这说明本文融合模型的方法是有效的。这是因为通过蒸馏弥补的参数并不一定就是更高剪枝率的网络所需要的重要参数，如果弥补了错误的参数，反而会影响网络的精度表现。本文在进行融合时计算了蒸馏网络与剪枝网络的相关性，使得弥补的参数对网络是有用的，并抑制了对网络来说不重要的值，因此在一定程度上提高了网络的稀疏性，这为后续的剪枝操作提供了帮助。

3）当使用上一时刻的网络进行蒸馏时，与始终使用未被剪枝的网络进行蒸馏相比能取得更好的精度，这说明来自上一时刻剪枝率较低网络在作为剪枝率更高的网络的教师网络时，其关于剪枝的先验知识能够对后者的剪枝操作起到启发式作用，因此帮助网络取得了更好的剪枝效果。

4）当在蒸馏时未加入稀疏约束时，网络的精度表现不及本文算法，当在仅对网络进行剪枝时，在网络的微调阶段在网络的损失函数内加入稀疏约束的效果虽不及本文算法，但优于不加入稀疏约束的仅进行剪枝的算法，这说明稀疏约束对剪枝算法来说是正向的提升。这是因为网络参数的重要性往往与其绝对值呈正相关，而稀疏约束控制了一部分参数的绝对值大小，使得剪去它们对网络的影响较小，从而得到更好的精度表现。

5）线性相加是模型融合时常见的方法，本文尝试将剪枝网络与蒸馏网络进行线性相加获得融合模型来替代本文的融合方法，其分类精度不及本文算法。这是因为线性相加虽然能够弥补网络中被错误减去的参数，但它对网络中所有参数均一视同仁地进行线性相加，而不能区分对网络来说重要的和不重要的参数，并进行针对性地增强与抑制，因此效果不及本文算法。

6）本文尝试在网络的每一层使用相同剪枝率的分层剪枝方法来代替本文根据参数的重要性在每层灵活分配剪枝率的全局剪枝方法，其分类精度不及本文算法。这是因为与分层剪枝相比，全局剪枝具有更高的自由度，在一些情况下可以达到与分层剪枝相同的结果，因此在正确的剪枝算法的约束下，全局剪枝的分类精度应不弱于分层剪枝，这说明了本文选择全局剪枝的正确性。

3.4 非结构剪枝分类精度与剪枝率关系实验

ResNet-56及ResNet-18在CIFAR-10及CIFAR-100上进行全局剪枝时网络的分类精度随剪枝率变化而变化的曲线如图4所示。可以看出，2个网络的分类精度随着剪枝率的上升均先提升一段时间后再下降，且在剪枝率超过60%后剪枝网络的分类精度开始严重下降。

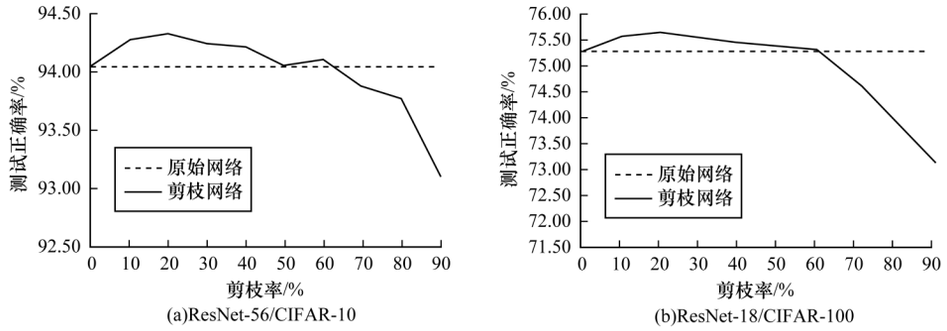

图4　剪枝网络精度随剪枝率变化曲线

Fig.4　Curve of pruning network precision with pruning rate

VGG 16和ResNet-56在CIFAR-10及ResNet-18在CIFAR-100上剪枝后在各个指标上的表现如表4所示。其中：网络的参数量代表了网络在硬件设备中占据的存储空间，参数量越少，网络对存储设备的要求就越低，因此剪枝后网络参数量的下降比例能够直观表示算法对网络的压缩能力；计算量为网络在进行推理过程时每秒的浮点计算数，计算量下降得越多，网络的复杂性就越低，在相同性能的设备中运行所需的时间就越少，在严格要求算法运行时长或时延的场景下对设备计算性能的要求就越低，这体现了算法对网络的加速能力。由表4可以看出，在经过本文的滤波器剪枝后，网络的参数量及计算量都有一定程度的下降，但与此同时网络的精度却仅有轻微的下降，这验证了本文剪枝算法的有效性。

表4　剪枝算法的压缩及加速效果

Table 4　Compression and acceleration effect of pruning algorithm

数据集	模型	剪枝率/%	参数量/10^5	参数量下降率/%	计算量/10^9	计算量下降率/%	精度/%	精度变化/百分点
CIFAR-10	VGG 16	0	147.00	—	30.7	—	93.90	—
		40	47.50	32.3	19.6	36.2	93.57	−0.33
	ResNet-56	0	5.93	—	8.71	—	94.06	—
		40	4.53	23.6	5.47	37.2	93.87	−0.19
CIFAR-100	ResNet-18	0	0.37	—	1.44	—	73.50	—
		40	0.23	37.8	0.87	30.6	71.27	−2.23

4　结束语

本文通过对网络进行非结构化的全局剪枝并使用前一阶段的剪枝网络，对当前阶段的剪枝网络进行稀疏蒸馏，再将两者相融合得到当前阶段的输出网络，如此迭代数轮得到非结构化剪枝的网络。然后对其进行结构化的滤波器剪枝来使网络能够结合结构化剪枝及非结构化剪枝的优点，从而兼顾网络，起到正则化的作用，使网络的精度得到提升。

本文分别使用不同的神经网络在CIFAR-10和CIFAR-100数据集上进行实验，结果表明本文算法能够在有效降低网络对存储及计算设备要求的同时保持网络的精度表现，从而拓宽网络的应用场景。但本文算法在剪枝率超过90%后精度下降较为严重，后续将考虑对算法进行优化，在对剪枝率要求较高的情况下获得较好的精度表现。

参考文献

[1] SZEGEDY C，LIU W，JIA Y Q，et al. Going deeper with convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2015：1-9.

[2] ALEX K，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[3] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego，USA：[s. n.]，2015：1-10.

[4] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：770-778.

[5] 郭子博，高瑛珂，胡航天，等. 基于混合架构的卷积神经网络算法加速研究[J]. 计算机工程与应用，2022，58(6)：88-94.
GUO Z B，GAO Y K，HU H T，et al. Research on acceleration of convolutional neural network algorithm based on hybrid architecture[J]. Computer Engineering and Applications，2022，58(6)：88-94.（in Chinese）

[6] HOWARD A G，ZHU M L，CHEN B，et al. MobileNets：efficient convolutional neural networks for mobile vision applications[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1704.04861.

[7] SANDLER M，HOWARD A，ZHU M L，et al. MobileNetV2：inverted residuals and linear bottlenecks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2018：4510-4520.

[8] WU J X, LENG C, WANG Y H, et al. Quantized convolutional neural networks for mobile devices[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 4820-4828.
[9] LUO J H, WU J X, LIN W Y. ThiNet: a filter level pruning method for deep neural network compression[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 5068-5076.
[10] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1412.6550.
[11] WANG W X, FU C, GUO J S, et al. COP: customized deep model compression via regularized correlation-based filter-level pruning[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1906.10337.
[12] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 2755-2763.
[13] CHIN T W, DING R Z, ZHANG C, et al. Towards efficient model compression via learned global ranking[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 1515-1525.
[14] FRANKLE J, CARBIN M. The lottery ticket hypothesis: finding sparse, trainable neural networks[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1803.03635.
[15] MAENE J, LI M X, MOENS M F. Towards understanding iterative magnitude pruning: why lottery tickets win[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2106.06955.
[16] SAVARESE P, HUGO S, MICHAEL M. Winning the lottery with continuous sparsification[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1912.04427v4.
[17] MALACH E, YEHUDAI G, SHALEV-SHWARTZ S, et al. Proving the lottery ticket hypothesis: pruning is all you need[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s. n.], 2020: 1-10.
[18] ZHOU H, LAN J, LIU R, et al. Deconstructing lottery tickets: zeros, signs, and the supermask[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1905.01067.
[19] FRANKLE J, DZIUGAITE G K, ROY D M, et al. Stabilizing the lottery ticket hypothesis[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1903.01611.
[20] PRAKASH A, STORER J, FLORENCIO D, et al. RePr: improved training of convolutional filters[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10658-10667.
[21] HE Y, KANG G L, DONG X Y, et al. Soft filter pruning for accelerating deep convolutional neural networks[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1808.06866.
[22] MISHRA A, LATORRE J A, POOL J, et al. Accelerating sparse deep neural networks[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2104.08378.
[23] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
[24] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1503.02531.
[25] XU Y G, QIU X P, ZHOU L G, et al. Improving BERT fine-tuning via self-ensemble and self-distillation[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2002.10345.
[26] KIM K, JI B, YOON D, et al. Self-knowledge distillation with progressive refinement of targets[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2022: 6547-6556.
[27] SIMONE S, COMMINIELLO D, HUSSAIN A, et al. Group sparse regularization for deep neural networks[J]. Neurocomputing, 2017, 241: 81-89.
[28] 韦越, 陈世超, 朱凤华, 等. 基于稀疏正则化的卷积神经网络模型剪枝方法[J]. 计算机工程, 2021, 47(10): 61-66.
WEI Y, CHEN S C, ZHU F H, et al. Pruning method for convolutional neural network models based on sparse regularization[J]. Computer Engineering, 2021, 47(10): 61-66. (in Chinese)
[29] WIMMER P, MEHNERT J, CONDURACHE A. COPS: controlled pruning before training starts[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2107.12673.
[30] LI Y W, GU S H, MAYER C, et al. Group sparsity: the hinge between filter pruning and decomposition for network compression[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 8015-8024.
[31] LIN M B, JI R R, WANG Y, et al. HRank: filter pruning using high-rank feature map[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 1526-1535.
[32] ZHAO C L, NI B B, ZHANG J, et al. Variational convolutional neural network pruning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 2775-2784.
[33] GUO Y, YUAN H, TAN J C, et al. GDP: stabilized neural network pruning via gates with differentiable polarization[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2022: 5219-5230.
[34] WANG Y L, ZHANG X L, XIE L X, et al. Pruning from scratch[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12273-12280.
[35] LIN M B, JI R R, ZHANG Y X, et al. Channel pruning via automatic structure search[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2001.08565v1.
[36] TANG Y H, WANG Y H, XU Y X, et al. SCOP: scientific control for reliable neural network pruning[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2010.10732.
[37] CHIN T W, DING R Z, ZHANG C, et al. Towards efficient model compression via learned global ranking[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 1515-1525.
[38] LEE N, AJANTHAN T, TORR P H S. SNIP: single-shot network pruning based on connection sensitivity[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1810.02340.
[39] MARQUES-SILVA J P, SAKALLAH K A. GRASP—a new search algorithm for satisfiability[M]//KUEHLMANN A. The best of ICCAD. Boston, USA: Springer US, 2003: 73-89.

编辑 金胡考