面向星上目标提取的卷积神经网络优化技术

卢　丹[1],孙永岩[1],郑幸飞[2,3],齐保贵[2,3],师　皓[2,3]
(1.上海卫星工程研究所,上海 201109;2.北京理工大学 信息与电子学院 雷达技术研究所,北京 100081;
3.嵌入式实时信息处理技术北京市重点实验室,北京 100081)

摘　要: 针对卷积神经网络规模庞大、参数数量众多、在资源受限的在轨场景中难以应用的问题,提出了一种基于知识蒸馏的剪枝压缩改进方法。该方法对训练好的网络进行基于权重和基于通道的混合参数剪枝,在保留网络重要连接的同时剔除冗余信息;采用知识蒸馏法,用原始网络学到的知识指导剪枝后网络的再训练过程,以恢复损失的精度;在遥感数据集上对VGG-16分类网络进行实验。结果表明:所提方法可以实现16～18倍的压缩效果,并且网络精度下降不到1%。这使得卷积神经网络的在轨应用成为可能,具有理论及现实意义。

关键词: 在轨应用; 卷积神经网络; 网络压缩; 参数剪枝; 知识蒸馏
中图分类号: TN 911.73; TP 391.9　　**文献标志码:** A　　**DOI:** 10.19328/j.cnki.1006-1630.2021.01.013

Convolutional Neural Network Optimization Technology for Satellite Target Extraction

LU Dan[1], SUN Yongyan[1], ZHENG Xingfei[2,3], QI Baogui[2,3], SHI Hao[2,3]
(1.Shanghai Institute of Satellite Engineering, Shanghai 201109, China; 2.Radar Research Lab, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
3.Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing 100081, China)

Abstract: Aiming at the problem that the volume of the convolutional neural network is large and the number of the parameters is numerous, which is difficult to be applied in the resource-limited on-orbit scenario, an improved pruning compression method based on knowledge distillation is proposed. The method adopts the mixed parameter pruning based on both the weights and the channels for a trained network. It reserves the important connection of the network and eliminates redundant information. Furthermore, the knowledge distillation method is used to guide the retraining process of the network after pruning with the knowledge learned from the original network in order to restore the loss accuracy. The VGG-16 classification network is tested on the remote sensing data-set. The results show that the proposed method can achieve 16-18 times compression effect, and the network accuracy is reduced by less than 1%. This makes the on-orbit application of convolutional neural network possible, and has theoretical and practical significance.

Key words: on-orbit application; convolutional neural network; network compression; parameter pruning; knowledge distillation

0　引言

随着遥感成像和信息处理技术的发展,在星上实时完成目标检测、场景分类等任务,直接从遥感数据中获取有用信息,具有十分重要的意义。卷积神经网络凭借其强大的特征提取能力,在遥感数据处理任务中可显著提升信息提取的效率和精度,逐

收稿日期: 2019-11-26; **修回日期:** 2020-03-02
基金项目: 上海航天科技创新基金(F-201812-0034)
作者简介: 卢　丹(1989—),女,硕士,主要研究方向为雷达卫星总体设计、通信、星上数据处理。
通信作者: 郑幸飞(1997—),女,硕士,主要研究方向为遥感图像处理、目标检测。

渐获得广泛应用[1]。

伴随着性能的提升,卷积神经网络的参数存储量和计算开销也越来越大,在低功耗、计算和存储资源有限的在轨场景中难以移植应用。因此,如何在不降低网络性能的条件下,缩小网络规模,减少计算消耗,成为当前亟待解决的问题。如果能将网络规模小型化,那么将卷积神经网络部署到机载、星载等遥感设备便成为可能,这无疑是从实验室走向应用的关键性进展。

通过模型压缩来降低存储量和计算成本已经有了大量的研究。目前提出的网络压缩方法主要可分为以下几类[2]:参数剪枝、权值量化、低秩分解、紧性卷积核设计和知识蒸馏。参数剪枝法通过对网络进行结构或非结构化的剪枝操作,将模型中冗余的参数进行裁剪,从而降低网络复杂度,并可在一定程度上防止过拟合。20 世纪末,CUN 等[3]和 HASSIBI 等[4]先后提出最优脑损伤(Optimal Brain Damage, OBD)和最优脑外科(Optimal Brain Surgeon, OBS)的参数剪枝方法,利用网络的二阶导数信息衡量参数的冗余性,从而对网络进行修剪。2015 年,韩松等[5]提出了一种基于低权值连接的剪枝方法,对网络中权值较小的参数进行修剪,将原始的稠密网络变得稀疏。但这种非结构化的剪枝方式不会带来显著的加速,LI 等[6]、LEBEDEV 等[7]通过对冗余滤波器进行裁剪,研究了结构化的参数剪枝方法。2017 年,LIU 等[8]提出了一种网络瘦身(Network Slimming)方法,利用 BN 层的缩放因子,结合正则化稀疏约束对网络中不重要的通道进行裁剪。权值量化的主要思想是用较低的比特位表示模型实际的浮点型权值,通过减少网络参数的存储量,实现模型加速和压缩。GUPTA 等[9]使用 16 位定点数表示 CNN 训练过程中的参数,不仅显著减少了内存使用和浮点运算量,且几乎对分类精度没有损失。

为了进一步压缩权重,二元[10]、三元[11]权值网络被提出,通过在训练过程中学习低比特权值或激活,大幅压缩网络规模。低秩分解法采用张量或矩阵分解技术,将原始卷积核拆解为多个小卷积核的乘积,从而降低网络运算量和参数规模。紧性卷积核设计法从网络结构出发,通过设计更精简的网络模块,如 SqueezeNet[12]、ShuffleNet[13]等,有效减小模型尺寸。2015 年,HINTON 等[14]提出了知识蒸馏的概念,从高性能复杂网络中学习有用信息来指导简单网络的训练,在保证性能相差不大的情况下实现模型压缩。

上述方法可不同程度地减小原始网络规模,但也带来了精度的下降。为了改善精度损伤的问题,本文提出一种基于知识蒸馏的剪枝压缩改进方法。该方法将参数剪枝与知识蒸馏技术相结合,先对网络进行混合参数剪枝以删除冗余信息,再用剪枝前网络指导剪枝后网络的再训练过程,以提升后者的性能。以遥感图像分类任务为例,在遥感数据集上对 VGG-16 网络进行压缩,在网络精度损伤不大的情况下,可实现 16~18 倍的压缩效果。

1 基于混合参数剪枝的卷积神经网络压缩方法

1.1 方法介绍

参数剪枝主要有两种途径:一种针对通道剪枝;一种针对权重剪枝。为了降低网络参数的冗余度,可将两种剪枝方法结合应用,其主要技术路线如图 1 所示[15]。

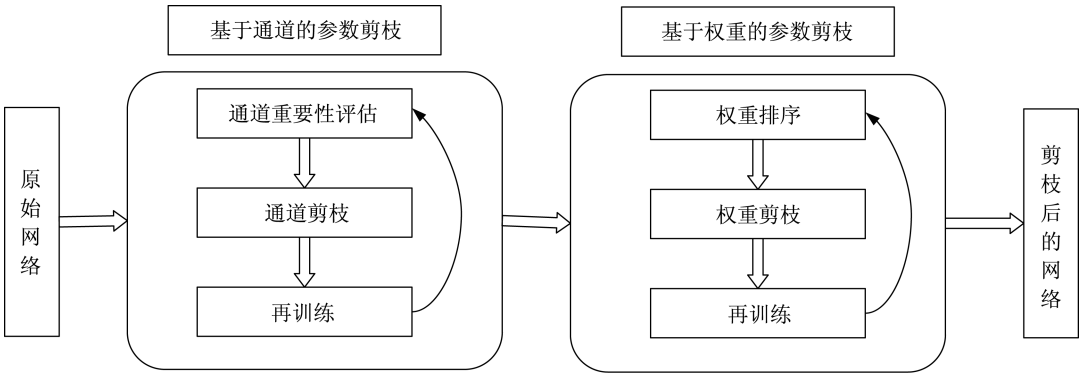


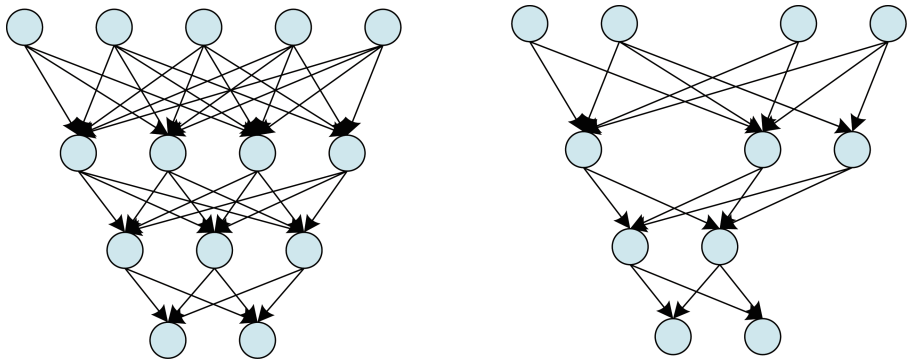
图 1 混合参数剪枝

Fig.1 Hybrid parameter pruning

首先对网络进行通道剪枝，实现模型的初步压缩，并对网络再训练以恢复损伤的精度。其次，在通道剪枝的基础上再进行权重剪枝，以进一步压缩网络的参数规模。

1.2　基于通道的参数剪枝

通道剪枝是一种结构化的参数剪枝方式，通过剪去网络中冗余的通道，即对网络贡献度较小的单个滤波器，实现对模型的压缩，如图 2 所示。



(a) 剪枝前　　(b) 剪枝后

图 2　通道剪枝前后

Fig.2　Before and after channel pruning

该方法可分为 3 个步骤：训练网络→裁剪冗余的通道→网络再训练。

如图 3 所示，$\boldsymbol{X}_i$ 为第 i 卷积层的输入特征图，$\boldsymbol{X}_{i+1}$ 为输出特征图，每个卷积层有 $n_i \times n_{i+1}$ 个二维的卷积核 $\boldsymbol{F}_{i,j}$，其中，n_i 和 n_{i+1} 分别表示输入、输出特征图的维度。

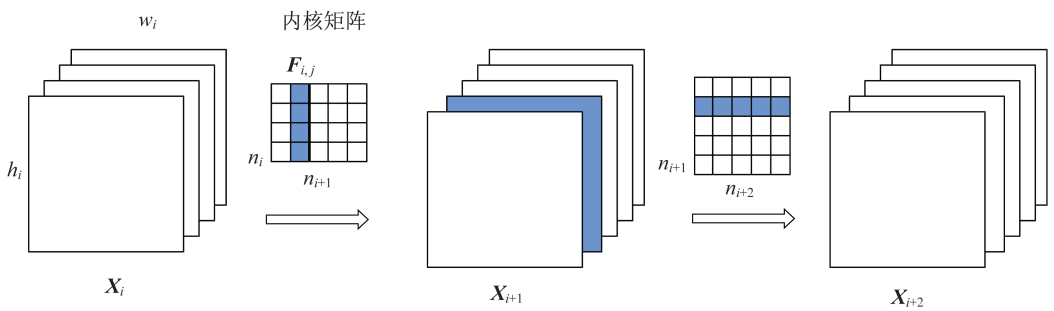


图 3　修剪一个滤波器会导致在下一层中删除其相应的特征图和相关的内核

Fig.3　Trimming a filter causes its corresponding feature map and associated kernel to be deleted in the next layer

首先评估通道的重要性，通过计算单个滤波器的 L_1 范数，即其绝对权重之和 $\sum\left|F_{i,j}\right|$ 来衡量每个层中滤波器的相对重要性。与层中其他滤波器相比，具有较小内核权重和的滤波器倾向于产生具有弱激活的特征映射。从第 i 个卷积层中修剪 m 个滤波器的步骤如下：

步骤 1　对于每个滤波器 $F_{i,j}$，计算其权重绝对值的总和 $s_j = \sum\left|F_{i,j}\right|$。

步骤 2　按 s_j 对滤波器进行排序。

步骤 3　剪去具有最小和值的 m 个滤波器及其对应的特征映射，下一个卷积层中与被修剪的特征映射相对应的内核也被删除。

步骤 4　第 i 层和第 $i+1$ 层创建新的内核矩阵，并将剩余的内核权重复制到新模型。

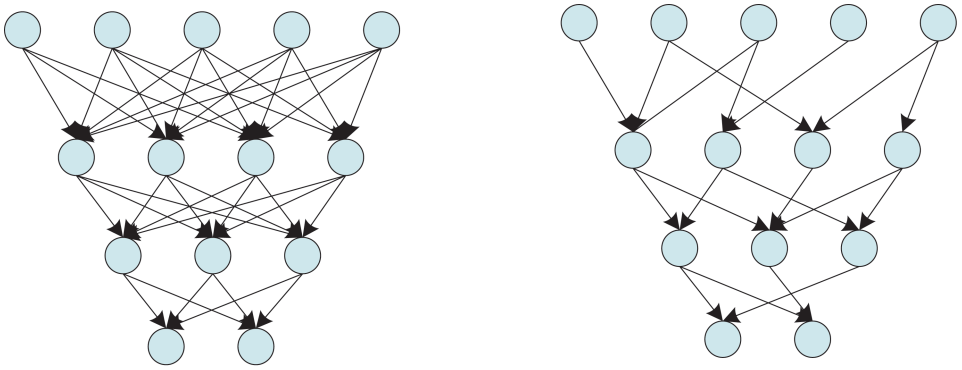
网络中不同卷积层对修剪的敏感度不同，通过逐层修剪并在测试集上评估剪枝网络的精度，可以判断各层对剪枝的敏感性，并确定每一卷积层通道剪枝的比例。

1.3　基于权重的参数剪枝

对网络进行通道剪枝后，得到较小规模的新网络，再训练使其恢复精度。将小规模新网络的参数

按照权值大小进行排序,权重越小代表其对网络的最终贡献越小。

通过删除权重小于指定阈值的连接,减少模型的参数量,如图4所示。



(a) 剪枝前　　(b) 剪枝后

图4　权重剪枝前后,网络由稠密变得稀疏

Fig.4　Before and after weight pruning, the network becomes sparse from dense

该方法仍然遵循“训练网络→剪枝→再训练”的3个步骤,剪枝的具体操作是将低于某一阈值的参数置零,再训练过程中已经被剪掉的参数将保持零值,不再进行参数更新。网络中不同层参数对剪枝的敏感度是不同的。由于网络的计算量主要集中在卷积层,通常卷积层比全连接层对参数剪枝更敏感。基于此,应仔细衡量每一层参数的剪枝比例,以获得最优的压缩效果。

2　基于知识蒸馏的剪枝压缩改进方法

由于上述混合参数剪枝方法会带来精度下降,因此本文提出了一种基于知识蒸馏的剪枝压缩改进方法,将知识蒸馏技术融合到混合参数剪枝中,通过再训练过程中的知识传递可以减少剪枝后的精度损失。具体步骤如图5所示。

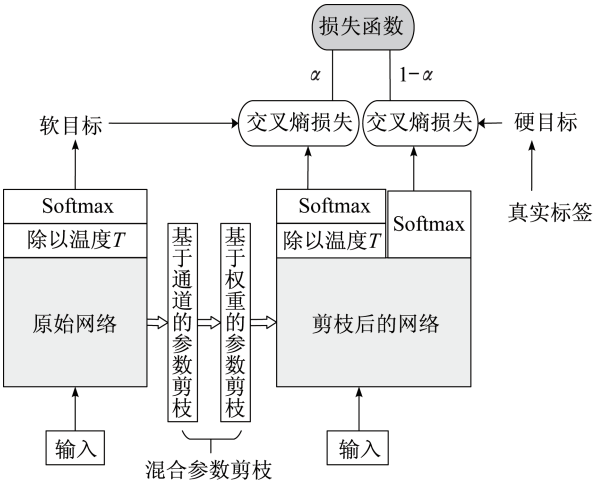


图5　基于知识蒸馏的剪枝压缩改进方法

Fig.5　Improved pruning compression method based on knowledge distillation

知识蒸馏法的主要思想是用大型复杂模型学习到的知识指导简单网络的训练,从而提升后者的表现。此时,复杂模型可以看作是“教师”,待训练的简单模型为“学生”。通过网络间的知识传递,使简单模型重现复杂模型的输出结果,达到规模压缩的目的。

神经网络学到的知识将输入数据映射为输出结果,对于处理分类任务的网络,其训练目标是将正确输出结果的概率最大化。但这种训练方式为每个错误结果也分配了概率,这些错误概率展示了网络进行分类判断的“思路”。神经网络通常使用Softmax输出层得到分类概率,函数定义如下:

$$q_i=\frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \tag{1}$$

式中:z_i为复杂模型分类结果的输出值;i为类别索引;j为类别总数;q_i为由Softmax函数计算得到的分类概率;T为温度,是一个可调节的超参数,通常为1,T值越大,类别的概率分布就越“软”,即更加均匀平和。

知识蒸馏法通过调整网络再训练过程中的损失函数,进行网络间的知识传递。训练小模型时,损失函数由两部分组成:第1项是包含“软目标”的

交叉熵函数，其目的是使小模型向经过蒸馏的复杂模型的输出结果优化；第 2 项是由小模型输出的结果与正确标签的差值计算得出的交叉熵函数，这与传统的训练过程相同，目的是使小模型向真实的标签值进行优化。损失函数的公式[16]为

$$L_{\mathrm{KD}}=\alpha T^{2}*L_{\mathrm{soft}}+(1-\alpha)*L_{\mathrm{hard}} \tag{2}$$

式中：L_{soft}、L_{hard}分别为优化方向为“软目标”和正确标签的交叉熵函数；α为损失函数的权重系数，$\alpha=0$时对应于不使用知识蒸馏法进行训练的情况。

以混合参数剪枝前的网络为“教师”，剪枝后的网络为“学生”，采用知识蒸馏法对网络进行再训练。相比于传统再训练方式，知识蒸馏法能获得更优的再训练效果，改善网络性能。

3　实验验证

3.1　基于混合参数剪枝的卷积神经网络压缩方法

以遥感图像分类任务为例，分别在两个公开的遥感数据集 NWPU-RESISC45 和 UC Merced Land-Use 上对 VGG-16 网络进行压缩实验。NWPU-RESISC45 数据集由西北工业大学研究团队于 2017 年提出，包含飞机、港口、教堂等 45 个不同的场景类别，每个场景中分别有 700 幅 256×256 的 RGB 图像，大部分测试图像的空间分辨率能达到每个像素 0.2 m。UC Merced Land-Use 数据集是由美国国家地质调查局航空拍摄的正射影像，包含棒球场、海滩、十字路口等 21 个不同的场景类别，每个场景包含 100 幅 256×256 的 RGB 图像，图像的空间分辨率为每个像素 0.3 m。实验时选取数据集中 20% 的图像作为训练集，剩余图像作为测试集，观察剪枝前后 VGG-16 网络的分类表现。

首先对 VGG-16 网络进行基于通道的参数剪枝，由于每一卷积层对剪枝的敏感性不同，需要先对单层修剪滤波器以判断每个卷积层对修剪的敏感度，从而设置恰当的通道剪枝比例。实验结果如图 6 和图 7 所示。由实验结果可知，在不同的数据集上，卷积层对通道剪枝的敏感性是相似的。第 7 卷积层对剪枝最敏感，第 1 卷积层对剪枝最不敏感。在本次实验中，对训练好的 VGG-16 网络的第 7 卷积层剪去 25% 的滤波器，对剩余卷积层均剪去 50% 的滤波器。

对通道剪枝后的网络进行再训练，使其恢复精度。接下来对网络进行基于权重的参数剪枝，以进一步减少冗余参数。首先对全连接层 90% 的参数进行剪枝，经过再训练，网络分类精度损伤不大，这说明剪掉的 90% 全连接层参数中大部分是冗余的。在此基础上，对卷积层参数进行剪枝，并逐渐增大卷积层参数的剪枝比例。在遥感数据集上对 VGG-16 网络进行混合参数剪枝，实验结果见表 1。

由表 1 数据可知，对 VGG-16 网络进行混合剪枝，网络精度下降 1% 左右时，能移除 93% 左右的

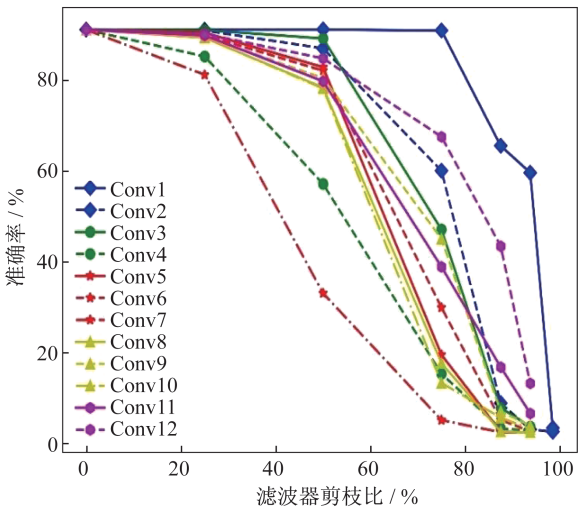


图 6　在 NWPU-RESISC45 数据集上，不同卷积层对通道剪枝的敏感性

Fig.6　Sensitivity of different convolutional layers to channel pruning on the NWPU-RESISC45 dataset

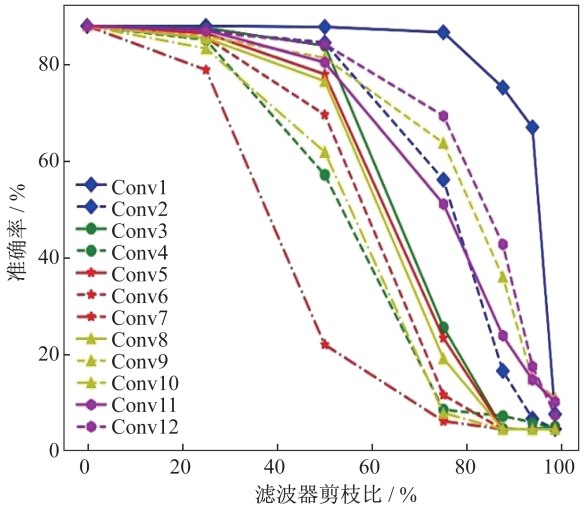


图 7　在 UC Merced Land-Use 数据集上，不同卷积层对通道剪枝的敏感性

Fig.7　Sensitivity of different convolutional layers to channel pruning on the UC Merced Land-Use dataset

表 1 混合参数剪枝前后网络对比

Tab.1 Network comparison before and after hybrid parameter pruning

遥感数据集	剪枝前分类精度	混合剪枝并再训练后分类精度	参数总下降比例/%
NWPU-RESISC45	89.69	88.79 (↓0.9)	92.79
		88.67 (↓1.02)	93.36
		88.53 (↓1.16)	93.92
		87.30 (↓2.39)	94.48
UC Merced Land-Use	87.86	87.08 (↓0.78)	92.47
		87.68 (↓0.18)	93.13
		86.25 (↓1.61)	93.78
		82.92 (↓4.94)	94.44

注:“↓”表示相对于原始网络,剪枝并再训练后网络分类精度的下降比例。

参数。这说明通过合理调整卷积层通道剪枝比例和通道剪枝后网络的权重剪枝比例,可以裁剪掉网络中的大部分冗余参数。

3.2 基于知识蒸馏的剪枝压缩改进方法

在 NWPU-RESISC45 数据集上,参数剪枝前 VGG-16 网络的分类精度为 89.69%。经混合参数剪枝剪去网络中 93.92% 的参数后,按普通方式对网络进行再训练,最终分类精度为 88.53%,相比于剪枝前,性能下降了 1.16%。为了使再训练后网络性能损失不超过 1%,使用知识蒸馏方式学习剪枝前网络的有用信息。

通过调整 Softmax 函数中参数 α、T 的值,可获得不同的知识蒸馏效果。网络再训练后的分类精度见表 2。由表 2 可见,采用知识蒸馏法训练的网络比普通训练方式得到的网络性能更好,这说明原始网络的分类结果对小网络的训练起到了正向的指导作用,提升了小网络的表现。当 $\alpha=0.8$、$T=5$ 时,对剪枝后的 VGG-16 网络重新训练,其精度可达到 89.63%;相比于剪枝前,性能只下降了 0.06%。

如图 8 所示,绘制剪枝后网络再训练过程中的精度变化曲线,红线代表 $\alpha=0.8$、$T=5$ 时使用知识蒸馏方法训练,蓝线代表普通训练方式。由曲线趋势可见,在相同迭代次数时,知识蒸馏法训练的网络性能有微弱的优势,最终分类精度比普通训练方式提升了 1 个百分点。

表 2 在 NWPU-RESISC45 数据集上,剪枝 93.92% 的参数,知识蒸馏效果

Tab.2 Knowledge distillation effect on the NWPU-RESISC45 dataset, pruning 93.92% of the parameters

α、T 取值	分类精度/%
$\alpha=0$(不使用知识蒸馏)	88.53 (↓1.16)
$\alpha=0.3$, $T=5$	88.87 (↓0.82)
$\alpha=0.7$, $T=5$	89.48 (↓0.21)
$\alpha=0.8$, $T=5$	89.63 (↓0.06)

注:“↑”“↓”表示相对于原始网络,剪枝并再训练后网络分类精度的上升或下降比例。

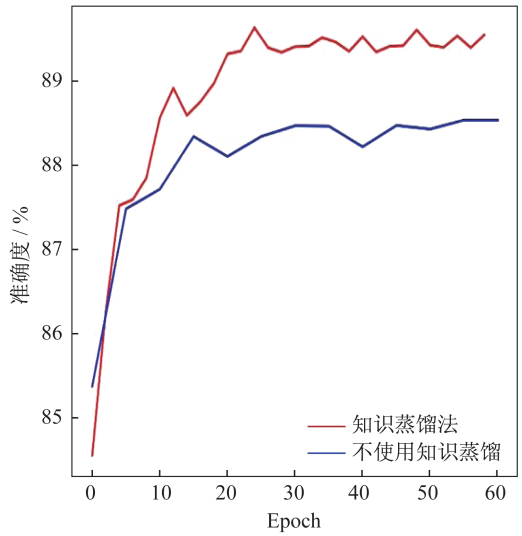


图 8 在 NWPU-RESISC45 数据集上,网络再训练过程中精度变化曲线

Fig.8 Accuracy curves during network retraining on the NWPU-RESISC45 dataset

在 UC Merced Land-Use 数据集上,参数剪枝前 VGG-16 网络的分类精度为 87.86%。经混合剪枝,剪去网络中 93.78% 的参数后,网络性能下降,经普通方式再训练精度可恢复到 86.25%,相比于剪枝前,精度下降了 1.61%。

采用知识蒸馏方式进行对剪枝后的网络进行再训练,训练结果见表 3。

由实验结果可知,采用知识蒸馏方法再训练的网络有了更好的性能。当 $\alpha=0.7$、$T=8$ 时,重新训练后网络的最佳分类精度可达到 87.86%,与剪枝前网络的精度相同。当 $\alpha=0.7$、$T=5$ 时,再训练后网络的最佳精度达到 88.21%,优于剪枝前的性能。这

表 3　在 UC Merced Land-Use 数据集上，剪枝 93.78% 的参数，知识蒸馏效果

Tab.3　Knowledge distillation effect on the UC Merced Land-Use dataset, pruning 93.78% of the parameters

α、T 取值	分类精度/%
$\alpha=0$（不使用知识蒸馏）	86.25（↓1.61）
$\alpha=0.5, T=5$	87.32（↓0.54）
$\alpha=0.7, T=8$	87.86
$\alpha=0.7, T=5$	88.21（↑0.35）

注：“↑”“↓”表示相对于原始网络，剪枝并再训练后网络分类精度的上升或下降比例。

表明剪枝后的网络接收到了原始模型学到的“知识”，在“教师”网络和有标签数据集的共同指导下，呈现出了更好的表现。

如图 9 所示，绘制在 UC Merced Land-Use 数据集上，剪枝后网络再训练过程中的精度变化曲线。图 9 中，红线代表 $\alpha=0.7$、$T=5$ 时使用知识蒸馏方法训练，蓝线代表普通训练方式。由图可知，在相同迭代次数时，知识蒸馏法训练的网络性能优于普通训练方式，在训练 20 个 epoch 后，知识蒸馏法训练的网络就达到了最高精度，所需迭代次数更少。

由以上实验结果可知，在遥感数据集上对

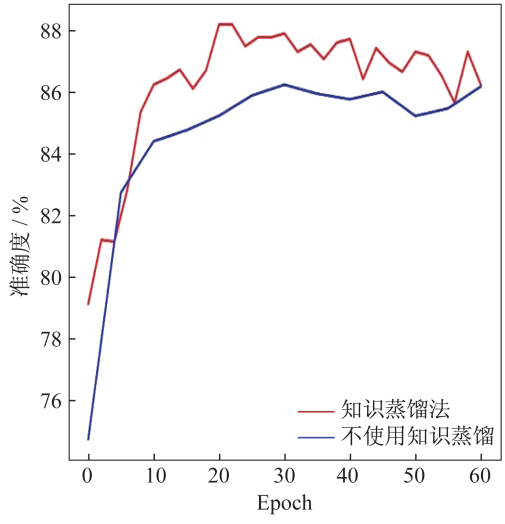


图 9　在 UC Merced Land-Use 数据集上，网络再训练过程中精度变化曲线

Fig.9　Accuracy curves during network retraining on the UC Merced Land-Use dataset

VGG-16 网络进行实验，基于知识蒸馏的剪枝压缩改进方法可在网络精度下降不到 1% 的条件下，移除 93%～94% 的参数量，实现 16～18 倍的压缩效果。在知识蒸馏过程中，当选择恰当的参数值时，剪枝后网络还可能获得优于原始网络的性能。

3.3　网络压缩前后性能对比

在 NWPU-RESISC45 数据集上训练 VGG-16 网络，比较网络压缩前后的性能变化，结果如图 10 所示。其中，Network Slimming 是 LIU 等[8]于 2017 年提出的基于正则化和 BN 层缩放因子的通道剪枝方法。取剪枝并再训练后网络精度下降不足 1% 的模型，对其参数量、浮点运算数（FLOPs）、内存读写开销三个维度进行比较。可以看出经剪枝后，Network Slimming 方法的 FLOPs 显著下降，而本文方法在减少参数量和内存读写开销上更有优势。这是由于网络运算量主要集中在卷积层，参数量集中在全连接层。对比模型中，Network Slimming 减去了约 70% 的卷积层，而本文方法只减去不足 50% 的冗余通道。通过多次进行剪枝和微调，本文方法有望进一步减少冗余通道数和 FLOPs。

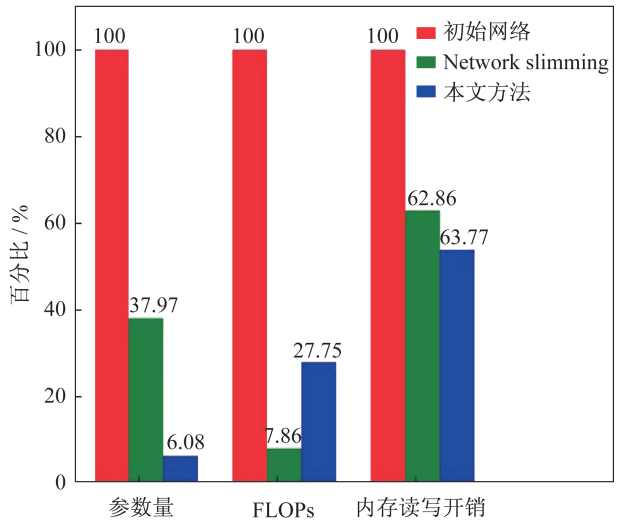


图 10　网络压缩前后性能变化

Fig.10　Performance changes before and after network compression

4　结束语

本文针对卷积神经网络规模庞大、难以在轨应用的问题，对深度神经网络压缩方法进行研究，提出了一种基于知识蒸馏的剪枝压缩改进方法。在

遥感图像分类任务中,该方法对VGG-16网络实现了16～18倍的压缩效果,并带来了运算量、内存读写开销的下降,且精度损失不到1%。本文提出的方法可扩展到其他深度神经网络上,如残差网络、Inception网络等,但在进行混合参数剪枝时需要根据网络的结构调整剪枝策略,以获得理想的压缩效果。此外,由于参数剪枝引入了不规则稀疏性,在网络计算过程中不会带来显著的加速。因此,后续可进行稀疏网络运算加速方面的研究,以满足实时性要求。

参考文献

[1] 余若峰,杨威,付耀文.基于测量数据积累的弱目标检测前跟踪算法[J].上海航天,2020,37(5):56-66.

[2] 纪荣嵘,林绍辉,晁飞,等.深度神经网络压缩与加速综述[J].计算机研究与发展,2018,55(9):47-64.

[3] CUN Y L, DENKER J S, SOLLA S A. Optimal brain damage [J]. Neural Information Proceeding Systems, 1990, 2(279):598-605.

[4] HASSIBI B, STORK D G. Second order derivatives for network pruning: optimal brain surgeon [J]. Advances in Neural Information Processing Systems, 1993, 5:164-171.

[5] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network [C]// Advances in Neural Information Processing Systems. Cambridge, UK: MIT Press, 2015: 1135-1143.

[6] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets [EB/OL]. (2017-03-10) [2019-11-04]. https://arxiv.org/pdf/1608.08710.pdf.

[7] LEBEDEV V, LEMPITSKY V. Fast ConvNets using group-wise brain damage [C]// IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016:2554-2564.

[8] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming [C]// Proceeding of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017:2736-2744.

[9] GUPTA S, AGRAWAL A, GOPALAKRISHNAN K, et al. Deep learning with limited numerical precision [C]// International Conference on Machine Learning. New York, USA: ACM, 2015:1737-1746.

[10] COURBARIAUX M, BENGIO Y, DAVID J P. Binary connect: training deep neural networks with binary weights during propagations [C]// Advances in Neural Information Processing Systems. Cambridge, UK: MIT Press, 2015:3123-3131.

[11] LI F, ZHANG B, LIU B. Ternary weight networks [EB/OL]. (2016-11-19) [2019-11-04]. https://arxiv.org/pdf/1605.04711.pdf.

[12] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size [EB/OL]. (2016-04-06) [2019-11-04]. https://arxiv. org/pdf/1602.07360v3.pdf.

[13] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018:6848-6856.

[14] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [J]. Computer Science, 2015, 14(7):38-39.

[15] 靳丽蕾,杨文柱,王思乐,等.一种用于卷积神经网络压缩的混合剪枝方法[J].小型微型计算机系统,2018,39(12):38-43.

[16] LI H. Exploring knowledge distillation of deep neural networks for efficient hardware solutions [R]. CS230 Final Report, 2018.