doi:10.6041/j.issn.1000-1298.2023.04.027

基于动态剪枝神经网络的杂草检测算法研究

亢　洁　刘　港　王　勍　夏　宇　郭国法　刘文波
(陕西科技大学电气与控制工程学院, 西安 710021)

摘要: 针对卷积神经网络模型巨大的参数量和计算量导致其实际应用时难度较大的问题,提出了一种基于注意力机制与动态稀疏约束的模型压缩方法。该算法首先借助 SENet(Squeeze and excitation networks,SENet)模块(可称为 SE 模块)评估出网络中各个通道的重要性,并施加稀疏正则化;然后提出一种网络稀疏度的自适应惩罚权重设计方法,根据模型学习效果,动态调整权重,将其添加到最终的训练目标上,实现模型动态压缩。最后,通过实验验证所提出的模型压缩方法,在经典的多分类数据集 CIFAR-10 上进行实验,证明了本文所提出的基于注意力机制与动态稀疏约束的模型压缩方法可降低网络的冗余度,使网络模型参数量减少 43.97%,计算量减少 82.94%,而分类准确率只比原始 VGG16 模型下降 0.04 个百分点。随后又将提出的模型压缩方法应用到杂草检测任务中,在甜菜与杂草数据集上进行实验,实验结果表明,剪枝模型相较于未剪枝模型的模型参数量减少 41.26%,计算量减少 45.77%,而平均检测精度均值只减少 0.91 个百分点,证明了该方法在杂草检测方面效果较好。

关键词: 杂草检测; 模型压缩; 注意力机制; 动态稀疏约束
中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2023)04-0268-08　　**OSID**:

Weed Detection Algorithm Based on Dynamic Pruning Neural Network

KANG Jie　LIU Gang　WANG Qing　XIA Yu　GUO Guofa　LIU Wenbo
(*School of Electrical and Control Engineering*, *Shaanxi University of Science and Technology*, *Xi'an* 710021, *China*)

Abstract: To address the problem that the convolutional neural network models are difficult to be applied in practice due to their vast number of parameters and computation, a model compression method based on attention mechanism and dynamic sparse constraint was proposed. Firstly, the importance of each channel in the network was evaluated with the help of the squeeze and excitation networks (SENet) module, and sparse regularization was applied; then an adaptive penalty weight design method for network sparsity was proposed. According to the learning effect of the model, the weight was dynamically adjusted and added to the final training target to realize the dynamic compression of the model. Finally, the proposed model compression method was verified by experiments on the classic multi-classification dataset CIFAR-10. It was proved that the proposed model compression method based on attention mechanism and dynamic sparse constraint can reduce the network redundancy, resulting in a 43.97% reduction in the amount of network model parameters and an 82.94% reduction in the amount of computation, while the classification accuracy was only 0.04 percentage points lower than that of the original VGG16 model. Then the proposed model compression method was applied to the weed detection task, and the experiment was carried out on the sugar beet and weed datasets. The experimental results showed that compared with the unpruned model, the pruned model reduced the model parameters by 41.26%, the calculation amount by 45.77%, and the average detection accuracy by only 0.91 percentage points, which proved that this method could also have a good effect on the weed detection task.
Key words: weed detection; model compression; attention mechanism; dynamic sparse constraint

0　引言

杂草与作物类似,它的成长也需要阳光和水资源等。如果田间的杂草密度太大,会严重影响农作物的产量和质量[1-3]。因此,在农业生产活动中,杂草控制具有十分重要的现实意义[4]。发展精准农

收稿日期: 2022-07-01　修回日期: 2022-12-22
基金项目: 陕西省重点研发计划项目(2021GY-022)和西安市科技计划项目(2019216514GXRC001CG002-GXYD1.7)
作者简介: 亢洁(1973—),女,副教授,博士,主要从事机器学习和模式识别研究,E-mail: kangjie@sust.edu.cn

业已成为现代农业的主流方向，杂草检测是精准防控杂草的前提和基础，是有效控制杂草的关键。基于杂草检测的农药精准变量喷洒研究已成为减少农药浪费、提高农药利用率的迫切需要[5]。

近年来，深度学习技术不断发展，在各种计算机任务中显示出强大的能力，越来越多优秀的模型被提出。大多数模型都在往更深更宽的方向发展，这样可以保证模型的性能，准确率较高。但也导致模型具有巨大的参数量和计算量，使得模型不易应用到实际中。虽然现在已经有如图像处理单元（Gaphic processing unit，GPU）或者神经网络处理单元（Nural network processing unit，NPU）等加速单元，但把巨大的模型移植到嵌入式设备或者移动端设备上，如可穿戴设备等，是难以实现的，无法满足商用需求。卷积神经网络的应用具有极大的时空限制，因此，研究者们通过模型压缩方法来降低网络冗余度，提出了许多解决上述问题的方案，包括低秩分解[6-9]、权重量化[10-12]、权重剪枝[13-14]、神经结构学习[15-17]和结构化剪枝等。其中结构化剪枝通过删除网络中的整个平面，以及它们连接的卷积核，计算成本可大大降低，且不会导致稀疏格式的矩阵。由于结构化剪枝方法不需要稀疏卷积库的支持，可以直接在现有的深度学习框架上运行，因此该方法受到广大学者的关注。

在结构化剪枝方法中[18-20]，通常采用一些剪枝策略对已经训练好的模型进行剪枝，从而得到紧凑的网络模型。LIU 等[21]提出了一种网络瘦身的方法来学习更加紧凑的网络模型，直接对批量归一化层（Batch normalization，BN）中的缩放因子施加稀疏正则化，用该缩放因子的值近似于通道对网络模型的贡献值，因此在训练过程中识别出不重要的通道后，该通道即可被修剪。HE 等[22]提出了一种软滤波修剪（Soft filter pruning，SFP）方法来加速深度卷积神经网络的推理过程。在训练过程中，所提出的 SFP 方法允许贡献值小的过滤器在之后的训练迭代过程中被更新，这种方法可以保持模型的容量，从而实现卓越的性能。GAO 等[23]提出了特征增强与抑制（Feature boosting and suppression，FBS），该方法在现有的卷积层中引入小的辅助连接，可以预测通道的显著性，并跳过那些对分类结果贡献较小的通道。LI 等[24]通过计算卷积滤波器权值的 L1 范数并排序，在网络训练过程中剪枝掉较小输入权重的滤波器，然后微调网络以恢复准确性。MOLCHANOV 等[25]基于权重泰勒级数展开的评价方式估计滤波器对网络的最终输出贡献，并丢弃评价分数较小的滤波器。LI 等[26]发现剪枝后不同的子网络与最终网络模型的精度关系密切，提出了一种基于自适应批归一化技术的评价方式，以快速找到具有高精度的子网络。

上述研究对通道的重要性评估过程较复杂，且通常对模型使用相同强度的稀疏约束，而不探究模型学习能力与模型冗余度之间的关系。为解决上述问题，本文提出一种基于注意力机制和动态稀疏约束的模型压缩方法，旨在提供一个简单的方法来实现卷积神经网络模型的通道级稀疏性。该方法首先借助 SE 模块评估出网络模型中各个通道的重要性，再根据模型目前的学习效果设计出一种动态稀疏约束，将其添加到最终的训练目标上，实现模型动态压缩。最后在甜菜和杂草数据集上进行实验，以验证提出的方法在目标检测任务中的有效性。

1　基于注意力机制与动态稀疏约束的模型压缩方法

1.1　基于注意力机制的模型压缩算法

随着深度学习不断发展，注意力机制被广泛研究和使用，深度学习中的注意力机制与人类特有的视觉注意力机制非常类似，可以从大量的信息中选择感兴趣的关键信息。其中，HU 等[27]提出了通道注意力机制——SE 模块。SE 模块可以逐通道重新标定通道的权重，显式地建立通道之间的相互关系。SE 模块使用起来也非常方便，在现有的卷积神经网络模型中，可以很方便地嵌入其中，即插即用。本文采用 SE 模块实现图像通道级的权重标定，进而实现模型通道剪枝。SE 模块的基本结构如图 1 所示。

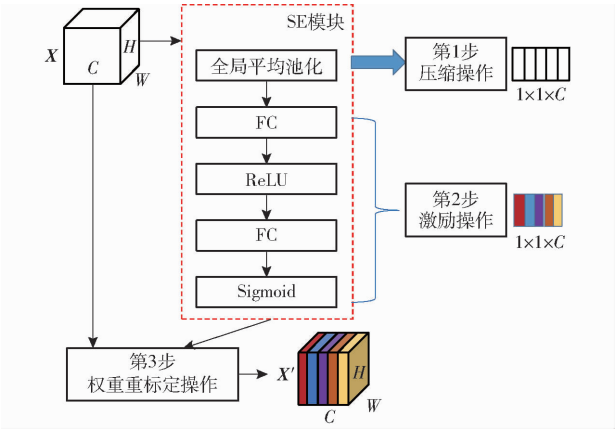


图 1　SE 模块结构图

Fig. 1　SE module structure diagram

图 1 中，$\boldsymbol{X}$ 和 $\boldsymbol{X}'$ 分别为输入特征图和输出特征图，C、H、W 分别为输入特征图的通道数、高和宽。SE 模块能够自动地了解各个通道的重要性，整个操作分为 3 步：

第 1 步为压缩（Squeeze）操作，此步骤的核心是

全局平均池化操作，将特征图 $\boldsymbol{X}$ 经全局平均池化后，把每个通道特征图的所有像素值相加，再求平均，用该平均值表示对应通道的特征图，则该数值具有全局的感受野，感受区域更广，此时特征图的尺寸变为 $1\times1\times C$。压缩操作可表示为

$$\boldsymbol{Z}=F_{sq}(\boldsymbol{X})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X(i,j) \tag{1}$$

式中 $\boldsymbol{Z}$——经第 1 步压缩操作后的输出特征

第 2 步激励（Excitation）操作，包括全连接层（FC）、激活函数（ReLU）和 Sigmoid 归一化，可为每个特征通道生成权重，此时特征图为尺寸 $1\times1\times C$ 的带有不同权重值的特征图，具体表达式为

$$\boldsymbol{S}=F_{ex}(\boldsymbol{Z},\boldsymbol{\omega})=\sigma(\boldsymbol{\omega}_2\delta(\boldsymbol{\omega}_1\boldsymbol{Z})) \tag{2}$$

式中 $\boldsymbol{S}$——经第 2 步激励操作后的输出特征

δ——ReLU 函数

$\boldsymbol{\omega}$——可训练权重

σ——Sigmoid 函数

$\boldsymbol{\omega}_1$——ReLU 函数的可训练权重

$\boldsymbol{\omega}_2$——Sigmoid 函数的可训练权重

从图 1 中可以看出，经过第 2 步激励操作后，每个通道具有不同的权重值，这主要归因于 ReLU 函数。根据 ReLU 函数的定义，当输入在横轴的负半轴时，ReLU 函数的输出结果恒为 0；当输入在横轴的正半轴时，函数值不变，为它本身。因此，经过 ReLU 激活函数后，此时特征图为带有不同权重的特征图。第 2 步中的 Sigmoid 函数主要起到归一化的作用，因为 Sigmoid 函数的取值范围为（0，1），当使用此函数后，特征图权重范围变为（0，1），得到归一化权重。

第 3 步为权重重标定（Reweight）操作，它是利用乘法运算，按通道把所获得的权重加权到原始特征上，从而在通道维度上实现重标定原始特征的权重，得到与原特征图尺寸相同，经通道信息强度重标定的特征图 $\boldsymbol{X}'$。

$$\boldsymbol{X}'=\boldsymbol{F}_{\text{reweight}}(\boldsymbol{X},\boldsymbol{S})=\boldsymbol{S}\boldsymbol{X} \tag{3}$$

经过 SE 模块，在通道维度上得到了具有不同权重的特征图。即完成了通道剪枝的第 1 步，评估出不同通道的重要性程度。

此时，每个通道具有不同的权重，用此权重表示各个通道对网络的贡献值，并对这些通道权重施加稀疏正则化。本文方法训练目标为

$$L=\sum l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y})+\lambda\sum g(\boldsymbol{S}) \tag{4}$$

其中 $$g(\boldsymbol{S})=\|\boldsymbol{S}\|_1$$

式中 $\boldsymbol{x}$——训练输入 $\boldsymbol{y}$——训练目标

$l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y})$——与任务相关的损失函数

λ——权重系数

$g(\boldsymbol{S})$——导致卷积滤波器通道稀疏性的 L1 范数，被广泛用于实现稀疏性

本文中某一通道的权重近似于该通道对网络的贡献值，因此对经 SE 模块的带有不同权重的输出施加稀疏正则化，使某些权重趋于 0，并随着模型的训练，权重不断发生变化，最终某些通道对应的权重将会变为 0，这些通道就可被修剪掉；权重系数 λ 用来平衡这 2 个损失，较大的 λ 会导致更稀疏的卷积核，从而获得更紧凑的网络。

1.2 基于动态稀疏约束的模型压缩算法

在式（4）中，权重系数 λ 对于确定稀疏性惩罚对通道显著性的强度至关重要，λ 越大，对网络的稀疏约束越大，即趋于 0 的值越多。然而，多数研究都使用相同的权重来训练网络，但如果当前网络未能很好地拟合输入的实例，则需要更大的网络容量来追求预测精度，而不是进一步推动稀疏性。因此，控制网络复杂度的稀疏性惩罚的权重应该随着损失的降低而增加，反之亦然。在极端情况下，对于那些未拟合的实例，不应对相应的网络提供稀疏性约束。因此，优化目标应分情况表述。图 2 为目标函数是否添加稀疏正则化的判别条件流程图。

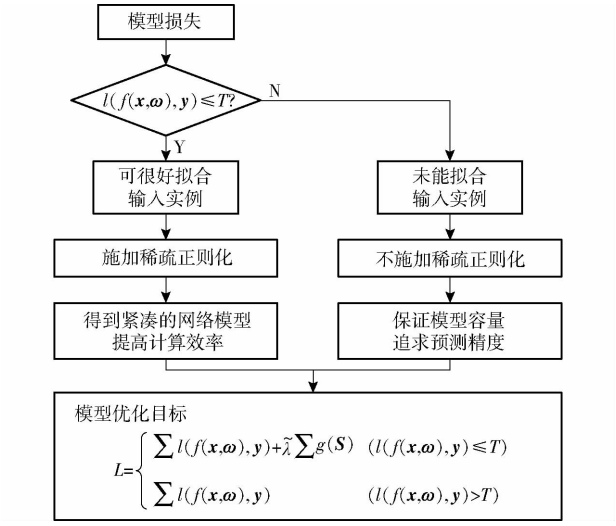


图 2 模型目标函数添加稀疏正则化的判别条件流程图

Fig. 2 Flowchart of the discriminant condition for adding sparse regularization to the model objective function

在本文中，通道显著性 $g(\boldsymbol{S})$ 的稀疏性决定了其中有效滤波器的数量，而稀疏的 $g(\boldsymbol{S})$ 会导致网络更紧凑，复杂度更低。因此，使用通道显著性的稀疏性作为网络复杂性的度量。

通过预定义阈值 T 来判断是否需要对网络实施稀疏性约束，本文中 T 的取值为 0.5。若损失大于 T，意味着此时损失较大，当前实例没有很好地拟合，则需要一个具有更强表示能力的网络来提取信息，不应施加稀疏性约束，因此优化目标为

$$L = \sum l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y}) \tag{5}$$

若损失值小于等于 T，优化目标为

$$L = \sum l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y}) + \widetilde{\lambda} \sum g(\boldsymbol{S}) \tag{6}$$

其中，$\widetilde{\lambda} = \lambda'(1 - l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y}))$，$\lambda'$是所有实例共享的超参数，用于平衡分类精度和网络稀疏性。因为控制网络复杂度的稀疏性惩罚的权重应该随着损失的降低而增加，符合减函数的特点，则在此处用 $1 - l(f(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{y})$ 表示。图 3 为动态稀疏约束曲线。

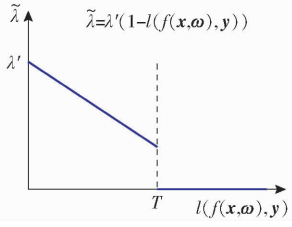


图 3　动态稀疏约束

Fig. 3　Dynamic sparse constraint

2　实验

2.1　实验设计

本文实验操作平台为 Ubuntu 16.04 系统，采用 PyTorch 深度学习框架，CPU 型号为 Intel(R) Xeon(R)E5 - 2678 v3@2.5 GHz，显卡(GPU)的型号为 NVIDIA GeForce RTX 2080 Ti，显卡内存 11 GB，编程语言为 Python。

在整个模型训练的过程中，批处理大小设置为 64。设置实验初始学习率为 0.1，学习率以几何退火的方式从 10^{-5} ~ 10^{-2} 自动调整，采用 Adam(Adaptive moment estimation)优化器更新网络参数，学习动量为 0.9，权重衰减率为 0.000 5。

2.2　评价指标

采用准确率(Accuracy)、参数量(Parameters)和计算量(FLOPs)对模型压缩算法进行评估。准确率是被正确分类的样本数占总样本数的比值，该值越大说明网络准确率越高。参数量可理解为空间复杂度，它决定了显存的使用量。参数量决定了网络规模。计算量可理解为时间复杂度，可用来衡量网络模型的复杂度。计算量决定了网络运行时间。

2.3　数据集及预处理

本文实验采用 CIFAR - 10 数据集和杂草数据集。CIFAR - 10 数据集是一个用于识别普适物体的小型开源数据集，CIFAR - 10 数据集中有 10 类目标物，这也是该数据集命名的由来，在图 4 中列举了这 10 类目标。数据集中图像尺寸为 32 像素 × 32 像素，训练集图像数量为 50 000 幅，测试集图像数量为 10 000 幅。CIFAR - 10 数据集的特别之处在于将识别目标转移到了普适对象上，且该数据集被应用于多分类任务，此前的数据集类别数很少。在 10 类目标物中，分别随机抽取了 10 幅图像展示在图 4 中。

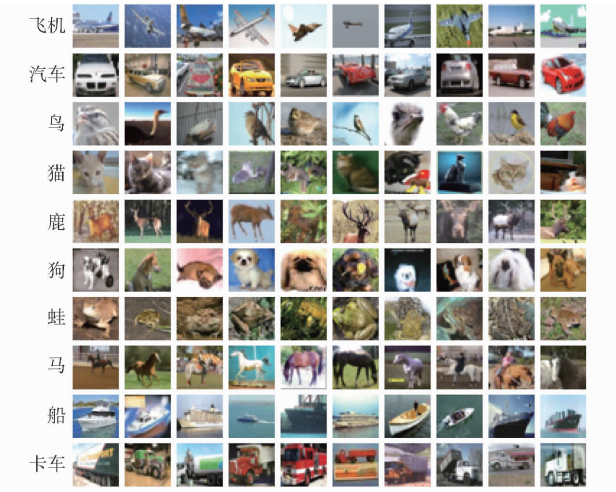


图 4　CIFAR - 10 数据集示意图

Fig. 4　Schematic of CIFAR - 10 dataset

本文实验的第 2 个数据集是由 CHEBROLU 等[28]制作的公开作物杂草数据集，选择自然环境下甜菜及其伴生杂草为实验对象，共包含 1 930 幅图像，该样本数据由田间农业机器人采集得到，从晴天干燥环境到阴天潮湿环境捕捉了不同的天气和土壤条件下的甜菜及杂草，能较好地反映自然环境下甜菜及伴生杂草的真实特点，下载地址为 http://www.ipb.uni-bonn.de/data/sugarbeets2016/。对数据集利用 LabelImg 图像标注工具进行标注，并通过颜色抖动、随机噪声和翻转的数据增强方法对下载的数据集进行扩充，增强模型的泛化能力，处理后部分图像如图 5 所示。扩充后数据集数量为 3 321 幅，并将原始图像压缩为 300 像素 × 300 像素的图像，作为训练模型的输入。

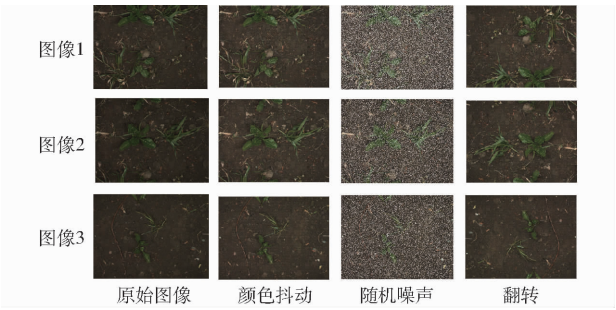


图 5　数据增强后的图像

Fig. 5　Image after data enhancement

2.4　实验结果与分析

2.4.1　图像分类任务

模型压缩方法多用于图像分类任务，因此首先将本文所提模型压缩算法部署在图像分类任务中，

使用 VGG16 模型(Baseline)来验证本文所提方法的有效性,将本文所提方法与 3 个经典的模型压缩方法进行了对比实验,表 1 为 4 种方法在 CIFAR - 10 数据集上的结果。分别对比 4 种方法在数据集上的准确率、与未做模型压缩的基准模型的参数量和计算量变化率(负值表示减少)。

表 1 VGG16 模型在 CIFAR - 10 上的剪枝结果

Tab. 1 Pruning results of VGG16 model on CIFAR - 10

模型	方法	动态剪枝	动态稀疏约束	准确率/%	参数量变化率/%	计算量变化率/%
VGG16	Baseline	-	-	91.64	0	0
	Slimming[21]	×	×	91.35	-36.27	-33.29
	SFP[22]	√	×	90.45	-35.17	-39.24
	FBS[23]	√	×	91.10	-37.23	-43.61
	本文方法	√	√	91.60	-43.97	-82.94

注:√表示添加此方法;×表示未添加此方法;-表示不使用剪枝方法。

由表 1 看出,只有本文方法添加了动态稀疏约束,其余 3 种方法没有此约束,其中 Slimming 算法是静态剪枝,其余 3 种方法是动态剪枝。在 CIFAR - 10 数据集上,VGG16 模型剪枝结果见表 1。未剪枝情况下,VGG16 模型的分类准确率为 91.64%,参数量为 $1.383\,6\times10^8$,计算量为 $1.548\,386\times10^{10}$,当使用文献[21 - 23]的剪枝方法时,分类准确率分别为 91.35%、90.45% 和 91.10%,模型参数量分别减少 36.27%、35.17% 和 37.23%,计算量分别减少 33.29%、39.24% 和 43.61%,而本文方法使模型参数量减少 43.97%,计算量减少 82.94%,可见,本文方法减少的模型参数量和计算量最多,而分类准确率只比原始 VGG16 模型下降 0.04 个百分点。由此可以推断,本文方法可以充分挖掘网络的冗余,从而得到紧凑但功能强大的高性能网络。

2.4.2 图像检测任务

如上所述证明了本文方法在图像分类任务中是有效的。为了进一步分析本文方法的通用性,将其部署在图像检测任务中。文献[29]主要提出了一种基于多尺度融合模块和特征增强的杂草检测方法,该方法对小目标作物和杂草、叶片交叠情况具有较好的检测效果,可实现作物与杂草的快速准确检测,能够为精准农业的发展提供支持。

将本文所提模型压缩方法应用在文献[29]的模型中,以此来得到更紧凑的杂草检测模型,数据集使用 2.3 节提到的杂草数据集进行实验。

文献[29]所提的模型使用 MobileNet 作为主干网络,并在 MobileNet 网络中添加了多尺度融合模块,以使其对小目标检测有更好的性能,其中,多尺度融合模块包括 SE 模块和扩张卷积。因此,首先分别统计 MobileNet 网络、MobileNet + SE 和 MobileNet + SE + 扩张卷积(MobileNet + SE + DC)的参数量,分别为 3 209 026、3 372 866 和 8 754 498。可以看出,扩张卷积的添加使参数量增加了很多。因此,在通道剪枝实验中,首先分析应该把通道剪枝算法添加到网络的哪一层。

将本文模型压缩算法应用在网络的各层,每层通道数对比如图 6 所示。可以看出,网络模型的前 5 层可剪枝通道数很少,说明浅层的网络模型冗余度不高。随着网络模型层数的增加,可剪枝通道数变多。若为网络的每层都添加通道剪枝算法,浅层剪枝数很少,对模型最终结果影响也不是很大,却增加了计算量。因此,综上分析,只给网络模型的后 6 层,即 6 ~ 11 层添加设计的通道剪枝算法。

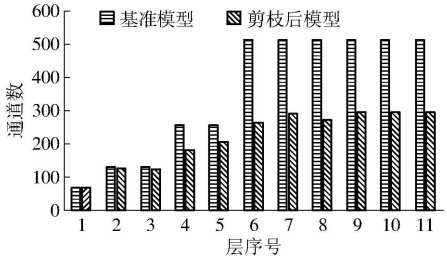


图 6 网络模型每层通道数对比

Fig. 6 Comparison of number of channels per layer of network model

经过实验,具体每层初始参数量、剪枝后参数量和通道数目变化率(负值表示减少)如表 2 所示。

表 2 网络模型通道剪枝结果

Tab. 2 Network model channel pruning details

层序号	初始参数量	剪枝后参数量	参数量变化率/%	通道数变化率/%
6	134 912	70 670	-47.62	-48.63
7	268 800	79 478	-70.43	-43.55
8	268 800	82 622	-69.26	-46.68
9	268 800	83 578	-68.91	-42.77
10	268 800	87 593	-67.41	-44.14
11	268 800	69 962	-73.97	-54.69

由表 2 可以看出,可以剪枝的网络模型通道数占比较多,这些通道对最终的网络模型贡献很小,通过本文方法,可以使网络模型变得更加紧凑。

图 7 为 SSD 原始模型(SSD - VGG)、文献[29]中基于多尺度融合模块和特征增强的模型(文献[29]模型)、通道剪枝模型在甜菜与杂草数据集上的训练损失曲线,由图 7 可以看出,在模型训练过程中,随着训练的进行,模型损失值不断降低直至最后收敛,本文所提模型的损失值和通道剪枝模型损失值均小于标准 SSD 模型。

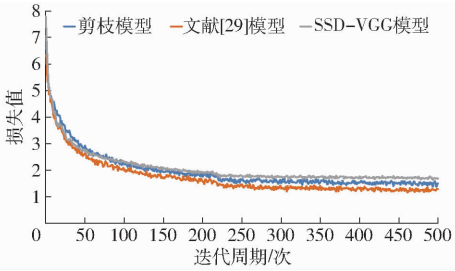


图 7　未剪枝模型与剪枝模型的训练损失曲线

Fig. 7　Training loss curves of unpruned model and pruned model

统计未剪枝模型和剪枝模型的平均检测精度均值(mAP)和模型参数量、计算量,如表 3 所示。

表 3　未剪枝模型与剪枝模型在甜菜与杂草数据集上的检测结果

Tab. 3　Detection results of unpruned and pruned models on sugar beet and weed dataset

方法	mAP/%	参数量	计算量(GFLOPs)
文献[29]模型	88.84	8 754 498	2.84
剪枝模型	87.93	5 142 737	1.54

由表 3 可知,剪枝模型有较好的结果。本文方法平均检测精度均值(mAP)虽低于基线 0.91 个百分点,但推理速度比基线快,模型参数量减少 41.26%,计算量减少 45.77%。表明本文方法对目标检测任务也具有很好的泛化能力。

图 8 和图 9 直观展示了模型经压缩后的作物与杂草图像的检测结果,共展示了 8 幅图像。图 8 中将 1 幅图像中的杂草与作物都检出,数据集中大部分图像的检测结果也如此。

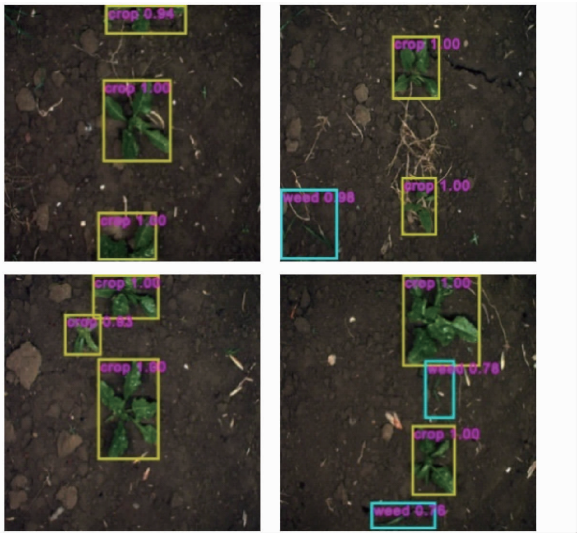


图 8　剪枝模型准确检测结果

Fig. 8　Accurate detection results of pruning model

检测结果中也存在一些未检测到作物或杂草的情况,图 9 展示了未剪枝模型和剪枝模型漏检的检测结果,用红色的虚线框框出了未检测出的植株。

(a) 未剪枝模型

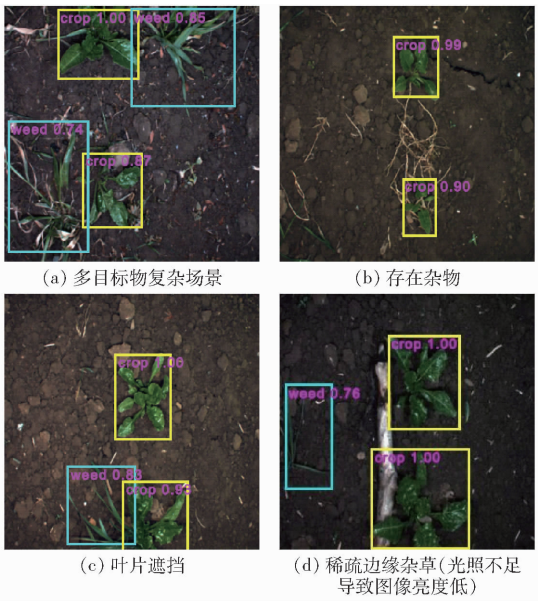
(b) 剪枝模型

图 9　漏检现象结果

Fig. 9　Results of missed inspection

从图 9 中可以看出,对于图中左侧的图像,未剪枝模型漏检了图中的一株作物,但准确检测出了杂草,而剪枝模型将杂草与作物检测为 1 株杂草,未能准确检测。对于图中右侧的图像,未剪枝模型漏检了 1 个目标,而剪枝模型有 2 处漏检。从而也可以得出,剪枝后模型的精度略有减小,和表 3 得出的结论一致。分析可知,未剪枝模型与剪枝模型因植物间遮挡等原因存在漏检现象,当植物间互相遮挡部分较多时,未能实现准确检测,这也是本文后续工作的研究方向。总体而言,剪枝模型可以达到与未剪枝模型相当的检测效果,且剪枝模型大大减少了模型参数量、计算量。

图 10 展示了剪枝模型在多目标物复杂场景、存



(a) 多目标物复杂场景　(b) 存在杂物

(c) 叶片遮挡　(d) 稀疏边缘杂草(光照不足导致图像亮度低)

图 10　剪枝模型在样本较复杂情况下的检测结果

Fig. 10　Detection results of pruning model in case of more complex samples

在杂物、叶片遮挡及光照不足导致图像亮度低等情况下的检测结果。可以看出,图 10a 存在较多的目标物,且作物与杂草紧挨在一起,剪枝模型可以准确检测出图中的作物与杂草。图 10b 存在很明显的杂物,与检测目标混杂在一起,且与植物形态很相像,剪枝模型可准确检测出目标物,且检测精度高。图 10c 中杂草和作物的根系交错在一起,剪枝模型可较准确地识别出杂草和作物。图 10d 中的杂草处于图像的边缘,且稀疏,剪枝模型依然检测出了杂草。图 10 表明本文模型可准确预测目标物位置及类别信息,具有良好的泛化能力和鲁棒性。

3 结论

(1)针对模型参数量、计算量巨大的问题,提出了一种基于注意力机制与动态稀疏约束的模型压缩算法。借助 SE 模块对网络模型中的通道重要性进行评估,通过 SE 模块评估通道重要性的方法相比于其他研究中评估通道重要性的方法更加简单;提出了一种基于动态稀疏约束的模型压缩算法,根据目前模型学习效果,动态调整权重,将其添加到最终的训练目标上,实现模型动态压缩。

(2)在经典多分类数据集 CIFAR-10 上进行了实验,从实验结果可以得出,本文方法可以充分挖掘网络的冗余,在几乎不损失模型精度的情况下,使模型参数量减少 43.97%,计算量减少 82.94%。最后又将提出的模型压缩方法应用到文献[29]中提出的杂草检测模型中,实验结果表明,剪枝模型相较于未剪枝模型的模型参数量减少 41.26%,计算量减少 45.77%,而平均检测精度均值只减少 0.91 个百分点,证明了本文方法在图像检测任务中也具有很好的效果,且像 MobileNet 这样的轻量化网络模型也可以被进一步压缩,压缩后的网络模型依然具有较高的检测精度。

参 考 文 献

[1] ZHANG Y, STAAB E, SLAUGHTER D, et al. Automated weed control in organic row crops using hyperspectral species identification and thermal micro-dosing[J]. Crop Protection, 2012, 41: 96-105.

[2] TANG J, CHEN X, MIAO R, et al. Weed detection using image processing under different illumination for site-specific areas spraying[J]. Computers and Electronics in Agriculture, 2016, 122: 103-111.

[3] HARKER K N, O'DONOVAN J T. Recent weed control, weed management, and integrated weed management[J]. Weed Technology, 2013, 27(1): 1-11.

[4] 孟庆宽, 张漫, 杨晓霞, 等. 基于轻量卷积结合特征信息融合的玉米幼苗与杂草识别[J]. 农业机械学报, 2020, 51(12): 238-245, 303.
MENG Qingkuan, ZHANG Man, YANG Xiaoxia, et al. Recognition of maize seedling and weed based on light weight convolution and feature fusion[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(12): 238-245, 303. (in Chinese)

[5] LIU B, BRUCH R. Weed detection for selective spraying: a review[J]. Current Robotics Reports, 2020, 1(1): 19-26.

[6] DENTON E L, ZAREMBA W, BRUNA J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[J]. Advances in Neural Information Processing Systems, 2014, 27: 1269-1277.

[7] LU Y, KUMAR A, ZHAI S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 5334-5343.

[8] LEBEDEV V, GANIN Y, RAKHUBA M, et al. Speeding-up convolutional neural networks using fine-tuned CP-decomposition[C]// International Conference on Learning Representations. ICLR, 2015: 149801.

[9] LIN S, JI R, CHEN C, et al. Holistic CNN compression via low-rank decomposition with knowledge transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(12): 2889-2905.

[10] CHEN W, WILSON J, TYREE S, et al. Compressing neural networks with the hashing trick[C]// International Conference on Machine Learning. PMLR, 2015: 2285-2294.

[11] RASTEGARI M, ORDONEZ V, REDMON J, et al. Xnor-net: imagenet classification using binary convolutional neural networks[C]// European Conference on Computer Vision. Cham: Springer, 2016: 525-542.

[12] 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/OL]. (2016-02-09)[2022-06-24]. https://arxiv.org/abs/1602.02830.

[13] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. NIPS, 2015: 1135-1143.

[14] SRINIVAS S, SUBRAMANYA A, VENKATESH B R. Training sparse neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2017: 138-145.

[15] JIN J, YAN Z, FU K, et al. Neural network architecture optimization through submodularity and supermodularity[J/OL]. (2018-02-21)[2022-06-24]. https://arxiv.org/abs/1609.00074.

[16] ZOPH B, LE Q V. Neural architecture search with reinforcement learning[C]// International Conference on Learning

Representations. MIT Press, 2017: 1 – 16.
[17] BAKER B, GUPTA O, NAIK N, et al. Designing neural network architectures using reinforcement learning[C]// International Conference on Learning Representations. MIT Press, 2017: 1 – 18.
[18] HE Y, ZHANG X, SUN J. Channel pruning for accelerating very deep neural networks[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 1389 – 1397.
[19] LUO J H, ZHANG H, ZHOU H Y, et al. Thinet: pruning CNN filters for a thinner net[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2019, 41: 2525 – 2538.
[20] LIN S, JI R, LI Y, et al. Accelerating convolutional networks via global & dynamic filter pruning[C]// Proceedings of International Joint Conference on Artificial Intelligence. IJCAI, 2018: 2425 – 2432.
[21] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 2736 – 2744.
[22] HE Y, KANG G, DONG X, et al. Soft filter pruning for accelerating deep convolutional neural networks[C]// Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI, 2018: 2234 – 2240.
[23] GAO X, ZHAO Y, DUDZIAK Ł, et al. Dynamic channel pruning: feature boosting and suppression[C]// International Conference on Learning Representations. MIT Press, 2019: 1 – 14.
[24] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[C]// Proceedings of International Conference on Learning Representations. MIT Press, 2017: 1 – 13.
[25] MOLCHANOV P, MALLYA A, TYREE S, et al. Importance estimation for neural network pruning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 11264 – 11272.
[26] LI B, WU B, SU J, et al. EagleEye: fast sub-net evaluation for efficient neural network pruning[C]// Proceedings of the European Conference on Computer Vision. ECCV, 2020: 639 – 654.
[27] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 7132 – 7141.
[28] CHEBROLU N, LOTTES P, SCHAEFER A, et al. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields[J]. The International Journal of Robotics Research, 2017, 36(10): 1045 – 1052.
[29] 亢洁, 刘港, 郭国法. 基于多尺度融合模块和特征增强的杂草检测方法[J]. 农业机械学报, 2022, 53(4): 254 – 260.
KANG Jie, LIU Gang, GUO Guofa. Weed detection based on multi-scale fusion module and feature enhancement[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(4): 254 – 260. (in Chinese)

(上接第 37 页)

[11] 张安琪,孟志军,陈立平,等. 小型方捆机草捆动态称量系统设计与试验[J]. 农业机械学报, 2020,51(10):170 – 175, 185.
ZHANG Anqi, MENG Zhijun, CHEN Liping, et al. Design and experiment of dynamic weighing system for small square baler[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020,51(10):170 – 175, 185. (in Chinese)
[12] 李伟,张小超,胡小安,等. 联合收获机称量式测产系统软件设计[J]. 农业机械学报, 2011,42(增刊):94 – 98.
LI Wei, ZHANG Xiaochao, HU Xiaoan, et al. Design of intelligent yield monitoring software for combine harvester[J]. Transactions of the Chinese Society for Agricultural Machinery, 2011,42(Supp.):94 – 98. (in Chinese)
[13] 张书慧,马成林,吴才聪,等. 一种精确农业自动变量施肥技术及其实施[J]. 农业工程学报,2003,19(1):129 – 131.
ZHANG Shuhui, MA Chenglin, WU Caicong, et al. Development and application of a variable rate fertilizer applicator for precision agriculture[J]. Transactions of the CSAE, 2003,19(1):129 – 131. (in Chinese)
[14] 黎永键,赵祚喜,黄培奎,等. 基于 DGPS 与双闭环控制的拖拉机自动导航系统[J]. 农业机械学报, 2017,48(2):11 – 19.
LI Yongjian, ZHAO Zuoxi, HUANG Peikui, et al. Automatic navigation system of tractor based on DGPS and double closed-loop steering control[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(2):11 – 19. (in Chinese)
[15] REYES J F, ESQUIVEL W, CIFUENTES D, et al. Field testing of an automatic control system for variable rate fertilizer application[J]. Computers and Electronics in Agriculture, 2015,113: 260 – 265.
[16] RETERU H I, KERSEBAUM K C. Chapter 27 applications in precision agriculture[J]. Geomorphometry: Concepts, Software, Applications, 2009, 33: 623 – 636.
[17] MALEKI M R, JAFARI J F, RAUFAT M H, et al. Evaluation of seed distribution uniformity of a multi-flight auger as a grain drill metering device[J]. Biosystems Engineering, 2006, 94(4): 535 – 543.
[18] DABBAGHI A, MASSAH J, ALIZADEH M. Effect of rotational speed and length of the fluted-roll seed metering device on the performance of pre-germinated paddy seeder unit[J]. International Journal of Natural and Engineering Sciences, 2010,4(3): 7 – 11.
[19] YU H F, DING Y Q, LIU Z, et al. Development and evaluation of a calibrating system for the application rate control of a seed-fertilizer drill machine with fluted rollers[J]. Applied Sciences, 2019, 9(24): 5434.
[20] 丁永前,刘卓,陈冲,等. 基于动态称量原理的泛函式播种施肥量检测方法[J]. 农业机械学报, 2021,52(10):146 – 154.
DING Yongqian, LIU Zhuo, CHEN Chong, et al. Functional detection method of application rate based on principle of dynamic weighing[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(10): 146 – 154. (in Chinese)