计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：融合权重与卷积核删减的 SSD 网络压缩
作者：韩佳林，王琦琦，杨国威，陈隽，王以忠
网络首发日期：2019-08-15
引用格式：韩佳林，王琦琦，杨国威，陈隽，王以忠．融合权重与卷积核删减的 SSD 网络压缩．计算机科学.
http://kns.cnki.net/kcms/detail/50.1075.TP.20190815.0855.002.html

网络首发时间：2019-08-15 09:22:25
网络首发地址：http://kns.cnki.net/kcms/detail/50.1075.TP.20190815.0855.002.html

融合权重与卷积核删减的 SSD 网络压缩

韩佳林[1] 王琦琦[1] 杨国威[1] 陈隽[2] 王以忠[1]
(天津科技大学电子信息与自动化学院 天津 300000)[1]
(麦克马斯特大学电子工程系 汉密尔顿 L8P3H9)[2]

摘要 目标检测是计算机视觉领域中重要的研究方向。近几年，深度学习在基于视频的目标检测领域取得了突破性研究进展。深度学习强大的特征学习和特征表达能力使其能够自动学习和提取相关特征并加以利用。然而，其复杂的网络结构使深度学习模型具有参数规模大、计算需求高、占用存储空间大等问题。基于深度神经网络的单发多框检测器(Single-shot Multi-box Detector 300, SSD300)能够对视频中的目标进行实时检测，但无法移植到嵌入式设备或移动终端以满足实际应用中的需求。为了解决该问题，文中提出了一种权重删减和卷积核删减融合的方法。首先，针对深度卷积神经网络模型权重参数过多导致模型过大的问题，采用权重删减的方法移除各卷积层中的冗余权重，确定各层权重的稀疏度；然后，针对卷积层计算量大的问题，根据各卷积层中的权重稀疏度对冗余卷积核进行删减，以减少冗余参数和计算量；最后，对删减后的神经网络进行训练以恢复其检测精度。为验证该方法有效性，在卷积神经网络框架 caffe 平台上对 SSD 网络模型进行验证实验研究。结果表明，压缩加速后的 SSD300 网络模型的大小为 12.5MB，检测速度最高可达 50FPS (frames per second)。实验实现了在网络检测准确率下降尽量小的前提下，将 SSD300 网络压缩了 8.4×，加速了 2×。权重删减和卷积核删减融合的方法为 SSD300 网络在视频检测中的智能化应用提供了可行性方案。

关键词 深度神经网络，单发多框检测器，网络压缩与加速，权重删减，卷积核删减

中图法分类号 TP183；TP391.4　文献标识码 A　　**DOI 10.11896/jsjkx.180901630**

SSD Network Compression Fusing Weight and Filter Pruning

HAN Jialin[1] WANG Qiqi[1] YANG Guowei[1] CHEN Jun[2] WANG Yizhong[1]

(School of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300000, China) [1]

(Department of Electronic Engineering, McMaster University, Hamilton L8P3H9, Canada)[2]

Abstract Object detection is an important research direction in the field of computer vision. In recent years, deep learning has achieved great breakthroughs in object detection which is based on the video. Deep learning has powerful ability of feature learning and feature representation. The ability enables it to automatically learn, extract and utilize relevant features. However, complex network structure makes the deep learning model have a large scale of parameter.The deep neural network is both computationally intensive and memory intensive. Single Shot MultiBox Detector300 (SSD300), a single-shot detector, produces markedly superior detection accuracy and speed by using a single deep neural network. But it is difficult to deploy it on object detection systems with limited hardware resources. To address this limitation, weight pruning and filter pruning method was proposed to reduce the storage requirement and inference time required by neural networks without affecting its accuracy. Firstly, in order to reduce the number of excessive weight parameters in the model of deep neural network, the weight pruning method is proposed. Network connections is pruned, in which weight is unimportant. Then, to reduce the large computation in convolution layer, the redundant filters are pruned according to the percentage of effective weights in each layer. Finally, the pruned neural network is trained to restore its detection accuracy. To verify the effectiveness of the method, the SSD300 was validated on caffe which is the convolutional neural network framework. After compression and acceleration, the storage of SSD300 neural network required is 12.5MB and the detection speed is 50FPS. The fusion of weight and filter pruning achieves the result by 2×speed-up, which reduces the storage required by SSD300 by 8.4×, as little increase of error as possible. The weight and filter pruning method makes it possible for SSD300 to be embedded in intelligent systems to detect and track objects.

Keywords Deep neural networks, Single-shot multi-box detector (SSD), Network compression and acceleration, Weight pruning, Filter pruning

1　引言

目标检测是计算机视觉领域中非常重要的一个研究方向。随着互联网、人工智能技术、智能硬件的迅猛发展，人类生活中存在着大量的图像

韩佳林（1992-），女，硕士生，主要研究方向为深度学习、网络压缩与加速；
王琦琦(1984－)，男，博士，讲师，主要研究方向为深度学习、室内定位，E-mail:wangqiqi@tust.edu.cn（通信作者）；
杨国威(1988－)，男，博士，讲师，主要研究方向为视觉检测、深度学习与人工智能；
陈隽（1978-），男，博士，教授，主要研究方向为信息论、数字通信、多媒体信号处理、分布式数据压缩和存储、机器学习以及大数据处理；
王以忠（1963-），男，博士，教授，主要研究方向为深度学习与人工智能。

和视频数据，这使得计算机视觉技术在人类生活中起到的作用越来越大，对计算机视觉的研究也越来越火热。近几年，深度神经网络在目标检测领域取得了突破性研究成果，其检测精度与速度均有显著的提升，如 R-CNN[1], Fast-RCNN[2], Faster-RCNN[3], YOLO[4], SSD[5]等。经大规模数据训练获得的深度网络模型由于具有高度计算密集型和内存密集型的特性，难以嵌入到计算资源和内存条件有限的硬件设备或移动终端中。当前性能最优的 SSD300 网络模型能够对视频中出现的目标进行实时检测，然而将其应在实际应用产品化的过程中仍具有两大难点：1）模型大，SSD300 网络模型的大小超过了 100M；2）计算量大，SSD300 网络模型具有成千上万的参数，其良好的的检测性能（74.3%的检测精度，59FPS 的检测速度）主要依赖于高端的硬件配置（Nvidia Titan X）。这两个问题就是将其应用到智能化视频检测系统中的主要障碍。对 SSD 网络进行压缩与加速以降低其硬件需求，具有极大的研究价值, 在视频监控、信息安全、自主驾驶、无人机导航、国防系统等领域具有极大的应用价值。

2 网络压缩与加速

深度神经网络的压缩与加速已经成为现阶段深度学习领域的主要研究方向之一。2016 年，Han 等[6-7]提出了一种网络剪枝的方法来对训练好的网络模型中的不重要连接进行删减，并采用再训练的方式恢复准确度。该方法的有效性在 AlexNet 和 VGGNet 网络上得到了很好的验证。2016 年，Rastegari 等[8] 和 Courbariaux 等[9]采用网络量化的方法，通过权重二值化操作，达到了减少模型存储空间的目的。Hinton 等[10]和 Sau 等[11]基于知识蒸馏方法，将预先训练好的复杂“教师”模型的输出作为监督信号去训练一个简单的“学生”网络，从而达到压缩网络的目的。低秩分解的压缩方法主要基于矩阵分解理论实现对神经网络中参数的压缩[12-13]。

为了使网络目标函数能够收敛到最优，一般情况下，设计的网络模型的容量往往会比实际所需容量大，因此，卷积核内部、卷积核之间存在很大的冗余。2016 年, Wen 等[14]提出了一种 SSL 的压缩方法来对网络结构进行稀疏化学习，对冗余卷积核、特征图、网络深度进行删减。2017 年，Liu 等[15]提出了名为 Slimming 的网络训练算法来对冗余特征图进行删减。2017 年，He 等[16]提出采用 Lasso 回归的算法删减网络中的冗余特征图。该类方法通过在目标函数中引入限定参数的方式删减网络模型中的冗余，过多参数的引入使得网络的调整和优化的过程较为复杂。

除了在原有网络上提出修剪或是加速的方法外，研究人员相继设计出了不同的轻量化网络结构[17-19]。轻量化网络对深度神经网络进行小型化处理，而其难点在于难以设计和训练。

根据以上研究基础，本文提出权重删减和卷积核删减融合的方法以对 SSD300 网络模型进行压缩与加速。本文的研究重点是减小 SSD300 网络的存储空间，加快其检测速度并保证其原有的检测精度。最后，在 PASCAL VOC 数据集[20]上进行大量实验以证实本文方法的有效性。

3 本文方法的设计

3.1 权重删减方法

权重删减在深度神经网络压缩中得到了广泛应用。权重删减可以约减不必要的权值连接，从而防止过拟合现象的发生[21-23]。Han 等[2]针对深度网络结构模型，通过权重删减方法，大大减少了网络参数量，结果显示该方法可以有效减小网络存储，达到网络压缩的目的。权重删减将移除所有绝对值小于阈值权重的连接，删减后的权重以零参数形式存在，即为一种带有稀疏性的零化操作。低于阈值的连接被删除后，稠密网络转换为稀疏网络。其原理如图 1 所示。

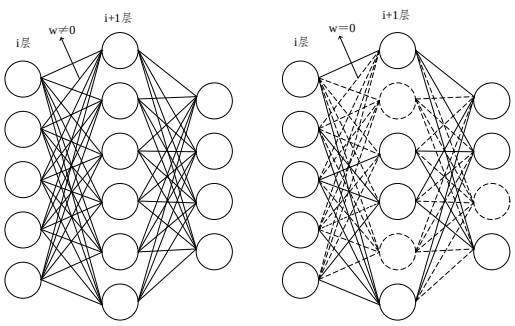


图 1　权重删减原理图

Fig.1 Schematic diagram of weight pruning

如图 1 所示，第 i 层输入神经元的个数为 N_{i-1}，输出神经元的个数为 N_i，权重 w 的个数 C 为：

$$C = N_{i-1} \cdot N_i \tag{1}$$

经权重删减后的权重个数仍为 C，然而其中更多冗余权重 w 以零值形式存在，非零权重个数为 C'，此时网络第 i 层的稀疏度为：

$$\eta = \frac{C'}{C} \tag{2}$$

权重删减流程图如图 2 所示。将删减后的权重值作为神经网络的初始值，对神经网络进行训练以完成对参数的更新。然后再次对网络进行删减，重复该过程，直至训练完成。

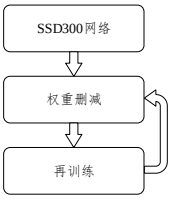


图 2 权重删减流程图

Fig.2 Flow diagram of weight pruning

当运算数据量很大且稀疏时，采用稀疏的数据存储格式可以节省大量的存储空间。删减后的权重是一个典型的稀疏矩阵，针对稀疏矩阵的存储，本文使用了按行存储（Compressed Sparse Row, CSR）格式。CSR 是比较标准的稀疏矩阵存储格式，需要 3 类数据进行表示：数值、列号以及行偏移。数值为矩阵中的所有非零数值，列号为数值对应的所在列号，行偏移表示某一行的第一个非零元素在数值里面的起始偏移位置，行偏移最后要补上总的非零个数。CSR 存储格式如图 3 所示，稀疏矩阵的行列存储格式需要存储 $2a+n+1$ 个数值，其中 $2a$ 为非零值个数，n 为行数[24]。

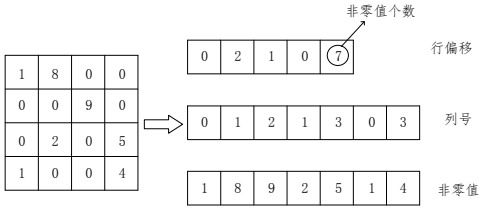


图 3 CSR 存储格式原理图

Fig.3 Schematic diagram of CSR storage format

3.2 卷积核删减方法

卷积是深度神经网络中的基础模块，其主要利用卷积核对输入图片进行处理，以提取拓扑对应性和鲁棒性较高的特征。为了使网络目标函数能够收敛到最优，一般设计的卷积中卷积核个数往往会比实际所需多，删减冗余卷积核可以有效减小网络模型的大小和计算量。卷积操作中卷积层 i 输入特征图的高、宽、通道数分别为 h_i，w_i，n_i，经卷积核操作，产生的输出特征图的高、宽、通道数分别为 h_{i+1}，w_{i+1}，n_{i+1}。每个输出特征图通道均由一个三维卷积核操作生成，每个卷积核对应输入特征图的 n_i 个通道，每个三维卷积核由 n_i 个 $k_i \cdot k_i$ 大小的二维卷积核组成。卷积层 i 的计算量和参数量分别如式（3）和式（4）所示：

$$G = k_i \cdot k_i \cdot n_i \cdot h_{i+1} \cdot w_{i+1} \cdot n_{i+1} \tag{3}$$

$$W = k_i \cdot k_i \cdot n_i \cdot n_{i+1} \tag{4}$$

删减冗余卷积核即可避免提取不必要的特征图，进而能够有效减少卷积层的计算量和参数量，其原理如图 4 所示。

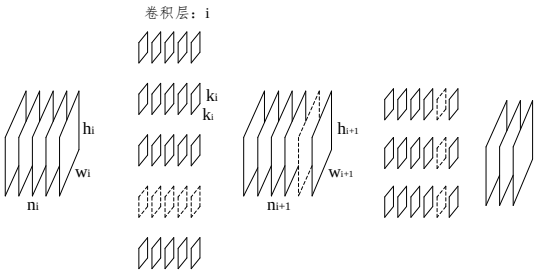


图 4 卷积核删减原理图

Fig.4 Schematic diagram of filter pruning

删减 i 层中 m 个卷积核，其相应的 m 个输出特征图通道也会随之移除。卷积层 i 中计算量的减少量和参数量的减少量分别如式（5）和式（6）所示：

$$\Delta G_i = k_i \cdot k_i \cdot n_i \cdot h_{i+1} \cdot w_{i+1} \cdot m \tag{5}$$

$$\Delta W_i = k_i \cdot k_i \cdot n_i \cdot m \quad (6)$$

此外，移除卷积层 i 特征图通道的同时，卷积层 $i+1$ 中与其对应的卷积核也会被随之删减，此时 $i+1$ 层中计算量的减少量和参数量的减少量分别如式（7）和式（8）所示：

$$\Delta G_{i+1} = k_i \cdot k_i \cdot n_i \cdot h_{i+1} \cdot w_{i+1} \cdot m \quad (7)$$

$$\Delta W_{i+1} = k_{i+1} \cdot k_{i+1} \cdot \mathrm{m} \cdot n_{i+2} \quad (8)$$

3.3 权重删减与卷积核删减融合方法

卷积神经网络的特性包括局部连接、权值共享等特性，且都蕴含着稀疏性，首先针对局部连接，相比于全连接策略，它更符合外侧膝状体到初级视觉皮层上的稀疏响应特性；其次权值共享的特性可进一步约束相似隐层单元具有同样的激活特性，使得局部连接后的权值具有结构特性，在实际应用中可进一步约减参数个数[25]。本文提出的权重删减和卷积核删减融合的方法能对 SSD300 网络进行有效压缩与加速。首先，使用权重删减的方法移除冗余权重连接；然后确定各卷积层删减后权重的稀疏度；最后根据权重的稀疏度对各层卷积核进行删减。权重删减和卷积核删减融合的方法是一种简单且有效的网络压缩与加速的方法。假设神经网络某一层中卷积核个数为 n_i，权重个数为 W，则：1）选取不同阈值对权重进行删减；

2）确定最佳阈值对权重进行删减，确定保留下来的有效权重个数为 W'；

3）计算有效权重稀疏度 $\eta = \frac{W'}{W}$；

4）确定卷积核保留个数 $n' = n \cdot \eta$，删减卷积核个数 $m = n - n'$。

5）删减后的网络需要通过再训练的方式恢复网络检测精度。

4 实验结果与分析

本文采用卷积神经网络开源框架 caffe 进行实验研究。服务器配置如下：

16.04-Ubuntu，CPU 为 Intel Xeon E5-2640 v4 处理器，主频为 2.1GHz，内存为 32GB，GPU 为 NVIDIATITAN Xp，显存为 12GB。

SSD300 网络以 VGG 网络为基础，并将 VGG 网络尾部的全连接层替换为了多个卷积层，整个网络不包括全连接层，主要由 35 个卷积层组成，其中 23 个卷积层用于图像特征提取，6 个卷积层用于类别信息预测，6 个卷积层用于位置信息预测，并且类别信息以及位置信息的预测均在底层产生的 6 层特征图上分别进行。

实验数据为 SSD300 模型采用的 PASCAL VOC（VOC2007，VOC2012）数据集，该数据集（VOC2012 trainval，VOC2007 trainval，VOC2007test）共包含 21503 张图像，并且标出了 20 个种类共计 52090 个物体。其中 16551 张（VOC2012 trainval、VOC2007 trainval）作为训练集，4952 张（VOC2007test）作为测试集。

本实验分别用权重删减方法和本文提出的权重删减与卷积核删减融合的方法对 SSD300 网络进行压缩加速研究，并将两种方法对 SSD300 网络的压缩加速效果进行了比较。

4.1 权重删减实验

1）训练 SSD300 网络。SSD300 网络经过 12 万次迭代后，平均精确度的均值（Mean Average Precision, mAP）为 77.0%，网络模型大小为 105.2M，检测速度为 25FPS(测试硬件为 16.04-Ubuntu，CPU 为 Intel Core i5-7400 处理器，主频为 3GHz，内存为 16GB，GPU 为 GeForceGTX1060，显存为 6GB)。

2）各卷积层选取不同的阈值进行实验。对各卷积层选取不同阈值进行权重删减，并测试其检测精度，确定其重要权重保留比例。SSD300 网络卷积层 conv1_1 在不同权重保留比例下的网络检测精度的大小如图 5 所示。网络检测精度随各卷积层权重保留比例的增大呈增长的趋势，最终趋于稳定。根据检测精度随权重保留比例的增长率的变化趋势确定卷积层 conv1_1 的权重保留比例为 70%。

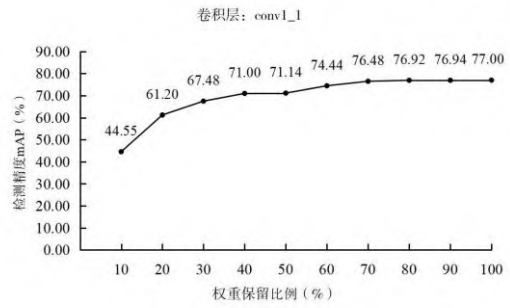
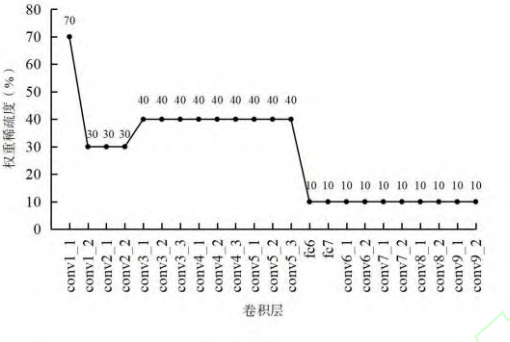


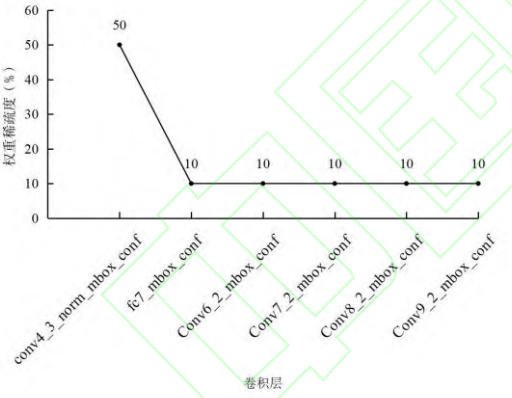
图 5　权重保留比例与网络检测精度关系图

Fig.5 Relationship diagram of weight spasity and mAP

3）确定各卷积层权重的稀疏度。根据 2）中的



(a) 特征提取层权重稀疏度



(b) 类别预测层权重稀疏度

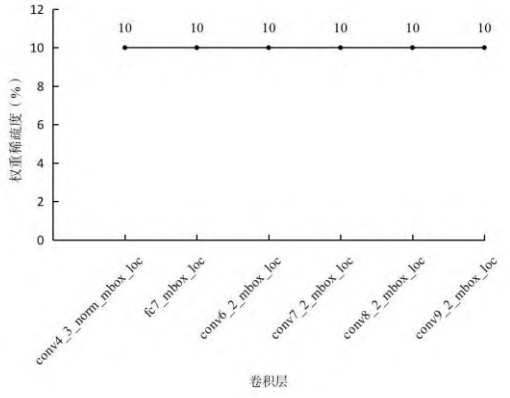


图 6　SSD300 网络卷积层权重稀疏度

Fig.6 Weight spasity of each convolution layer in SSD300

4）对删减后的网络进行再训练以恢复网络检测精度。以 0.0001 的学习率训练网络 1500 次进行一次删减，计作一次迭代。训练结果如表 1 所列。

表 1　权重删减实验的训练结果

Table1 Training result of weight pruning experiment

迭代次数	检测精度/%
5	71.3
10	74.7
15	75.4
20	75.6
27	76.1

4.2　卷积核删减实验

根据删减后网络各卷积层权重的稀疏度，确定各层卷积核的保留个数。特征提取层中卷积核删减前后的个数如图 7 所示。系列 1 为删减前各卷积层中卷积核的个数，系列 2 为删减后各卷积层中卷积核的个数。设计好的 SSD300 网络模型各层卷积核均存在较大的冗余，顶层卷积层中的卷积核可移除掉 90%。

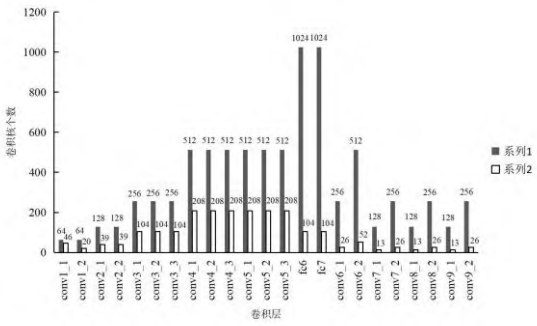


图 7　SSD300 网络删减前后卷积核的个数

Fig.7 Weight spasity of each convolution layer in SSD300

对删减后的 SSD300 网络进行训练以恢复检测精度，首先以学习率 0.001 迭代 350000 次，再以学习率 0.0001 迭代 50000 次，最后以学习率 0.00001 迭代 10000 次，训练结果如表 2 所列。

表 2　卷积核删减实验训练结果

Table 2 Training result of filter pruning experiment

迭代次数	检测精度/%
50000	37.2
100000	56.6
150000	61.6
200000	65.4
250000	67.1
300000	69.5
350000	69.8
370000	70.5

4.3 SSD 网络压缩加速结果

分别采用权重删减的方法以及权重删减和卷积核删减融合的方法对 SSD 网络进行压缩加速研究，实验结果如表 3 所列。在保证网络检测精度损失尽可能小的前提下，权重删减的方法将 SSD300 网络存储大小减小 3 倍。采用权重删减和卷积核删减融合方法，SSD300 网络模型压缩了 8.4×，加速了 2×。

表 3 不同方法对 SSD300 网络压缩加速效果对比

Table 3 Compression and acceleration results of different methods for SSD300 network

网络模型	SSD300 模型	权重删减得到的 SSD300 模型	融合删减得到的 SSD300 模型
模型大小	105.2MB	35.9MB	12.5MB
检测速度	25FPS	25FPS	50FPS
检测精度	77.0%	76.1%	70.5%
压缩率	-	2.9×	8.4×
加速率	-	0×	2×

使用权重删减方法对冗余权重进行零化操作，使网络连接具有稀疏性，可有效减少参数存储量，但零化后的权重仍会参与到网络计算中，该方法对神经网络具有压缩效果，没有加速效果。而使用权重删减和卷积核删减融合的方法对冗余卷积核进行删减，可以有效减少卷积核的个数，从而减少卷积层参数量和计算量，能对网络进行有效压缩和加速。

结束语

深度神经网络已经在目标检测领域获得了前所未有的成功，然而其参数量大，计算复杂的特性是将其部署到移动端上的主要障碍。SSD300 网络的检测速度快且检测精度高，但同样具有以上缺陷。本文提出的权重删减和卷积核删减的融合方法可有效移除卷积层中冗余的卷积核，减小模型的存储内存，减少网络计算量。在精度损失尽可能小的情况下，该方法可将 SSD300 网络压缩 8.4×，加速 2×，这证实了本文方法的有效性。

该方法操作简单方便，能够有效删减卷积神经网络中的冗余卷积核。但该方法也有一定的缺点，即其虽然具有良好的压缩加速效果，但牺牲了一定的检测精度。该方法还有待改进。

参考文献

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014:580-587.

[2] GIRSHICK R. Fast R-CNN [C]// Proceedings of the IEEE Conference on International Conference on Computer Vision. Boston: IEEE, 2015:1440-1448.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 779-788.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]// Proceedings of European Conference on Computer Vision. Amsterdam: Springer International Publishing, 2016:21-37.

[6] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks [C]// Neural Information Processing Systems. 2015: 1135-1143.

[7] HAN S, POOL J, DALLY W J, et al. Deep Compression: compressing deep neuralnetworks with pruning, trained quantization and huffman coding[C]//Proceedings of Conference on Learning Representations. San Juan: IEEE, 2016: 233-242.

[8] MOHAMMAD R, VICENTE O, JOSEPH R, et al. XOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. [C]// Proceedings of European Conference on Computer Vision. Amsterdam: ECCV, 2016:525-542.

[9] MATTHIEU C, ITAY H, DANIEL S, et al. Binarized Neural Networks: Training Neural

Networks with Weights and Activations Constrained to +1 or -1. [EB/OL]. https://arxiv.org/abs/1704.04861.pdf.

[10] GEOFFREY H, ORIOL V, JEFF D, et al. Distilling the knowledge in a Neural Network. [C]// Proceedings of Conference on Advances in Neural Information Processing Systems. Montreal: IEEE, 2014:2644-2652.

[11] BHARAT BHUSAN S, VINEETH N. B. Deep Model Compression: Distilling Knowledge from Noisy Teachers. [EB/OL].https://arxiv.org/abs/1610.09650.pdf.

[12] MAX J, ANDREA V, ANDREW Z, et al. Speeding up Convolutional Neural Networks with Low Rank Expansions[J]. Computer Science, 2014, 4(4):1-7.

[13] VIKAS S, TARA N. S, SANJIV K,et al. Structured Transforms for Small-Footprint Deep Learning. [EB/OL]. https://arxiv.org/abs/1510.01722.pdf.

[14] WEN W, WU C, WANG Y, et al. Learning structured sparsity in deep neural networks[C]// Advances in Neural Information Processing Systems. 2016:2074-2082.

[15] LIU Z, SHEN Z, HUANG G, et al. Learning efficient convolutional networks through network slimming [C]// Proceedings of the IEEE International Conference on Computer Vision(ICCV). 2017:2755-2763.

[16] HE Y, ZHANG X, SUN J, et al. Channel pruning for accelerating very deep neural networks [EB/OL]. https://arxiv.org/abs/1707.06168.pdf.

[17] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet:AlexNet-level accuracy with 50x fewer parameters and < 0.5MB model size [C]// Proceedings of International Conference on Learning Representations. San Juan: ICLR, 2016.

[18] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. https://arxiv. org/abs/1704. 04861.pdf.

[19] ZHANG X, ZHOU X, LIN M, et al.ShuffleNet: an extremely efficient convolutional neural network for mobile devices [EB/OL]. https:// arxiv.org/abs/1707. 01083.pdf.

[20] EVERINGHAM M, VAN G L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge[J].International journal of computer vision, 2010, 88(2):303-338.

[21] HANSON S J, PRATT L Y. Comparing biases for minimal network construction with back-propagation [C]// Neural Information Processing Systems.1989:177-185.

[22]CUN Y L, DENKER J S, SOLLA S A, et al. Optimal brain damage [C]// Neural Information Processing Systems. Morgan Kaufmann Publishers Inc.,1990:598-605.

[23] HASSIBI B, STORK D G. Second Order derivatives for network pruning: optimal brain surgeon [C]// Neural Information Processing Systems.1992:164-171.

[24] HAN Y F, JIANG T H, MA Y P, et al. Compression of deep neural networks [J]. Application Research of Computers, 2018, 35(10): 2894-2897. (in Chinese)
韩云飞，蒋同海，马玉鹏，等.深度神经网络的压缩研究[J].计算机应用研究，2018, 35(10): 2894-2897.

[25] JIAO L C. Deep Learning, Optimization and Recognition, [M]. Beijing : Tsinghua University Press, 2017:104 (in Chinese)
焦李成. 深度学习、优化与识别[M]. 北京：清华大学出版社，2017:104.