神经网络模型压缩方法综述

张 弛 田 锦 王永森 刘宏哲
(北京联合大学北京市信息服务工程重点实验室 北京 100101)

摘 要 深度神经网络在计算机视觉、语音识别等领域取得了巨大成功。然而,目前的深度神经网络模型需要消耗大量的计算资源和存储空间,限制了在移动终端和车载设备等低存储、低延迟需求环境下的应用。因此,需要在保证准确率的前提下对神经网络进行压缩、加速和优化。文中主要讨论如下3种方法:1)在已有的网络结构下进行参数压缩,包括剪枝、量化和低秩分解;2)使用更加紧凑的网络结构;3)采用知识迁移。对于每一类方法的每个分支,文中都详细介绍了其性能、优缺点及应用,并且列举出了相关领域的最新研究成果。最后,总结了现有的成果并讨论了将来可能的发展方向。

关键词 神经网络,深度学习,模型优化,参数压缩
中图法分类号 TP311 **文献标识码** A

Survey of Model Compression Method for Neural Networks

ZHANG Chi TIAN Jin WANG Yong-sen LIU Hong-zhe
(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)

Abstract Deep neural networks have achieved great success in computer vision and speech recognition. However, existing deep neural network models are computationally expensive and memory intensive, hindering their deployment in devices with low memory resources or in applications with strict latency requirements such as mobile terminal and vehicle equipment. Therefore, it is necessary to compress, accelerate and optimize neural networks without decreasing performance. This paper mainly discussed three kinds of methods: to compress the parameters under the existing network structure including pruning, quantization and low-rank factorization, to use a more compact network structure, and to use knowledge transfer. For each branch of each method, this paper provided insightful analysis regarding the performance, advantages, disadvantages, and applications. Then this paper listed out the latest research progress involved. Finally, the existing methods were summarized and possible future directions were discussed.

Keywords Neural networks, Deep learning, Model optimization, Parameter compression

1 引言

近年来,深度神经网络模型已经成为人工智能领域的研究热点,且被广泛应用在计算机视觉、自然语言处理等许多不同的领域,并在诸如图像分类、目标检测、语义分割等很多实际任务上都取得了巨大的性能提升。这些成功都依赖于深度神经网络成千上万的参数,以及具有强大性能的计算设备。然而,深度神经网络模型复杂、计算量巨大的特点会导致很多问题:1)训练模型十分困难;2)使用模型也很困难。训练上的困难是指一个模型的训练周期非常长,短则数小时,长则数天,并且需要性能非常强大的机器和大量的数据作为支撑。使用上的困难是指现在很多实际应用场景并不具备很高的计算能力和

本文受北京市属高校高水平教师队伍建设支持计划项目(IDHT20170511),英国皇家工程院牛顿基金:面向智能车产业化的人才培养与合作(UK-CIAPP\324)资助。

张 弛(1992—),男,硕士生,主要研究方向为神经网络模型压缩,田 锦(1994—),男,硕士生,主要研究方向为计算机视觉、深度学习;王永森(1994—),男,硕士生,主要研究方向为数字图像处理;刘宏哲(1971—),女,教授,主要研究方向为语义计算、数字图像处理、人工智能。

存储条件，并且对实时性有硬性需求，例如移动终端和驾驶环境，这使得深度神经网络很难实际部署到这些低存储、低功耗的小型设备上。

因此，如何在保证现有神经网络模型性能不变的情况下，有效减小神经网络模型的计算量和存储空间，成为了一个亟待解决的重要问题。我们将这个问题的解决方案分为 3 类：1）在已有的网络结构基础上进行参数压缩，通过剪枝、量化、低秩分解等方式压缩模型大小，并保证精度；2）重新设计更加紧凑的网络结构，减小模型的同时不显著改变模型的性能；3）使用知识迁移，使用大模型中的知识指导小模型的训练过程，从而提升小模型的性能。这 3 类方法都有很多重要的研究成果，我们将在下文中详细介绍。

2 神经网络模型的参数压缩方法

现有的很多神经网络模型具有海量的参数，而这些参数中往往存在大量的冗余。参数压缩方法旨在消除这些冗余的部分，包括信息上的冗余以及空间上的冗余。其中，剪枝的方法是通过搜索模型中冗余的参数并将之修剪掉；量化的方法是通过减小每个权重的比特数来压缩原始网络的存储空间；而低秩分解的方法是将卷积神经网络中的卷积核看成一个矩阵或张量，通过对其进行低秩分解来消除模型中的冗余部分。

2.1 剪枝

神经网络参数压缩的第一种方式是剪枝，即通过修剪神经网络中的连接，将一个复杂度很高的网络转变为一个复杂度较低的网络，有效地压缩了模型的大小，并一定程度上解决了过拟合的问题。文献[1]提出了一种随机的剪枝方法，通过设定一个阈值，随机修剪不重要的连接，再重新训练网络，从而达到参数压缩的目的。在实际应用中，随机剪枝的方法往往不能起到很好的效果，结构化剪枝应运而生。文献[2]通过对权重加上稀疏正则来进行剪枝，首先使用组稀疏方法对分组特征添加稀疏正则来修剪权重矩阵的列，再通过排他性稀疏来增强组间竞争，两者结合取得了很好的剪枝效果。文献[3]提出了一种对特征图通道进行剪枝的算法，通过给每个通道添加一个尺度因子，并对尺度因子添加稀疏正则来对通道进行剪枝。文献[4]通过 lasso 回归对通道进行了剪枝，并使用最小二乘来重构模型。文献[5]对网络训练过程中的梯度信息进行剪枝，从而加快了网络的训练过程并减少了过拟合。表 1 对比了几种典型的剪枝方法。

表 1 典型剪枝方法的对比

剪枝算法	修剪对象	修剪方式	效果
Deep Compression	权重	随机修剪	50 倍压缩
Structured Pruning	权重	组稀疏＋排他性稀疏	性能提升
Network Slimming	特征图通道	根据尺度因子修剪	节省计算资源
mProp	梯度	修剪幅值小的梯度	加速

2.2 量化

权值量化是神经网络参数压缩的另一种重要方法，通过减少每个权重的比特数来压缩原始网络。文献[1]中提出的方法在剪枝的基础上对权重进行聚类，将权重共享后对生成的编码本进行霍夫曼编码，从而实现压缩。量化的极限形式是将每个权值都表达为一位，即二值化神经网络。文献[6-8]是二值化网络的一系列成果，其主要思想是将卷积神经网络的权值、激活函数和特征图进行二值化，并采用一位同或门运算代替矩阵点乘运算，大大减小了计算量和存储空间。文献[9]对反向传播中的梯度实现了量化，将权值用一位表示，激活函数用两位表示，梯度用四位表示，提高了并行计算的效率。图 1 展示了几种不同的二值化神经网络。

2.3 低秩分解

对于深层的卷积神经网络而言，其主要的计算量在于卷积操作。卷积核可以看成是一个矩阵或张量，通过低秩分解的方法来减少矩阵或张量的运算量。对于二维矩阵，文献[10]研究了几种低秩分解方法，如将二维矩阵进行 SVD 分解，三维张量转化为二维矩阵进行分解，以及单色卷积分解和聚类法低秩分解等，通过减少卷积核的冗余来减少计算量，在损失 1%精度的条件下取得了两倍的加速效果。文献[11]采用 CP 分解法将一层复杂网络分解为五层相对简单的网络，使用非线性最小二乘法来计算，以较低的性能损失实现了较大的速度提升。文献[12]使用广义奇异值分解（GSVD）对卷积神经网络的非线性单元进行分解，在增加极小的分类误差情况下实现了 4 倍加速。其他方法还有 Tucker 分解[13]、Tensor Train 分解[14]以及 Block Term 分解[15]等，也都取得了不错的模型压缩效果。图 2 展示了一个低秩分解的例子。

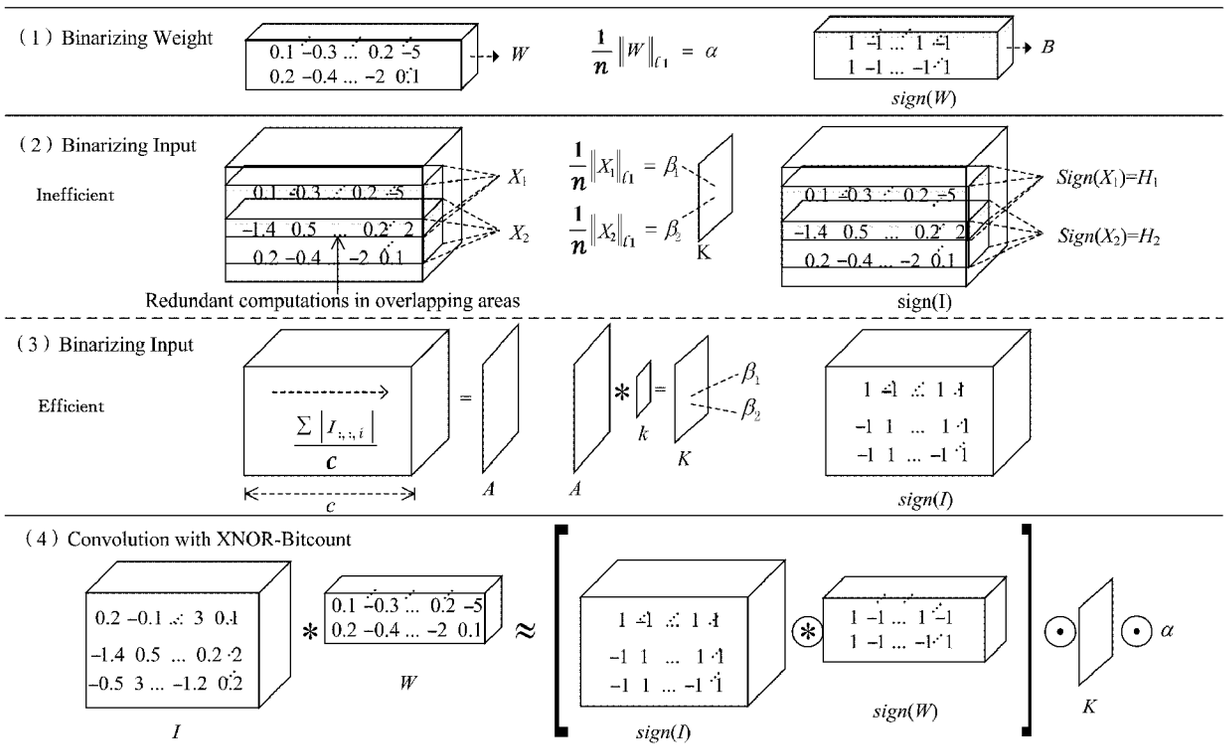

图 1　不同的二值化神经网络

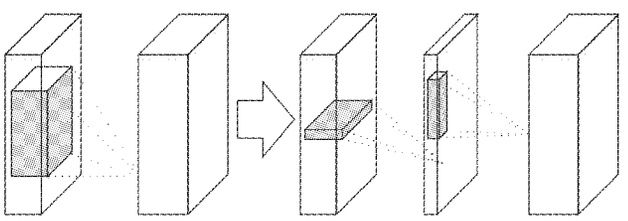

图 2　低秩分解示意图

3　设计更加紧凑的网络结构

相对于在已有的复杂网络基础上进行参数压缩，我们也可以直接设计一些更加轻量化的神经网络，通过改变网络结构和使用更加高效的计算方式使其比复杂网络更加紧凑，在保证网络性能的基础上大幅降低计算量并缩减存储空间。

SqueezeNet[16]提出了一种被称作 Fire 的模块，它分为两部分：一个由 1×1 卷积核构成的压缩层以及一个由 1×1 和 3×3 卷积核组成的扩张层，由这种模块组成的神经网络 SqueezeNet 实现了 50 倍压缩。文章还进一步使用 DeepCompression[1]将模型进一步压缩至 0.5 MB，并且可以达到 AlexNet 级别的性能。

Google 提出的 MobileNets[17]模型是专门针对移动端和嵌入式设备开发的轻量级卷积神经网络。其中的核心思想是使用 Depthwise 卷积，即将一个 $D\times D\times M$ 的卷积核分解为 $D\times D\times 1$ 的逐通道卷积和 $1\times 1\times M$ 的逐点卷积，其中 D 是卷积核尺寸，M 是通道数。对于 3×3 卷积核而言，计算量减小至 1/9～1/8。MobileNet V2[18]模型在 V1 的基础上引入了残差结构，并使用线性函数代替 Relu 以减少特征损失，进一步提升了 Mobilenet 模型的性能。

ShuffleNet[19]模型使用分组卷积的方式减少了计算量，但分组卷积的每一组只接收上一层的部分信息，而组间信息无法流通，于是使用 Channel Shuffle 来打破组间隔阂，解决了信息流通问题。表 2 对几种轻量化网络结构进行了对比。

表 2　几种轻量化网络结构的对比

网络结构	TOP 1 准确率/%	参数量 /M	CPU 运行时间/ms
MobileNet V1	70.6	4.2	123
ShuffleNet (1.5)	69.0	2.9	—
ShuffleNet (x2)	70.9	4.4	—
MobileNet V2	71.7	3.4	80
MobileNet V2(1.4)	74.7	6.9	149

4　知识迁移

神经网络模型压缩的思路是减小模型的体量，但往往小模型的精度相比大模型会有所下降，知识迁移(Knowledge Transfer)也称知识蒸馏(Knowledge Distillation)的方法就通过大模型学到的知识来指导小模型的训练，从而提升小模型的性能。文献[20]提出了一种教师-学生网络(见图 3)，将学生模型的优化目标分为两部分：1)学生模型的 Hard

Target，即学生模型输出的类别概率与真值的交叉熵；2）学生模型与教师模型的 Soft Target 计算交叉熵。将这两个优化目标进行组合，使学生模型能够模仿教师模型输出的概率分布，让学生模型在训练中的信息更丰富，有助于提高学生模型的性能。

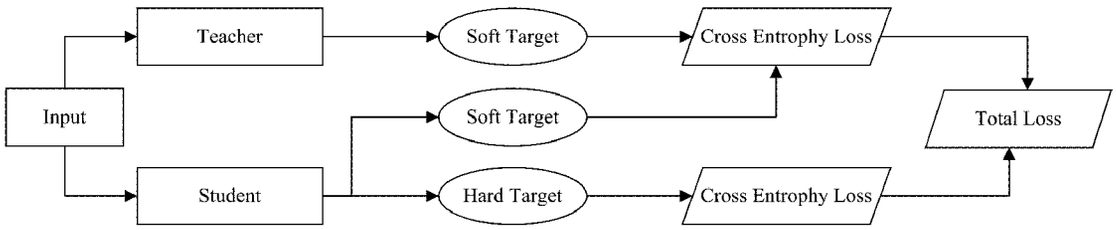

图 3 教师-学生网络[20]

文献[21]认为仅通过让学生模型在输出端模仿教师模型的方法对于深层神经网络比较困难，文中提出的 FitNets 模型在网络的中间层添加监督信号，先训练学生模型的前半部分参数，让网络中间层的输出也尽可能一致，再使用知识蒸馏的方法训练全体参数。文献[22]希望使用学生模型来模仿教师模型层与层之间的关系，这种关系通过层与层之间的内积来定义，教师和学生网络的层间关系定义为 FSP 矩阵，优化目标为教师与学生网络对应层之间 FSP 之差的 L_2 范数。文献[23]将知识迁移作为域自适应问题，通过匹配教师模型和学生模型特征图的激活分布来实现知识迁移，使用最大平均差异（MMD）定义损失函数，并通过与其他知识迁移方法的结合达到了很好的效果。

结束语 现有的神经网络模型压缩方法主要是针对卷积神经网络（CNN），未来可能会有针对其他深度神经网络（RNN 和 LSTM 等）的研究成果。现在很多通用的参数压缩方法如剪枝、量化等应用在硬件上时往往达不到很好的效果，所以今后可能会结合硬件来研究一些有针对性的参数压缩方法。另外，现有的模型压缩方法大多针对分类问题，未来在目标检测和语义分割等任务上也会有类似的工作。而且，现阶段模型压缩的方法都需要进行 fine-tuning，需要一些有标签的训练样本，未来或许会有一些小样本甚至无监督的模型压缩方法出现。

参 考 文 献

[1] HAN S, MAO H, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149, 2015.

[2] YOON J, HWANG S J. Combined group and exclusive sparsity for deep neural networks[C]// International Conference on Machine Learning. 2017: 3958-3966.

[3] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 2755-2763.

[4] HE Y, ZHANG X, SUN J. Channel pruning for accelerating very deep neural networks[C]// International Conference on Computer Vision (ICCV). 2017: 6.

[5] SUN X, REN X, MA S, et al. meProp: Sparsified back propagation for accelerated deep learning with reduced overfitting[J]. arXiv preprint arXiv:1706.06197, 2017.

[6] COURBARIAUX M, BENGIO Y, DAVID J P. Binaryconnect: Training deep neural networks with binary weights during propagations[C]// Advances in Neural Information Processing Systems. 2015: 3123-3131.

[7] 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]. arXiv preprint arXiv:1602.02830, 2016.

[8] 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.

[9] ZHOU S, WU Y, NI Z, et al. DoReFa-Net: Training low bit-width convolutional neural networks with low bit-width gradients[J]. arXiv preprint arXiv:1606.06160, 2016.

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

[11] LEBEDEV V, GANIN Y, RAKHUBA M, et al. Speeding-up convolutional neural networks using fine-tuned cp-decomposition[J]. arXiv preprint arXiv:1412.6553, 2014.

[12] ZHANG X, ZOU J, HE K, et al. Accelerating very deep convolutional networks for classification and detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 1943-1955.

（下转第 24 页）

4.3 教学视频的录制过程

工作人员根据前期编写好的文字稿本搭建录制工作的平台，主要由摄像机、切换台、音频系统、教师电脑、展示屏幕以及非线性编辑系统等构成。“录制”阶段要根据计划，安排室内或室外场景录制视频，录制过程中，教师电脑和展示屏幕刷新频率的设置需要同步，避免摄像机镜头拍摄的画面出现抖屏现象。

教师需要提前试讲，调整状态，适应场景，还要确保素材的正确性，提高对素材的熟悉程度。在开始拍摄课程前，摄像师需要依据教师的讲课活动范围固定好全景机位和近景机位，调整摄像机的镜头，保持画面的构图能突出主体，并调整好摄像机的光圈和白平衡，使主体的全景画面和近景画面的光纤效果一致，前后焦点清晰。注意音频、灯光、机位的选择以及画面构图的美观性等关键点，以提高视频的质量。

4.4 教学视频的后期编辑

视频的后期编辑要依据教学设计和稿本设计，利用非线性编辑系统对录制的整个视频进行剪辑与特效处理，同时适当加入与教学相关的媒体素材，进行声音降噪，提高视频的声音清晰度，并利用软件添加字幕、设计片头片尾，最终形成一部主题突出、教学情节连贯、媒体展示丰富、画面流畅的视频课程作品。在导出视频时，需确认 MOOC 平台对视频文件格式的要求。

结束语 MOOC 作为互联网时代的新型教育形式，很多学科课程都进行了大胆尝试，视频也将成为主流教学活动的重要素材，而 MOOC 课程视频的设计与制作是整个 MOOC 课程开发的关键核心环节。MOOC 视频制作的根本目的是提高学习者的学习效果，在视频设计与制作中不能偏离这个主题。一个良好的 MOOC 视频应该能满足学习者的一般学习需求，这就要求除 MOOC 视频本身知识内容完整、科学外，还要求 MOOC 视频具有美的展现形式，促使学习者能坚持不懈地继续学习，达到良好的学习效果。

参 考 文 献

[1] 胡进，白冰，王亚君，等. MOOC 背景下网络视频课程的设计与开发研究[J]. 中国教育信息化，2016(8)：12-14.

[2] 马清. 高校 MOOC 视频资源设计与开发研究[J]. 辽宁省交通高等专科学校学报，2018(2)：73-75.

[3] 伍文燕，张振威. MOOC 环境下视频资源呈现形式研究[J]. 中国教育信息化，2017(11)：45-47.

[4] 何敬，潘宇，王艳萍. MOOC 视频设计与制作研究[J]. 软件导刊教育技术，2016(9)：25-27.

[5] 梁金兰. 高校 MOOC(慕课)视频制作探讨[J]. 软件导刊教育技术，2018(5)：95-96.

(上接第 4 页)

[13] KIM Y D, PARK E, YOO S, et al. Compression of deep convolutional neural networks for fast and low power mobile applications[J]. arXiv preprint arXiv: 1511.06530, 2015.

[14] NOVIKOV A, PODOPRIKHIN D, OSOKIN A, et al. Tensorizing neural networks[C]// Advances in Neural Information Processing Systems. 2015: 442-450.

[15] WANG P, CHENG J. Accelerating convolutional neural networks for mobile applications[C]// Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016: 541-545.

[16] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size[J]. arXiv preprint arXiv: 1602.07360, 2016.

[17] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv: 1704.04861, 2017.

[18] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[19] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[J]. arXiv preprint arXiv: 1707.01083, 2017.

[20] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. ar-Xiv preprint arXiv: 1503.02531, 2015.

[21] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv: 1412.6550, 2014.

[22] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[23] HUANG Z, WANG N. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer[J]. arXiv preprint arXiv: 1707.01219, 2017.