卷积神经网络算法模型的压缩与加速算法比较

李思奇

（北京物资学院，北京 101149）

摘 要：随着深度学习网络的不断发展，卷积神经网络在图像识别与处理领域的正确率已达到甚至超越人类水平。但是，越来越复杂的网络结构导致庞大的计算模型体积和计算量，不利于模型的移植利用。基于此，分别介绍了网络压缩加速的典型方法并进行比较，在保证算法准确率损失最少的前提下，尽可能使算法具有可移植性，充分体现卷积神经网络算法的应用价值。

关键词：卷积神经网络；网络压缩；网络加速；模型移植

中图分类号：TP18 **文献标识码：**A **文章编号：**1003-9767（2019）11-021-03

Comparison of Compression and Acceleration Algorithms for Convolutional Neural Network Model

Li Siqi

(Beijing Wuzi University, Beijing 101149, China)

Abstract: With the continuous development of deep learning network, the accuracy of convolutional neural network in image recognition and processing has reached or even surpassed human level. However, more and more complex network structure leads to huge computing model volume and computation, which is not conducive to the transplantation and utilization of the model. Based on this, the typical methods of network compression and acceleration are introduced and compared. On the premise of guaranteeing the least loss of accuracy, the algorithm can be transplanted as far as possible, which fully reflects the application value of convolutional neural network algorithm.

Key words: convolutional neural network; network compression; network acceleration; model transplantation

0 引言

目前，深度神经网络已被广泛应用于众多领域，尤其在图像处理领域，基于卷积神经网络的模型训练效果明显优于其他传统方法。但是，模型复杂度不断提高，有限的存储空间和计算能力成为进一步训练大规模任务的瓶颈。因此，如何使神经网络具有可移植性，在更多的应用场景下体现价值，成为了研究重点之一。

1 模型压缩方法介绍

如果想要在移动平台上运用卷积神经网络模型，有两个关键点，一个是模型参数要少，一个是运行时间短。深度神经网络的加速方法主要包括两大类，硬件加速和软件加速。硬件加速对算法的改动较小，如结构优化，通过整型数据代替浮点数数据，减小模型体积，或使用专门应用于深度学习领域的硬件。软件加速通过修改原始模型实现结果，分为两类，压缩结构和压缩参数。本文主要研究基于软件的模型加速方法。

1.1 基于压缩结构的模型加速方法

压缩结构的方法是基于现有卷积神经网络结构修改原始模型的卷积运算，重新设计网络结构，实现减少参数量和降低计算复杂性的目标。设计轻量级网络模型的核心是设计一种更有效的“网络计算方法”，主要用于卷积方法，从而减少网络计算量和参数，且不损失网络性能[1]。

1.1.1 标准卷积操作

卷积神经网络的基础是标准卷积，其是卷积神经网络最基本的运算。假设 $H\times W$ 表示输入特征空间尺寸，H 和 W 分别代表特征图的高度和宽度，输入和输出特征空间尺寸不变，N 是输入特征通道数，$K\times N$ 表示卷积核尺寸，卷积核的数量表示输出特征图通道数，用 M 表示。因为每个通道卷积运算后都要对结果求和，所以该值和每个通道特征都有关。标准卷积计算量是 $HWNK^2M$，标准卷积层的参数量为 K^2MN。

作者简介：李思奇（1993—），男，河北保定人，硕士研究生在读。研究方向：深度学习网络压缩与加速。

1.1.2 可分离卷积操作

深度可分离卷积实际是把标准卷积变换为一个深度卷积和一个逐点卷积的运算。对于相同的输入特征 $H\times W\times N$，标准卷积核尺寸为 $K\times K\times N$，数量为 M，则标准卷积的计算复杂度为 $HWNK^2M$，标准卷积的参数量为 K^2MN。将标准卷积分解成深度卷积和逐点卷积后，深度卷积核尺寸为 $K\times K\times 1$，$K\times K$ 为卷积核空间尺寸，通道数为 1，卷积核数量为 N，则输入特征通过深度卷积运算的计算复杂度为 $HWNK^2$，深度卷积层的参数量为 K^2N。通过深度卷积后进行 1×1 逐点卷积，卷积核通道数为 N，卷积核数量为 M，输入特征通过逐点卷积运算后计算复杂度为 $HWNM$，参数量为 MN。因此，深度可分离卷积总的计算复杂度为 $HWNK^2+HWNM$，是标准卷积的 $\frac{1}{M}+\frac{1}{K^2}$ 倍，因为网络结构中 $M\gg K^2$。MobileNet 就是基于此设计的网络结构，极大减小了模型规模，降低了计算复杂度。

1.2 基于压缩参数的模型加速方法

压缩参数的方法分为四种，网络剪枝、模型量化、低秩估计和模型蒸馏。基于网络修剪的方法侧重于探索模型参数冗余部分，并尝试去除冗余和不重要的参数。基于模型的量化方法关注参数的最佳表示，并降低参数的复杂性。基于低秩估计的方法使用矩阵 / 张量分解，估计深 CNN 中的信息量参数。模型蒸馏方法学习提炼一个新的模型，该模型结构更加简单、紧凑，可再现大型网络的输出 [2]。

1.2.1 网络剪枝

裁剪训练好的模型是模型压缩中使用最广泛的方法，通常根据有效的评判手段，判断参数的重要性，裁剪不重要的网络连接或者卷积核连接进行，减少模型冗余。网络剪枝一般通过三个步骤进行训练，如图 1 所示。通过常规网络训练方法训练，但学习的结果不是最终模型参数的权重，而是学习哪些连接更具有权重。其次，对网络进行剪枝，去掉权重较低的连接。设置阈值，删除权重较低的连接，降低网络复杂度，将稠密网络转换为稀疏网络，如图 2 所示。最后，重新训练网络剩余稀疏连接的最终权重。在不断稀疏的网络参数中学习正确的连接是一个迭代过程。许多轮迭代后，可以找到最小数目的连接。修剪后的网络必须经过重新训练，否则准确性会受到显著影响。

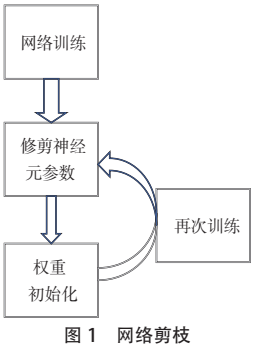

图 1 网络剪枝

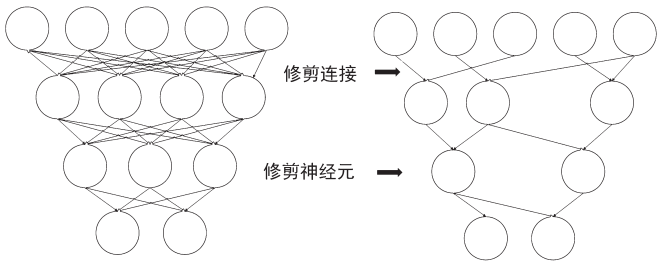

图 2 剪枝前网络和剪枝后网络

1.2.2 模型量化

模型量化着眼于参数本身，没有改变模型的计算量，主要有两个研究方向，权值共享和权值精简。

权值共享的基本思想是多个网络连接的权重共用一个权值。其方法是对每一层权重矩阵利用 K-means 聚类算法聚类成若干个簇，使用 $\arg\min\sum_{i=1}^{k}\sum_{w\in c_i}|w-c_i|$ 计算得到每个聚类中心值，代表该 cluster 的权重。由于同一簇的权重共享一个权重大小，因此只需存储权值的簇的索引值即可。通过查表获取该索引值对应的权重大小，为了减少精度损失，再通过训练微调对权重进行补偿：

$$\frac{\partial l}{\partial c_k}=\sum_{i,j}\frac{\partial l}{\partial w_{i,j}}\cdot\frac{\partial w_{i,j}}{\partial c_k}=\sum_{i,j}\frac{\partial l}{\partial w_{i,j}}\left(I_{i,j}=k\right) \quad (1)$$

式中，$\frac{\partial l}{\partial c_k}$ 表示参数微调的损失函数，$\frac{\partial l}{\partial W_{i,j}}$ 为损失函数对权重的偏导值，$\frac{\partial W_{i,j}}{\partial c_k}$ 为权重对每一个聚类中心的偏导，当属于 $I_{i,j}=k$ 时，结果为 1，不属于结果为 0。

所有的梯度信息按照权重矩阵之前的分组进行。权值精简的思想十分简单。网络模型中的权重和偏移量，都是用单精度 4 字节的 32 位浮点数或者双精度 8 字节的 64 位浮点数表示，转化为更低位的整数运算，常见的是 8 比特的整形数运算或 1 比特布尔运算。以 8 比特整形数为例，记录当前参数的最大值和最小值，设为 max$i$、min$i$，p 为压缩大小的超参数，a、b 为压缩后的最大最小值。则所有参数量化结果为：

$$p_q=\frac{2^8-1}{\max i-\min i}\times\left(p-\min i\right)=a\left(p-b\right) \quad (2)$$

通过降低参数精度降低存储空间。因为网络模型经过训练后，对噪声和较小的扰动具有鲁棒性，故此方法十分实用 [3]。

1.2.3 低秩估计

低秩估计的方法分为矩阵分解和张量分解。矩阵分解主要采用 SVD 和 PCA 方法。基于矩阵分解的方法通过运用矩阵分解和矩阵乘法结合律，加快运算速度。两个维数分别为 $a\times b$ 的 A 矩阵和 $b\times c$ 的 B 矩阵相乘，每个向量相乘所需要的计算次数为 b 次乘法，b-1 次加法，共 $a\times(2b\text{-}1)\times c$ 次运算。因为加法速度远快于乘法的速度，故计算量可简化为 $a\times b\times c$。通过 SVD 分解的方法，$B=U\Lambda V^{\mathrm{T}}$，故：

$$A\times B=A(U\Lambda V^{\mathrm{T}})=(AU)\Lambda V^{\mathrm{T}} \quad (3)$$

其中，U 是通过 SVD 从 B 中分解而来。基于张量分解的方法主要有 CP 和 Tucker。Tucker 方法将原卷积分解为 3

个计算量更小的卷积，对于输入为 $W_{in}H_{in}C_{in}$ 的向量，通过卷积核 $C_{in}K_HK_WC_{out}$ 运算，总计算量为：

$$N_{conv}=H_{out}W_{out}C_{out}K_HK_WC_{in} \quad (4)$$

将卷积核分成三组大小分别为 $C_{in}\times1\times1\times R_3$、$R_3\times K_H\times K_W\times R_4$、$R_4\times1\times1\times C_{out}$ 的卷积操作，R 为压缩通道数，则计算量为：

$$N_{Tucker}=H_{out}\times C_{in}\times(R_3\times C_{in}+R_3\times R_4\times K_H\times K_W+R_3\times C_{out}) \quad (5)$$

低秩估计方法的优点是基础运算的结构没有变化，不需要定义额外的新操作，分解后的网络仍由卷积运算实现，应用范围更广。有许多分解方法适用于较多场景，但与网络剪枝相同，需在优化后微调网络，以保证网络的最终效果 [4-5]。

1.2.4 模型蒸馏

模型蒸馏的方法不同于之前的三种方法，其通过构建新的小网络达到加速的效果。其思想是先训练一个结构复杂、计算量大但准确率较高的模型，之后通过大网络模型的计算结果训练一个结构简单的小网络。复杂网络可看为老师，精简网络可看为学生，老师经过长时间的“训练”总结出一套适用于某项任务的方法，之后将该方法提炼成“知识”传授给学生，帮助学生更快学会处理类似任务。模型蒸馏思想主要依据以下三点：（1）大规模神经网络在保证训练准确性的前提下，提炼出对应问题的数据结构的相似性；（2）基于先验知识的小网络，只需 , 很少的迭代次数和训练样本即可快速收敛；（3）Softmax 分类函数的结果随着温度变量的升高分布更均匀：

$$q_i=\frac{\exp(z_i/T)}{\sum_j\exp(z_i/T)} \quad (6)$$

其中，T 为温度参数，q_i 是第 i 类的概率值，z_i 为对应的结果。整体结构如图 3 所示。

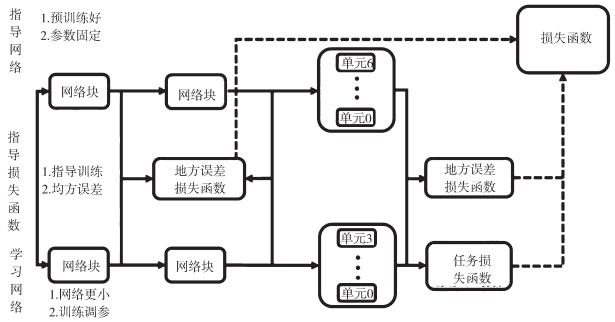

图 3 模型蒸馏

第一，为了使神经网络能够产生更均匀分布的软目标，使用较大的温度参数训练模型。第二，小规模的神经网络基于同样的温度参数，训练大网络产生的目标结果，并通过迭代学习使小网络逐渐接近大网络的软目标，输出相同的结果。第三，温度参数逐渐减小至 1，完成模型训练，使类别概率偏向正确类别。

2 结 语

基于卷积神经网络的计算机视觉任务，如物体检测、图像识别等任务，相关研究已展示了效果。但是，在普及的移动设备上，由于硬件限制，想要实现相同效果，必须通过网络模型加速方法进行加速，使好的基础模型可以在移动设备上高效运行。

第一，压缩结构设计领域，将普通卷积转变为计算量低的卷积操作，具有很好的效果。MobileNet、ShuffleNet 等网络结构，不仅进一步降低了计算复杂度，在准确度方面也能达到甚至超过 VGG-16。这说明进行模型轻量化不一定以牺牲准确度为代价，降低计算复杂度、减少参数量的同时，模型准确度能得到提升。

第二，压缩参数领域，网络剪枝方法简单，压缩效果可控，但精细的剪枝需要平台的额外支持。低秩估计十分灵活，在各个平台均可实现，但压缩效果一般。参数量化的方法基本适用于所有场景，且操作十分简单。模型蒸馏的效果很好，可以提升小网络的准确性，但调参困难。

实际应用中，这些方法可以相互融合，针对特定网络结构分析特点，处理一个网络。对于深层神经网络，可以分层使用网络剪枝和低秩估计，且已达到最佳平衡。卷积网络模型的加速的研究刚刚开始，还有许多值得探索的方法有待研究。

参考文献

[1]He Y,Zhang X,Sun J.Channel Pruning for Accelerating Very Deep Neural Networks[C].// International Conference on Computer Vision,2017.

[2] 陈伟杰 . 卷积神经网络的加速及压缩 [D]. 广州 : 华南理工大学 ,2017:81.

[3]Qin Z,Zhang Z,Chen X,et al.FD-MobileNet:Improved MobileNet with a Fast Downsampling Strategy[C].// 2018 25th IEEE International Conference on Image Processing,2018.

[4] 毕鹏程 , 罗健欣 , 陈卫卫 . 轻量化卷积神经网络技术研究 [J]. 计算机工程与应用 ,2019(10):14.

[5]Hinton G,Vinyals O,Dean J.Distilling the Knowledge in a Neural Network[J].Computer Science,2015,14(7):38-39.