MobileNet网络的改进

黄萱昆[1]，董远[1]，白洪亮[2]

[1] 北京邮电大学信息与通信工程学院，北京 100876

[2] 北京飞搜科技有限公司，北京 100082

摘要：基于卷积神经网络的模型加速在是近几年是一个研究热点，本文首先介绍了目前常用的模型加速的方法以及这些方法之间的区别，另外介绍了谷歌公司设计的一个轻量级网络MobileNet，采用通道分离卷积作为主要结构，极大的降低计算复杂度和参数量，能够灵活用于移动端的各类视觉应用中。在此基础上，本文提出了一种新的改进原始MobileNet网络的方法，通过在通道分离卷积结构间引入残差学习连接，在不额外增加网络计算复杂度和参数量的情况下，提升网络收敛效果，从而进一步提高网络准确率，并最终在Imagenet目标分类数据集上验证了这一方法的有效性。

关键词：人工智能，机器学习，卷积神经网络，模型加速，MobileNet网络

中图分类号： TP391.41

The improvements of MobileNet network

HUANG Xuan-Kun[1], DONG Yuan[1], BAI Hong-Liang[2]

[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876

[2] Beijing FaceALL Technology Ltd., Beijing 100082

Abstract: In recent years, model acceleration based on convolutional neural network has been a research hotspot. This paper firstly introduces the current commonly used methods in model acceleration and the difference between these methods, and also introduces a lightweight network structure designed by Google company called MobileNet, which is mainly consisted of depthwise separable convolution layers and can be used in mobile vision applications by greatly reducing the computational complexity and the amount of parameters of the model. Therefore, this paper proposes some methods to improve the MobileNet network efficiently. By calling for the addition of some residual connections in the depthwise separable convolution structure without further increase the complexity and the amount of parameters of the model, these methods can dramatically imporve the convergence effect and thus enhance the accuracy of the MobileNet. Finally, the effectiveness of these methods is verified in the Imagenet object classification dataset.

Key words: Artificial Intelligence, Machine Learning, Convolutional Neural Network, Model Acceleration, MobileNet Network

---

基金项目： 国家自然科学基金重点项目(61532018)

作者简介： 黄萱昆（1993-），男，硕士研究生，主要研究方向：深度学习理论，邮箱：mumaren@bupt.edu.cn。通信作者：董远（1970-），男，教授，主要研究方向：机器学习、图像与视频搜索，邮箱：yuandong@bupt.edu.cn。

0 引言

在2012年的ILSVRC的分类任务上，多伦多大学的Hinton团队利用卷积神经网络提出的AlexNet [1]结构赢得了第一名的成绩，极大的推动了深度学习技术的发展。近年来，为了达到更高的识别准确率，多层卷积神经网络结构开始变得越来越深，模型复杂度也越来越高，如GoogLeNet [2]、VGG [3]和残差网络[4]等，但是这样带来的是模型存储开销的增大和模型计算时间的增加，使得很难把一个深度学习模型应用到计算和存储资源都有限的移动端设备中，因此，合理的设计一个网络结构，通过高效的参数利用使得计算量和模型存储能够得到极大的降低，同时又保持模型识别精度尽可能的高。

因此在工业实践中，基于卷积神经网络的模型加速研究变得尤为重要，近年来的很多研究也都在致力于构建一个轻量级并且高效的卷积神经网络，从而降低深度模型对硬件资源（内存、CPU、硬盘）的开销。很多不同的模型加速方法可以大致分为两种，一种是在预训练好的模型基础上压缩和加速模型，另一种是直接设计并训练一个较小的网络。

因此，本论文的主要贡献如下：

- 本论文介绍了目前主流的一些基于卷积神经网络模型加速的方法，极大的降低模型参数存储和模型计算复杂度，并介绍了谷歌新提出的轻量级MobileNet [5]网络，该网络与其他主流的深度网络具有大致相同的精度表现，但是在计算复杂度和参数量上具有极大的优势。
- 本论文提出了一种有效改进原有MobileNet结构的方法，通过在通道分离卷积模块间加入跳跃连接，使得连接区间内的层去学习输入与输出的残差函数，降低单个模块学习难度，从而有效的提升原有结构的精度表现。
- 最后，本论文在Imagenet [6]数据集上做实验，验证了该方法能有效的提升原网络的准确率，并且不会带来模型计算量和参数量的增加。

1 相关工作

1.1 模型加速思路

近年来，在基于卷积神经网络的模型压缩和加速的研究方向上涌现了很多新的方法，这些方法大致可以归为两类，一种是构建一个全新的小规模网络结构，另一种是在现有深度模型的基础上用一些特定策略去压缩模型参数和降低计算复杂度来获得模型的提速。

Flattened networks [7]通过对三维卷积核进行每个维度上的完全分解，构建了一个高效扁平化的网络结构。与此类似，Factorized Networks [8]介绍了类似的卷积层分解方法，并加入拓扑连接的优化，其设计的分类网络精度与GoogLeNet [2]和VGG-16 [3]相当，但是模型参数和计算量大幅减少。SqueezeNet [9]主要通过加入1*1卷积层的降维作用来减少3*3卷积层的计算量，简化网络复杂度，其设计的整体结构印证了即使很小的CNN网络也能达到很好的识别精度。类似的，MobileNet [5]通过引入通道分离卷积结构来构建了一个轻量级且高效的网络。除

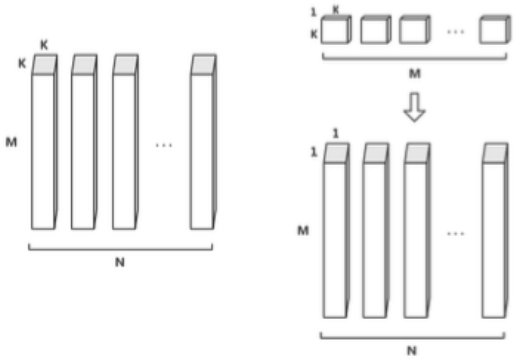

图 1: 左侧为标准卷积，右侧为通道分离卷积模块

Fig. 1: Left: Standard convolution filers. Right: Depthwise separable convolution block

此之外，运用迁移学习原理的蒸馏模型[10]，通过采用预先训练好的复杂模型（老师模型）的输出作为监督信号去训练另外一个简单的网络（学生模型），这样得到的学生模型具有与老师模型类似的表现能力，同时复杂度大大降低，实现了模型的压缩和提速。

由于深度模型参数量庞大，且具有较大的信息参数冗余，因此还可以通过一定的技术方法对已有的复杂模型进行去冗余。一种是权重值的正则化，具体包括权重层的剪枝、权值量化、二值化或多值化编码等。另外一种就是基于张量的低秩分解，即认为一个低秩的多维矩阵可以被分解为多个矩阵的形式，具体包括Low-Rank分解和CP分解等。

1.2 MobileNet网络

MobileNet [5]网络是由谷歌公司于2017年提出的一个轻量级的卷积神经网络结构，其最主要的模块结构叫做通道分离卷积。具体如图1所示，左边为标准的卷积层，每个卷积核尺寸为K*K*M；右边为通道分离卷积模块，第一级卷积层每个卷积核尺寸为K*K*1，卷积核数量M与输入通道数相等，第二级每个卷积核为1*1*M，输出通道数为N。通过对第一个K*K卷积层按输入通道数进行分组卷积，再级联一个1*1的卷积层，即将传统的卷积层拆分成两个级联的卷积，组合起来很好的逼近普通的三维卷积，计算量主要集中分布在1*1卷积层，极大的降低了K*K卷积层的计算量和参数量。整个网络的准确率与经典分类网络GoogLeNet [2]和VGG-16 [3]保持在同一水准，但是在计算量和参数量上，该网络比GoogLeNet小大概2.5倍，比VGG-16小大概30倍，优势显著。

1.3 深度残差网络

在2015年由何凯明团队提出的深度残差网络[4]一举获得了ILSVRC2015竞赛各项任务的冠军，将神经网络的深度又提高了一个量级。其主要贡献在于提出了一个用来训练非常深的深度网络又十分简洁优雅的框架，通过加入恒等连接，如图2，原始卷积层模块的输出由F(x)改写为F(x)+x，使得每层去学习除了本征外的残差信息，降低每层的学习难度，从而对于即使很

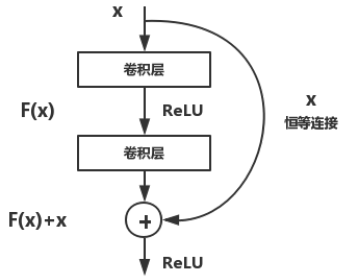

图 2: 残差学习模块
Fig. 2: Residual learning block

深的网络也能得到很好的优化，在一定程度上增加网络深度的同时也能获得较高的精准度。

2 改进MobileNet网络

本论文主要目的在于充分利用原始MobileNet [5]网络的参数，进一步提升原网络的识别准精度，同时维持原网络的计算复杂度和参数量不变。在这一章节，本论文主要从以下三方面来改进MobileNet网络：

1. 对于网络中一部分输出维度大小与输入不变的通道分离卷积模块（DW模块），添加一条简单的跳跃连接，即恒等连接。
2. 对于输出维度大小有变化的DW模块，添加一条带有1*1卷积的跳跃连接，称作复杂连接，控制输出的空间大小和通道数与原模块相同。
3. 进一步研究在加入了跳跃连接的DW模块中，不同的激活函数使用方式对模型准确率的影响，并选取最优的一种。

2.1 跳跃连接

MobileNet [5]网络由13个DW模块串联堆叠起来，整体流线型的架构与VGG网络类似，依据深度残差网络的思想，对每个DW模块的输入与输出间加入一条跳跃连接来优化原始的结构。这里本论文把DW模块中的3*3通道分离卷积和级联的1*1卷积作为一个整体，不考虑这两个卷积层各自的跳跃连接。

对于加入了跳跃连接的DW模块，可以用公式表示如下：

$$y_i = h(x_i) + F(x_i, w_i) \tag{1}$$

$$x_{i+1} = f(y_i) \tag{2}$$

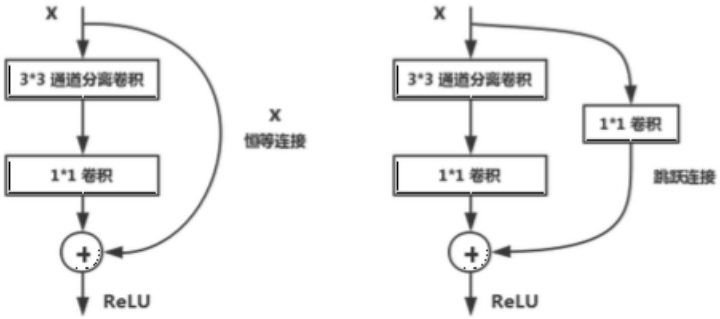

图 3: 改进后的通道分离模块

Fig. 3: The improved depthwise separable block

其中，x_{i+1}和x_i为第i个模块的输出与输入值，F为模块中两个卷积层的线性变换函数，f为ReLU非线性激活函数，$h(x_i)$为跳跃连接函数。因此，当输出与输入数据维度大小相同时，令$h(x_i) = x_i$, 即输出直连输入，为恒等连接，如图3左侧；当输出与输入数据维度不同时，令$h(x_i) = g(x_i, w_i)$，这里$g(x_i, w_i)$表示通过一个1*1卷积层，控制跳跃连接的输出通道数和空间分辨率与模块输出相等，如图3右侧。

得益于跳跃连接在深度残差网络中出色的性能表现，本论文认为对于MobileNet网络，适当加入一些跳跃连接也能起到一定效果，通过迫使每层只去学习一些输出与输入间的残差信息，降低每层的学习难度，同时也能避免深度网络在梯度后馈回传的过程中消失的问题，有利于网络的快速优化和最终的精度表现。

2.2 激活函数的影响

不同的激活函数使用方式会对网络产生不同的正则化约束效果，同时对网络的优化学习也有一定的影响，这里的激活函数使用方式指BatchNorm [11]层或ReLU层在DW模块中的不同摆放位置。为了探究其对MobileNet的影响，对于带有恒等连接的这些模块，依据ResNet-v2 [12]的分析，我们只尝试两种可能方式：原激活方式和前激活方式，见图4。

2.3 整体结构

原始MobileNet [5]网络由大量的通道分离卷积模块堆叠而成，本论文只是对每个模块做了改进，因此改进后的网络整体与原始的基本一致。根据上述针对跳跃连接和激活函数的使用方式两方面的改进策略，我们得到四种新网络:MobileNet-simple-connect、MobileNet-full-connect、MobileNet-simple-connect-pre-act和MobileNet-full-connect-pre-act，详细的网络结构定义见表1。

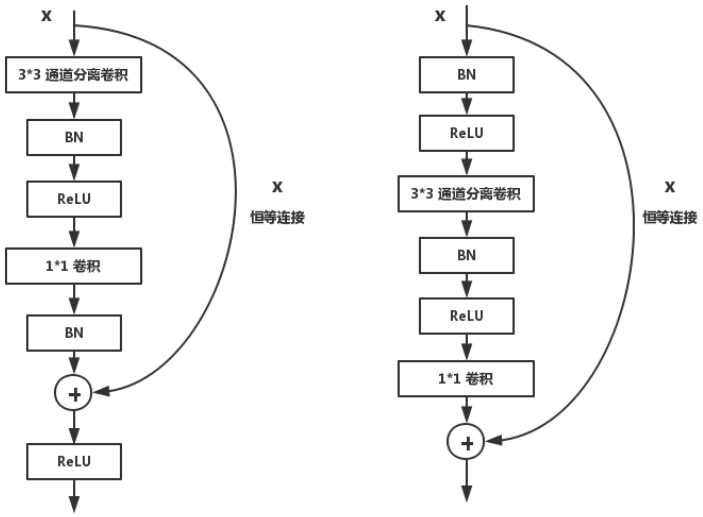

图 4: 不同的激活函数使用方式，左侧为原激活方式，右侧为前激活方式

Fig. 4: Various usages of activation. Left: Original activation. Right: Pre-activation

表 1: 改进后的MobileNet网络结构

Tab. 1: The improved Mobilenet body architecture

层名称/类型	输出尺度	跳跃连接(直连/1*1卷积)			
Input data	224*224*3	-	-	-	-
Conv_1	112*112*32	-	-	-	-
Conv_2 DW	112*112*64	-	1*1卷积	-	1*1卷积
Conv_3 DW	56*56*128	-	1*1卷积	-	1*1卷积
Conv_4 DW	56*56*128	直连	直连	直连	直连
Conv_5 DW	28*28*256	-	1*1卷积	-	1*1卷积
Conv_6 DW	28*28*256	直连	直连	直连	直连
Conv_7 DW	14*14*512	-	1*1卷积	-	1*1卷积
5 * Conv DW	14*14*512	直连	直连	直连	直连
Conv_13 DW	7*7*1024	-	1*1卷积	-	1*1卷积
Conv_14 DW	7*7*1024	直连	直连	直连	直连
Global Pool	1*1*1024	-	-	-	-
FC	1*1*1000	-	-	-	-
Softmax	1*1*1000	-	-	-	-
激活方式		原始激活		前激活	
MobileNet		simple-connect	full-connect	simple-connect-pre-act	full-connect-pre-act

表 2: 不同的跳跃连接方式

Tab. 2: Various usages of skip connection

模型	Top-1准确率	计算量	参数量
原始1.0 MobileNet-224	70.6%	569M	4.2M
1.0 MobileNet-224-simple-connect	72.3%	569M	4.2M
1.0 MobileNet-224-full-connect	72.7%	698M	4.9M

3 实验结果

为了实验方便，本文以原始的1.0 MobileNet-224 [5]网络作为基准，改进后的网络与原网络保持相同的网络深度和宽度，通过控制变量来验证改进方法的合理性。

3.1 训练配置

本文的改进的MobileNet [5]网络训练均是基于MxNet [13]分布式深度学习框架，采用两块英伟达GTX1080显卡，每块显卡批尺寸（Batch Size）设为32，采用随机梯度下降法训练100个Epoch。本文只在Imagenet [6] 1000类目标识别数据集上实验，由于该数据集较大（训练集128万张图片，验证集5万张图片），对于小网络不容易产生过拟合，所以在训练时较少地使用正则化约束和数据增强等策略，网络优化参数基本与Inception-v3 [14]一致。在测试时，本文使用Imagenet验证集，对每张图片中心裁剪到224*224大小，前馈输入到单模型网络得到分类结果。

3.2 跳跃连接的影响

本文对原网络中大量使用的DW模块加入跳跃连接，这里分两种连接方式：1*1卷积连接和恒等连接两种，具体实现见表1。按照上述训练配置方法训练得到改进后的网络，在Imagenet数据集上测试得到结果，见表2。

通过与原网络对比可以发现,对于仅加入了简单的恒等连接的网络1.0 MobileNet-224-simple-connect，准确率较原始网络提升了1.7%，且维持原网络的计算量和参数量不变；对于同时加入了恒等连接和1*1卷积连接的网络1.0 MobileNet-224-full-connect，准确率较原始网络提升2.1%，且比1.0 MobileNet-224-simple-connect网络继续提升了0.4%，但是带来的是网络计算量增加22.6%，权重参数增加14.3%。

3.3 激活函数的影响

不同的激活函数使用方式会对网络的优化训练以及最终的准确率产生一定的影响，前激活方式是指ReLU层和BatchNorm [11]层在卷积层之前，而表1中的三个模型均为原始激活方式，即BatchNorm层接在卷积层之后，ReLU层接BatchNorm层之后。

表 3: 不同的激活函数使用方式

Tab. 3: Various usages of activation

模型	Top-1准确率	计算量	参数量
原始1.0 MobileNet-224	70.6%	569M	4.2M
1.0 MobileNet-224-simple-connect	72.3%	569M	4.2M
1.0 MobileNet-224-simple-connect-pre-act	72.6%	569M	4.2M
1.0 MobileNet-224-full-connect	72.7%	698M	4.9M
1.0 MobileNet-224-full-connect-pre-act	72.9%	698M	4.9M

表 4: 改进的MobileNet网络与常见网络对比

Tab. 4: The improved MobileNet comparison to popular models

模型	Top-1准确率	计算量	参数量
原始1.0 MobileNet-224	70.6%	569M	4.2M
1.0 MobileNet-224-simple-connect-pre-act	72.6%	569M	4.2M
1.0 MobileNet-224-full-connect-pre-act	72.9%	698M	4.9M
GoogLeNet	69.8%	1550M	6.8M
VGG16	71.5%	15300M	138M

本节对加入了不同连接方式的两个网络1.0 MobileNet-224-simple-connect和1.0 MobileNet-224-full-connect改用前激活方式进行实验，结果见表3。由实验结果可以看出，将原激活方式改为前激活方式后，准确率均有进一步提升，目前性能表现最好的1.0 MobileNet-224-full-connect-pre-act网络准确率到达了72.9%，比同样采用前激活方式的提升0.2%。

因此，相比于原始激活方式，前激活方式有效的保证了后馈梯度在两个带有恒等连接的DW模块间更直接的传递，在网络训练过程中更容易学习优化参数，也能在一定程度上减少过拟合，使得网络的测试准确率得到稳定的提升。

表4为将改进后的网络与原始的MobileNet [5]、VGG16 [3]以及GoogLeNet [2]作对比。对于改进后的1.0 MobileNet-224-simple-connect-pre-act网络相比于VGG16，参数量小约32倍，计算量小约27倍，计算量相比GoogLeNet也小约2.5倍，在准确率上仍具有一定的优势和提升。

4 结论与展望

本论文首先介绍了目前基于深度神经网络的模型加速的一些主流方法，并在此基础上介绍了一款近期提出的轻量级的小网络MobileNet，它在大幅降低了模型参数和计算量的同时，维持了和目前主流的一些深度网络的精度水准,效果显著。本论文以MobileNet网络为基准，提出了一些改进的策略，在不改变原网络深度和宽度、参数和计算量大小的同时，进一步提升原网络的识别精度。最后在Imagenet大规模数据集上，通过对比原网络，验证了这些改进策略的合

理性和有效性。

本文仅提供了一些改进MobileNet网络的思路和初步尝试，相信会有一些更优的策略，能进一步提升原网络的精度，或者可以尝试改进网络中的主要结构通道分离卷积模块，优化结构，降低计算量，达到进一步加速的效果。

5 致谢

特别感谢导师董远教授和白洪亮博士对本研究的支持与指导，并对审稿专家表示最诚挚的谢意。

参考文献（References）

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 1097-1105, 2012.

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-9, 2015.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778, 2016.

[5] 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.

[6] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database [J]. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 248-255, 2009.

[7] Jin J, Dundar A, Culurciello E. Flattened convolutional neural networks for feedforward acceleration [J]. arXiv preprint arXiv:1412.5474, 2014.

[8] Wang M, Liu B, Foroosh H. Factorized Convolutional Neural Networks [J]. arXiv preprint arXiv:1608.04337, 2016.

[9] 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.

[10] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. arXiv preprint arXiv:1503.02531, 2015.

[11] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. arXiv preprint arXiv:1502.03167, 2015.

[12] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks [J]. European Conference on Computer Vision. Springer International Publishing, 630-645, 2016.

[13] Chen T, Li M, Li Y, et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems [J]. arXiv preprint arXiv:1512.01274, 2015.

[14] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818-2826, 2016.