计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):330−335 [doi: 10.15888/j.cnki.csa.007943]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

用于大规模图像识别的特深卷积网络①

李　荟, 王　梅

(东北石油大学 计算机与信息技术学院, 大庆 163318)
通讯作者: 李　荟, E-mail: lihui_dqpi@163.com

摘　要: 卷积网络深度对大规模图像识别的准确性有不可忽视的影响. 使用具有非常小 (3×3) 卷积滤波器的架构, 我们对深度不断增长的网络进行了全面评估. 通过将深度推到 16–19 重量层可以实现对现有技术配置的显着改进. 通过比对其他卷积滤波器架构的卷积网络, 我们验证了我们提出的网络对大规模图像识别的改进效果. 同时为了避免训练数据集内在的偏倚, 我们还使用了其他数据集对网络进行了验证, 在这些数据集中, 它们可以获得最先进的结果.

关键词: 深度学习; 卷积网络; 图像识别; 卷积滤波器

引用格式: 李荟,王梅.用于大规模图像识别的特深卷积网络.计算机系统应用,2021,30(9):330–335. http://www.c-s-a.org.cn/1003-3254/7943.html

Extra Deep Convolutional Networks for Large-Scale Image Recognition

LI Hui, WANG Mei

(College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: The convolutional network depth is crucial to accurate large-scale image recognition. In this work, we thoroughly evaluate the networks with increasing depth using the architecture with quite small (3×3) convolution filters. The prior-art configurations can be improved significantly after the depth is pushed to 16–19 weight layers. The comparison with the convolution networks of other convolution filter architectures verifies the effectiveness of the proposed network for large-scale image recognition. In addition, the network verification is conducted with some other data sets to avoid the inherent bias of training data sets. As a result, the most advanced results can be obtained from these data sets.

Key words: deep learning; convolutional network; image recognition; convolutional filter

1　引言

伴随着大数据时代的到来以及各种强大的计算设备的发展, 深度学习是可以充分利用海量数据, 包括标注数据、弱标注数据和无标注数据类型, 对抽象的知识表达进行完全自动地学习. 目前, 深度学习改进了图像处理、语音处理和文本处理等众多领域的算法设计思想, 逐步形成了一套基于训练数据, 通过一种端到端的模型, 最后得到最终结果的思路, 这是一种既简单又高效的处理方式, 并且深受认同. 随着不断深入的研究与应用, 也出现了很多精良设计的深度网络结构, 可以解决传统机器学习并不好解决的很多复杂问题. 对比传统的机器学习, 深度学习的本质就是一种特征学习方法, 它把原始的数据经过简单并且非线性的模型转换成为层次更高的、更加抽象的表达. 虽然深度学习能够很好的建立起输入和输出之间的映射关系, 可是却不能较好地发现其内在物理联系. 相比应用研究来

① 基金项目: 国家自然科学基金 (51774090); 黑龙江省自然科学基金 (LH2019F004); 东北石油大学引导基金 (2020YDL-15)
Foundation item: National Natural Science Foundation of China (51774090); Natural Science Foundation of Heilongjiang Province (LH2019F004); Guiding Fund of Northeast Petroleum University (2020YDL-15)
收稿时间: 2020-10-06; 修改时间: 2020-10-30; 采用时间: 2020-11-12; csa 在线出版时间: 2021-09-02

说, 深度学习的理论研究可以做的更多.

近年来, 卷积神经网络受到了广泛关注, 其在大规模图像识别和视频识别领域都获得了较于以往非常好的效果. 随着大规模的图像库和高性能的 GPU 的发展, 识别率有了显著的提升. 值得一提的是以 ImageNet 为代表的大规模图像识别大赛, 其在深度学习领域有较强的推动作用[1], 此大赛已经成为了最近几年较大型实验平台.

目前, ConvNets 已经变成了图像识别的商品之一, 研究人员努力改进 Krizhevsky 等人创建的初始架构. 目的是为了提升其在 abid 中的准确率[2]. 本文我们主要讨论的是卷积神经网络设计时的一个非常重要的参数, 即网络的架构深度, 我们还尝试了架构中的其他参数的阈值, 试图加入更多的卷积层, 以此来保持增加网络深度的稳定, 并且证明了其可行性. 我们在所有的层里都采用了 verysmall (3×3) 的卷积滤波器, 还提出了一种基于 verysmall 卷积滤波器的架构模型, 通过实验证明, 该架构在多种图像识别数据集中都有较好的识别率[3,4], 可以为接下来的工作打下基础.

2 卷积网络配置

为了衡量卷积网络深度在公平环境中所带来的改进, 我们所有的卷积网络层配置均采用与 Krizhevsky 等人相同的设计原则.

2.1 架构

该卷积神经网络训练时的输入是 224×224 像素的 RGB 图像, 并且大小固定. 训练前的预处理过程是将训练数据集的 RGB 算出平均值, 再把每个像素都减去平均值完成预处理操作. 将图像经过一些卷积 (转换) 层, 包含 verysmall 接受模板的过滤器. 该模板是 3×3 的, 因为它是可以分出上下左右中心的最小的尺寸. 我们采用了 1×1 的卷积滤波器在其中的一种配置里, 其作用相当于对输入通道做线性变换 (马上再进行非线性变换). 将卷积步长设定成 1 个像素, 并且将空间层的输入设定为卷积操作后原有的分辨率, 也就是对 3×3 转换, 填充 1 个像素层[5]. 用 5 个最大池化层来填充空间层. 最大池化层的操作在 2×2 像素窗口上, 设定步长为 2. 通过上述的卷积神经网络后, 图像信号经过一叠三层的完全连接层, 即前两个 4096 个通道, 第 3 个有 1000 个通道, 架构的最后一层是 Softmax 层. 在任何神经网络中配置全连接层的方法都是相同的[6,7].

众所周知, 全部的隐藏层都具有非线性整流特征. 我们的网络中, 无一例外都没有局部响应归一化 (LRN) 层. 第 4 节中会介绍归一化层部分[8]. 这样架构会使得内存的消耗变多, 并且运行时间会增多, 但是不会改变数据集的性能[9].

2.2 配置

本文评估的卷积网络配置在表 1 中列出, 每列一个. 在下面的论述中, 我们通过他们的名称 (A–E) 来指代相应的网络. 所有配置均遵循第 2.1 节中提出的通用设计, 并且仅在深度上不同: 从网络 A 中的 11 个权重层—包含 8 个转换层和 3 个完全连接层到网络 E 中的 19 个权重层—包含 16 个转换层和 3 个完全连接层. 转换层的宽度 (通道数) 相当小, 从第一层中的 64 开始, 然后在每个最大化池层之后增加 2 倍, 直到达到 512[10].

表 1 卷积网络单尺度演化性能测试

网络配置	小图像		Top1错误率(%)	Top5错误率(%)
	训练	测试		
A	256	256	29.4	11.2
A-LRN	256	256	29.4	10.9
B	256	256	28.6	9.8
C	256	256	28.3	9.9
	384	384	27.7	8.5
	512	256	27.5	8.5
D	256	256	26.9	8.1
	384	384	26.3	9.1
	512	512	25.4	9.2
E	256	256	26.4	8.8
	384	384	25.8	8.4
	512	512	24.9	8.6

3 分类框架

3.1 训练

网络权重的初始化很重要, 因为糟糕的初始化可能会使深度网络中梯度的不稳定性停止学习. 为了避免这个问题, 我们开始训练配置 A (表 1), 其足够浅以便随机初始化进行训练. 然后, 在处理更深层次的体系结构时, 初始化了前 4 个卷积层, 最后 3 个完全连接层用网 A 初始化 (中间层随机初始化). 预初始化层的学习率没有调低, 并允许它们在训练期间改变. 对于随机初始化 (如果适用), 从具有零均值和 10^2 方差的正态分布中对权重进行采样[11]. 偏差初始化为零. 我们发现使用文献 [12] 的随机初始化程序可以在没有预训练的情况下初始化权重.

为了获得固定大小的 224×224 卷积网络输入图像, 从重新缩放的训练图像中随机裁剪它们 (每个 SGD 迭代每个图像一个裁剪). 为了进一步增加训练集, crop 经历了随机水平翻转和随机 RGB 色移操作[7]. 下面解释了训练图像重新缩放.

训练图像尺寸: 设 S 是各向同性重新缩放的训练图像的最小边, 从中裁剪出卷积网络输入 (我们也将 S 称为训练比例). 虽然作物大小固定为 224×224, 但原则上 S 可以采用不小于 224 的任何值: 对于 S=224, 作物捕获全图像统计, 完全跨越训练图像的最小侧; 当 S>>224 时裁剪将对应于图像的一小部分, 包含一个小对象或一个对象部分.

我们考虑两种设置训练量表 S 的方法. 第一种是修正 S, 它对应于单一规模的训练 (注意, 采样作物中的图像内容仍然可以代表多尺度图像统计)[12]. 在我们的实验中, 我们评估了在两个固定尺度下训练的模型: S=256 (已在现有技术中广泛使用[7,13,14]) 和 S 在给定卷积网络配置的情况下, 我们首先使用 S=256 训练网络. 为了加速 S=384 网络的训练, 使用 S=256 预训练的权重对其进行初始化, 并且我们使用较小的初始学习率 10^{-3}.

设置 S 的第 2 种方法是多尺度训练, 其中通过从特定范围 [S_{min}, S_{max}](我们使用 S_{min}=256 和 S_{max}=512) 随机采样 S 来单独地缩放每个训练图像[15]. 由于图像中的物体可以具有不同的尺寸, 因此在训练期间考虑这一点是有益的. 这也可以被视为通过尺度抖动的训练集增强, 其中训练单个模型以识别各种尺度上的对象. 出于速度原因, 我们通过使用相同的配置微调单尺度模型的所有层来训练多尺度模型, 使用固定的 S=384 进行预训练.

3.2 测试

在测试时, 给定经过训练的卷积网络和输入图像, 按以下方式对其进行分类. 首先, 它被各向同性地重新缩放到预定义的最小图像侧, 表示为 Q (我们也将其称为测试标度). 我们注意到 Q 不一定等于训练量 S (正如我们将在第 4 节中所示, 每个 S 使用几个 Q 值可以提高性能). 然后, 以类似于文献 [14] 的方式将网络密集地施加在重新缩放的测试图像上. 即, 首先将完全连接的层转换为卷积层 (第一 FC 层到 7×7conv. 层, 最后两个 FC 层到 1×1 转换层). 然后将得到的完全卷积网应用于整个 (未剪切的) 图像. 结果是一个类得分图, 其中通道数等于类的数量, 并且可变空间分辨率取决于 inputimage 大小. 最后, 为了获得图像的类别得分的固定大小的矢量, 类别得分图被空间平均 (求和). 我们还通过水平翻转图像来增加测试集; 对原始图像和翻转图像的 Softmax 类后验进行平均以获得图像的最终分数[16,17].

3.3 实现

我们的实现源自公开的 C ++ Caffe 工具箱, 但包含许多重要的修改, 允许我们对安装在单个系统中的多个 GPU 进行训练和评估, 以及训练并在多个尺度上评估完整尺寸 (未剪切的) 图像. 多 GPU 训练利用数据并行性, 并通过将每批训练图像分成几个 GPUbatches 来执行, 并在每个 GPU 上并行处理. 在计算 GPU 批量梯度之后, 他们被平均获得完整批次的梯度[18]. 梯度计算在 GPU 之间是同步的, 因此结果与在单个 GPU 上训练时的结果完全相同.

虽然最近提出了加速卷积网络培训的更复杂方法, 采用了网络不同层次的模型和数据并行性, 但我们发现概念上更简单的方案已经提供了 3.75倍的加速与使用单个 GPU 相比, 现成的 4-GPU 系统. 在配备有 4 个 NVIDIA Titan Black GPU 的系统上, 根据架构的不同, 训练一个网络需要 14~21 天[19,20].

4 分类实验

数据集: 在本节中, 我们将介绍由描述的 ConcevNet 架构在 ILSVRC-2012 数据集上实现的图像分类结果. 该数据集包括 1000 个类的图像, 并分为 3 组: 训练组 (1.3 M 图像数据), 验证组 (50 K 图像数据) 和测试组 (具有保持类标签的 100 K 图像数据). 使用两个度量评估分类性能: top-1 和 top-5 错误. 前者是多级分类错误, 即错误分类图像的比例; 后者是 ILSVRC 中使用的主要评估标准, 并且计算为图像的比例, 使得地面实况类别在 前 5 个预测类别之外.

4.1 单尺度演化

我们首先使用 Sect 中描述的层配置, 以单一规模评估各个卷积网络模型的性能[21]. 测试图像尺寸设定如下: Q=S 表示固定 S, Q=0.5(S_{min}+S_{max}) 表示抖动 S∈[S_{min}, S_{max}]. 结果如表 1 所示.

首先, 我们注意到使用本地响应规范化 (A-LRN 网络) 并没有改进没有任何规范化层的模型 A. 因此, 我们不在深层结构 (B~E) 中采用标准化.

其次, 我们观察到分类误差随着卷积网络深度的增加而减小: 从 A 中的 11 层到 E 中的 19 层. 值得注意的是, 尽管深度相同, 但配置 C (包含 3 个 1×1 转换层) 的性能更差比配置 D, 它在整个网络中使用 3×3 转换层. 这表明虽然附加的非线性确实有帮助 (C 比 B 更好), 但使用 conv 捕获空间上下文也很重要. 具有非平凡接收字段的过滤器 (D 优于 C). 当深度达到 19 层时, 我们的架构的错误率会饱和, 但更深的模型可能对更大的数据集有益. 我们还将净 B 与浅网进行了比较, 其中 5 个为 5×5 转换. 通过 3×3 转换的替换对来自 B 的层. 具有单个 5×5 转换的层. 层是指在 2.3 节中具有相同的感受区域. 在中心作物上, 测量浅网的前 1 个误差比 B 的高 1%, 这证实了具有小过滤器的深网优于具有较大过滤器的浅网.

最后, 即使在测试时使用单个尺度, 在训练时间 ($S\in[256; 512]$) 的尺度抖动导致对具有固定最小边 (S=256 或 S=384) 的图像的显着更好的结果. 这证实了通过尺度抖动的训练集增加确实有助于捕获多尺度图像统计.

4.2 多尺度演化

在单一规模评估卷积网络模型后, 评估规模抖动对时间的影响, 包括在测试图像的几个重新缩放版本上运行模型 (对应于不同的 Q 值), 然后对得到的类后验进行平均[22]. 考虑到训练和测试量表之间的巨大差异导致性能下降, 在 3 个测试图像大小上评估具有固定 S 的模型, 接近训练一: $Q=\{S-32, S, S+32\}$. 同时, 在训练时刻度抖动允许网络在测试时应用于更宽范围的尺度, 因此模型训练变量 $S\in[S_{min}, S_{max}]$, 结果如表 2 所示.

表 2 卷积网络多尺度演化性能测试

网络配置	小图像		Top1错误率(%)	Top5错误率(%)
	训练	测试		
A	256	256	29.1	10.2
A-LRN	256	256	28.4	10.4
B	256	256	28.9	9.7
C	256	256	29.3	9.5
	384	384	28.7	8.9
	512	384	27.9	8.9
D	384	384	27.6	8.2
	512	512	27.3	9.2
	512	512	26.4	9.1
E	256	384	27.4	8.7
	384	512	26.8	8.5
	512	512	25.9	8.6

表 2 中显示的结果表明, 在测试时刻度的抖动导致更好的性能 (与在单一规模上评估相同模型相比, 如表 3 所示). 与以前一样, 最深的配置 (D 和 E) 表现最佳, 并且比例抖动优于使用固定最小边 S 的训练. 我们在验证集上的最佳 top-1 / top-5 是 24.8%、7.5% 错误. 在测试集上, 配置 E 达到 7.3% 的前 5 个错误.

4.3 Multi-crop 演化

在表 3 中, 我们将密集的卷积网络评估与多作物评估进行比较. 我们还通过平均其软最大输出来评估两种评估技术的互补性[23]. 可以看出, 使用多种作物的表现略好于密集评价, 这两种方法确实是互补的, 因为它们的组合优于每一种. 如上所述, 我们假设这是由于对卷积边界条件的不同处理.

表 3 网络演化方法比较

网络配置(参照表1)	演化方法	Top1错误率(%)	Top5错误率(%)
C	密集	24.7	7.6
	Multi-crop	24.4	7.4
	Multi-crop&密集	24.3	7.5
D	密集	24.6	7.2
	Multi-crop	24.2	7.1
	Multi-crop&密集	24.1	7.1

4.4 卷积网络混合演化

到目前为止, 我们评估了各个卷积网络模型的性能. 在这部分实验中, 我们将几个模型的输出结合起来, 通过平均它们的 Softmax 类后验. 由于模型的互补性, 这提高了性能, 并且在 2012 年和 2013 年的顶级 ILSVRCs 提交中使用[24]. 结果显示在表 4 中.

表 4 不同网络模型错误率比较 (%)

网络模型	Top-1	Top-5	Top-5测试
VGG	23.7	6.8	6.8
Clarifai	24.7	14.2	11.7
MSRA	27.9	9.1	9.1
OverFeat	34.0	13.2	13.6
Krizhevsky	38.1	16.4	16.4

到 ILSVRC 提交时, 我们只训练了单级网络, 以及多尺度模型 D (仅通过微调全连接层而不是所有层). 由此产生的 7 个网络集合有 7.3% 的 ILSVRC 测试错误. 提交后, 我们考虑了仅有两个性能最佳的多尺度模型 (配置 D 和 E) 的集合, 使用密集评估将测试误差降低到 6.8% 使用综合密集和多作物评估. 作为参考, 我们表现最佳的单模具有 7.1% 的误差.

实验发现, 从深层网络的角度出发, 不同的隐层对学习速度的差异很大. 当靠近输出层时, 其相应权值矩阵学习的情况很好, 而靠近输入层时, 其权值矩阵学习很慢, 有时训练了很久, 前几层的权值矩阵仍然和随机初始化的值差不多. 因此, 深度学习中梯度消失问题的根源在于反向传播算法. 为了摆脱反向传播思想的限制, 有研究人员提出了 CapsuleNet, 充分地利用数据中组件的朝向和空间上的相对关系, 并使用动态路由算法计算胶囊的输出. 但是该网络并没有完全地摆脱反向传播算法, 因为网络中的转换矩阵仍然用成本函数通过反向传播进行训练. 近年来, 关于梯度消失问题, 研究人员提出了一系列改良方案, 如精调结合的训练策略和预训练、梯度剪切、权重正则, 使用不同的激活函数 (如 ReLU), 使用批量归一化技巧, 使用残差结构, 使用 LSTM 网络等[25]. 为了从本质上解决梯度消失问题, 设计避免局部极值和鞍点的高效优化算法成为目前深度学习研究的重点.

5 结论

在这项工作中, 我们评估了非常深的卷积网络 (最多 19 个权重层), 用于大规模图像分类. 已经证明, 表示深度有利于分类准确性, 并且使用传统的卷积网络架构可以实现 ImageNet 挑战数据集上的最先进性能. 深度大幅增加. 我们还展示了我们的模型很好地概括了广泛的任务和数据集, 匹配或优于围绕不太深的图像表示构建的更复杂的识别管道. 我们的结果再次证实了深度视觉表征的重要性.

参考文献

1 Deng L, Li JY, Huang JT, *et al*. Recent advances in deep learning for speech research at microsoft. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada. 2013. 8604–8608.

2 Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: Delving deep into convolutional nets. Proceedings of British Machine Vision Conference. Nottingham, UK. 2014.

3 Cimpoi M, Maji S, Vedaldi A. Deep convolutional filter banks for texture recognition and segmentation. arXiv: 1411.6836, 2014.

4 Ciresan DC, Meier U, Masci J, *et al*. Flexible, high performance convolutional neural networks for image classification. Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain. 2011. 1237–1242.

5 朱庆生, 周冬冬, 黄伟. BP 神经网络样本数据预处理应用研究. 世界科技研究与发展, 2012, 34(4): 624–626. [doi: 10.3969/j.issn.1006-6055.2012.04.024]

6 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–255.

7 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, CA, USA. 2012. 1097–1105.

8 Everingham M, Eslami SMA, van Gool L, *et al*. The Pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98–136. [doi: 10.1007/s11263-014-0733-5]

9 郭洋. 利用向量量化的优化混合法分形图像压缩 [硕士学位论文]. 西安: 西安交通大学, 2004.

10 胡保清, 关柯. 一种改进的神经网络数据预处理方法及其在建筑管理中的应用. 土木工程学报, 2004, 37(5): 106–110. [doi: 10.3321/j.issn:1000-131X.2004.05.019]

11 Gkioxari G, Girshick R, Malik J. Actions and attributes from wholes and parts. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 2470–2478.

12 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy. 2010. 249–256.

13 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. arXiv:1311.2901, 2013.

14 Sermanet P, Eigen D, Zhang X, *et al*. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229. 2014.

15 Goodfellow IJ, Bulatov Y, Ibarz J, *et al*. Multi-digit number recognition from street view imagery using deep convolutional neural networks. Proceedings of the 2nd International Conference on Learning Representations. Banff, AB, Canada. 2014.

16 Griffin G, Holub A, Perona P. Caltech-256 object category dataset. Technical Report 7694, Pasadena: California Institute of Technology, 2007.

17 Mollahan A. Estimation of reservoir water saturation using sup-port vector regression in an Iranian carbonate reservoir. American Rock Mechanics Association. 2013.

18 Hoai M. Regularized max pooling for image categorization. Proceedings of British Machine Vision Conference. Nottingham, UK. 2014.

19 Howard AG. Some improvements on deep convolutional neural network based image classification. Proceedings of 2nd International Conference on Learning Representations. Banff, AB, Canada. 2014.

20 Surhone LM, Tennoe MT, Henssonow SF, *et al*. Histograms of oriented gradients. Betascript Publishing, 2010, 12(4): 1368–1371.

21 李卫. 深度学习在图像识别中的研究及应用 [硕士学位论文]. 武汉: 武汉理工大学, 2014.

22 Lee H, Battle A, Raina R, *et al*. Efficient sparse coding algorithms. Proceedings of the 19th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada. 2007. 801–808.

23 Sánchez A V D. Advanced support vector machines and kernel methods. Neurocomputing, 2003, 55(1–2): 5–20. [doi: 10.1016/S0925-2312(03)00373-4]

24 Campbell C. Kernel methods: A survey of current techniques. Neurocomputing, 2002, 48(1–4): 63–84. [doi: 10.1016/S0925-2312(01)00643-9]

25 赵进. 稀疏深度学习理论与应用 [博士学位论文]. 西安: 西安电子科技大学, 2019.