DenseNet详解

原创： YCWU

广而告之

SIGAI飞跃计划第二期等你来挑战！（点击有惊喜）

SIGAI-AI学习交流群的目标是为学习者提供一个AI技术交流与分享的平台。操作指引：关注本微信公众号，回复“芝麻开门”，即可收到入群二维码，扫码即可。

同时在本微信公众号中，回复“SIGAI”+日期，如“SIGAI0515”，即可获取本期文章的全文下载地址（仅供个人学习使用，未经允许，不得用于商业目的）。

一、概述

作为CVPR2017年的Best Paper, DenseNet脱离了加深网络层数(ResNet)和加宽网络结构(Inception)来提升网络性能的定式思维,从特征的角度考虑,通过特征重用和旁路(Bypass)设置,既大幅度减少了网络的参数量,又在一定程度上缓解了gradient vanishing问题的产生.结合信息流和特征复用的假设,DenseNet当之无愧成为2017年计算机视觉顶会的年度最佳论文.

卷积神经网络在沉睡了近20年后,如今成为了深度学习方向最主要的网络结构之一.从一开始的只有五层结构的LeNet, 到后来拥有19层结构的VGG, 再到首次跨越100层网络的Highway Networks与ResNet, 网络层数的加深成为CNN发展的主要方向之一.

随着CNN网络层数的不断增加,gradient vanishing和model degradation问题出现在了人们面前,BatchNormalization的广泛使用在一定程度上缓解了gradient vanishing的问题,而ResNet和Highway Networks通过构造恒等映射设置旁路,进一步减少了gradient vanishing和model degradation的产生.Fractal Nets通过将不同深度的网络并行化,在获得了深度的同时保证了梯度的传播,随机深度网络通过对网络中一些层进行失活,既证明了ResNet深度的冗余性,又缓解了上述问题的产生. 虽然这些不同的网络框架通过不同的实现加深的网络层数,但是他们都包含了相同的核心思想,既将feature map进行跨网络层的连接.

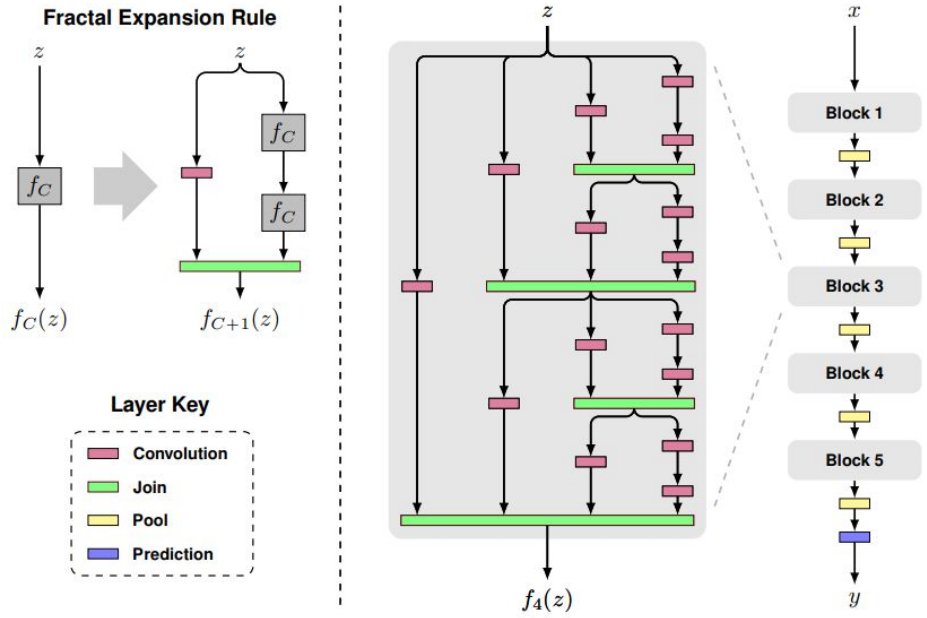

Figure 1: **Fractal architecture.** *Left:* A simple expansion rule generates a fractal architecture with C intertwined columns. The base case, $f_1(z)$, has a single layer of the chosen type (*e.g.* convolutional) between input and output. Join layers compute element-wise mean. *Right:* Deep convolutional networks periodically reduce spatial resolution via pooling. A fractal version uses f_C as a building block between pooling layers. Stacking B such blocks yields a network whose total depth, measured in terms of convolution layers, is $B \cdot 2^{C-1}$. This example has depth 40 ($B = 5, C = 4$).

DenseNet作为另一种拥有较深层数的卷积神经网络,具有如下优点:

(1) 相比ResNet拥有更少的参数数量.

(2) 旁路加强了特征的重用.

(3) 网络更易于训练,并具有一定的正则效果.

(4) 缓解了gradient vanishing和model degradation的问题.

何恺明先生在提出ResNet时做出了这样的假设:若某一较深的网络多出另一较浅网络的若干层有能力学习到恒等映射,那么这一较深网络训练得到的模型性能一定不会弱于该浅层网络.通俗的说就是如果对某一网络中增添一些可以学到恒等映射的层组成新的网路,那么最差的结果也是新网络中的这些层在训练后成为恒等映射而不会影响原网络的性能.同样DenseNet在提出时也做过假设:与其多次学习冗余的特征,特征复用是一种更好的特征提取方式.

二、DenseNet

假设输入为一个图片X_0, 经过一个L层的神经网络, 其中第i层的非线性变换记为$H_i(*)$, $H_i(*)$可以是多种函数操作的累加如BN、ReLU、Pooling或Conv等. 第i层的特征输出记作X_i.

传统卷积前馈神经网络将第i层的输出X_i作为i+1层的输入,可以写作$X_i = H_i(X_{i-1})$. ResNet增加了旁路连接,可以写作

$$X_l=H_l(X_{l-1})+X_{l-1}$$

ResNet的一个最主要的优势便是梯度可以流经恒等函数来到达靠前的层.但恒等映射和非线性变换输出的叠加方式是相加, 这在一定程度上破坏了网络中的信息流.

为了进一步优化信息流的传播,DenseNet提出了图示的网络结构

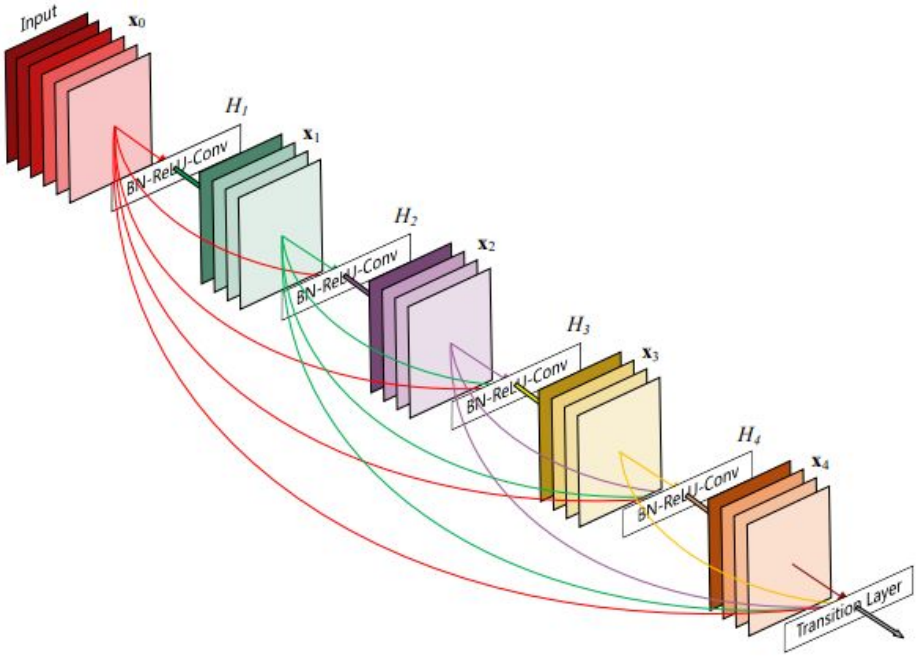

Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

如图所示,第i层的输入不仅与i-1层的输出相关,还有所有之前层的输出有关.记作:

$$X_l=H_l([X_0,X_1,\cdots,X_{l-1}]),$$

其中[]代表concatenation(拼接),既将X_0到X_{l-1}层的所有输出feature map按Channel组合在一起.这里所用到的非线性变换H为BN+ReLU+ Conv(3×3)的组合.

由于在DenseNet中需要对不同层的feature map进行cat操作,所以需要不同层的feature map保持相同的feature size,这就限制了网络中Down sampling的实现.为了使用Down sampling,作者将DenseNet分为多个Denseblock,如下图所示:

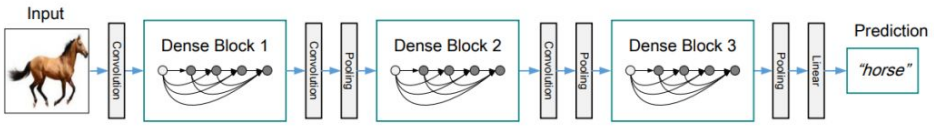

Figure 2: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling.

在同一个Denseblock中要求feature size保持相同大小,在不同Denseblock之间设置transition layers实现Down sampling, 在作者的实验中transition layer由BN + Conv(1×1) +2×2 average-pooling组成.

在Denseblock中,假设每一个非线性变换H的输出为K个feature map, 那么第i层网络的输入便为K0+(i-1)×K,这里我们可以看到DenseNet和现有网络的一个主要的不同点:DenseNet可以接受较少的特征图数量作为网络层的输出,如下图所示

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Table 1: DenseNet architectures for ImageNet. The growth rate for all the networks is $k = 32$. Note that each "conv" layer shown in the table corresponds the sequence BN-ReLU-Conv.

原因就是在同一个Denseblock中的每一层都与之前所有层相关联,如果我们把feature看作是一个Denseblock的全局状态,那么每一层的训练目标便是通过现有的全局状态,判断需要添加给全局状态的更新值.因而每个网络层输出的特征图数量K又称为Growth rate,同样决定着每一层需要给全局状态更新的信息的多少.我们之后会看到,在作者的实验中只需要较小的K便足以实现state-of-art的性能.

虽然DenseNet接受较少的k,也就是feature map的数量作为输出,但由于不同层feature map之间由cat操作组合在一起,最终仍然会是feature map的channel较大而成为网络的负担.作者在这里使用1×1 Conv(Bottleneck)作为特征降维的方法来降低channel数量,以提高计算效率.经过改善后的非线性变换变为BN-ReLU-Conv(1×1)-BN-ReLU-Conv(3×3),使用Bottleneck layers的DenseNet被作者称为DenseNet-B.在实验中,作者使用1×1卷积生成channel数量为4k的feature map.

为了进一步优化模型的简洁性,我们同样可以在transition layer中降低feature map的数量.若一个Denseblock中包含m个feature maps,那么我们使其输出连接的transition layer层生成⌊θm⌋个输出feature map.其中θ为Compression factor, 当θ=1时,transition layer将保留原feature维度不变.

作者将使用compression且θ=0.5的DenseNet命名为DenseNet-C, 将使用Bottleneck和compression且θ=0.5的DenseNet命名为DenseNet-BC

三、 算法分析

由于DenseNet对输入进行cat操作,一个直观的影响就是每一层学到的feature map都能被之后所有层直接使用,这使得特征可以在整个网络中重用,也使得模型更加简洁.

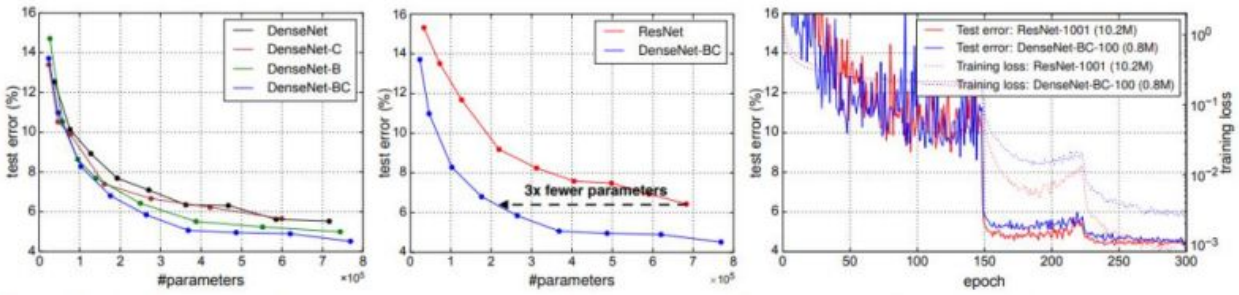

Figure 4: *Left:* Comparison of the parameter efficiency on C10+ between DenseNet variations. *Middle:* Comparison of the parameter efficiency between DenseNet-BC and (pre-activation) ResNets. DenseNet-BC requires about 1/3 of the parameters as ResNet to achieve comparable accuracy. *Right:* Training and testing curves of the 1001-layer pre-activation ResNet [12] with more than 10M parameters and a 100-layer DenseNet with only 0.8M parameters.

从上图中我们可以看出DenseNet的参数效率:左图包含了对多种DenseNet结构参数和最终性能的统计,我们可以看出当模型实现相同的test error时,原始的DenseNet往往要比DenseNet-BC拥有2-3倍的参数量.中间图为DenseNet-BC与ResNet的对比,在相同的模型精度下,DenseNet-BC只需要ResNet约三分之一的参数数量.右图为1001层超过10M参数量的ResNet与100层只有0.8M参数量的DenseNet-BC在训练时的对比,虽然他们在约相同的训练epoch时收敛,但DenseNet-BC却只需要ResNet不足十分之一的参数量.

解释DenseNet为何拥有如此高性能的另一个原因是网络中的每一层不仅接受了原始网络中来自loss的监督,同时由于存在多个bypass与shortcut,网络的监督是多样的.Deep supervision的优势同样在deeply-supervised nets (DSN)中也被证实.(DSN中每一个Hidden layer都有一个分类器,强迫其学习一些有区分度的特征).与DSN不同的是,DenseNet拥有单一的loss function, 模型构造和梯度计算更加简易.

在设计初,DenseNet便被设计成让一层网络可以使用所有之前层网络feature map的网络结构,为了探索feature的复用情况,作者进行了相关实验.作者训练的L=40,K=12的DenseNet,对于任意Denseblock中的所有卷积层,计算之前某层feature map在该层权重的绝对值平均数.这一平均数表明了这一层对于之前某一层feature的利用率,下图为由该平均数绘制出的热力图:

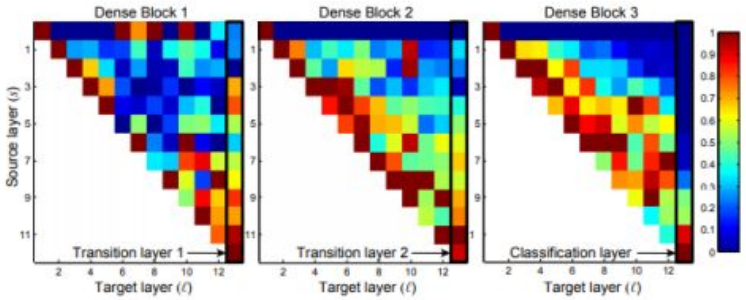

Figure 5: The average absolute filter weights of convolutional layers in a trained DenseNet. The color of pixel (s, ℓ) encodes the average $L1$ norm (normalized by number of input feature-maps) of the weights connecting convolutional layer s to ℓ within a dense block. Three columns highlighted by black rectangles correspond to two transition layers and the classification layer. The first row encodes weights connected to the input layer of the dense block.

红色代表strong use

蓝色代表almost no use

横坐标为选定层

纵坐标为选定层的之前层

最右侧列以及第一行为transition layer

从图中我们可以得出以下结论:

a) 一些较早层提取出的特征仍可能被较深层直接使用

b) 即使是Transition layer也会使用到之前Denseblock中所有层的特征

c) 第2-3个Denseblock中的层对之前Transition layer利用率很低,说明transition layer输出大量冗余特征.这也为DenseNet-BC提供了证据支持,既Compression的必要性.

d) 最后的分类层虽然使用了之前Denseblock中的多层信息,但更偏向于使用最后几个feature map的特征,说明在网络的最后几层,某些high-level的特征可能被产生.

四、实验结果

作者在多个benchmark数据集上训练了多种DenseNet模型,并与state-of-art的模型(主要是ResNet和其变种)进行对比:

Method	Depth	Params	C10	C10+	C100	C100+	SVHN
Network in Network [22]	-	-	10.41	8.81	35.68	-	2.35
All-CNN [32]	-	-	9.08	7.25	-	33.71	-
Deeply Supervised Net [20]	-	-	9.69	7.97	-	34.57	1.92
Highway Network [34]	-	-	-	7.72	-	32.39	-
FractalNet [17]	21	38.6M	10.18	5.22	35.34	23.30	2.01
with Dropout/Drop-path	21	38.6M	7.33	4.60	28.20	23.73	1.87
ResNet [11]	110	1.7M	-	6.61	-	-	-
ResNet (reported by [13])	110	1.7M	13.63	6.41	44.74	27.22	2.01
ResNet with Stochastic Depth [13]	110	1.7M	11.66	5.23	37.80	24.58	1.75
	1202	10.2M	-	4.91	-	-	-
Wide ResNet [42]	16	11.0M	-	4.81	-	22.07	-
	28	36.5M	-	4.17	-	20.50	-
with Dropout	16	2.7M	-	-	-	-	1.64
ResNet (pre-activation) [12]	164	1.7M	11.26*	5.46	35.58*	24.33	-
	1001	10.2M	10.56*	4.62	33.47*	22.71	-
DenseNet ($k = 12$)	40	1.0M	**7.00**	5.24	**27.55**	24.42	1.79
DenseNet ($k = 12$)	100	7.0M	**5.77**	**4.10**	**23.79**	**20.20**	1.67
DenseNet ($k = 24$)	100	27.2M	**5.83**	**3.74**	**23.42**	**19.25**	**1.59**
DenseNet-BC ($k = 12$)	100	0.8M	**5.92**	4.51	**24.15**	22.27	1.76
DenseNet-BC ($k = 24$)	250	15.3M	**5.19**	**3.62**	**19.64**	**17.60**	1.74
DenseNet-BC ($k = 40$)	190	25.6M	-	**3.46**	-	**17.18**	-

Table 2: Error rates (%) on CIFAR and SVHN datasets. k denotes network's growth rate. Results that surpass all competing methods are **bold** and the overall best results are **blue**. "+" indicates standard data augmentation (translation and/or mirroring). $*$ indicates results run by ourselves. All the results of DenseNets without data augmentation (C10, C100, SVHN) are obtained using Dropout. DenseNets achieve lower error rates while using fewer parameters than ResNet. Without data augmentation, DenseNet performs better by a large margin.

由上表我们可以看出,DenseNet只需要较小的Growth rate(12,24)便可以实现state-of-art的性能,结合了Bottleneck和Compression的DenseNet-BC具有远小于ResNet及其变种的参数数量,且无论DenseNet或者DenseNet-BC,都在原始数据集和增广数据集上实现了超越ResNet的性能.

理解计算：从√2到AlphaGo ——第4季 凛冬将至

【获取码】SIGAI0802

机器学习和深度学习中值得弄清楚的一些问题

【获取码】SIGAI0824

浓缩就是精华--SIGAI机器学习蓝宝书

数学类

【获取码】SIGAI0417

学好机器学习需要哪些数学知识

【获取码】SIGAI0511

理解梯度下降法

【获取码】SIGAI0518

理解凸优化

【获取码】SIGAI0531

理解牛顿法

机器学习类

【获取码】SIGAI0428

用一张图理解SVM的脉络

【获取码】SIGAI0505

理解神经网络的激活函数

【获取码】SIGAI0522

【实验】理解SVM核函数和参数的作用

【获取码】SIGAI0601

【群话题精华】五月集锦—机器学习和深度学习中一些值得思考的问题

【获取码】SIGAI0602

大话AdaBoost算法

【获取码】SIGAI0606

理解主成分分析（PCA）

【获取码】SIGAI0611

理解决策树

【获取码】SIGAI0613

用一句话总结常用的机器学习算法

【获取码】SIGAI0618

理解过拟合

【获取码】SIGAI0627

k近邻算法

【获取码】SIGAI0704

机器学习算法地图

【获取码】SIGAI0706

反向传播算法推导—全连接神经网络

【获取码】SIGAI0711

如何成为一名优秀的算法工程师

【获取码】SIGAI0723

流形学习概述

【获取码】SIGAI0725

随机森林概述

深度学习类

【获取码】SIGAI0426

卷积神经网络为什么能够称霸计算机视觉领域？

【获取码】SIGAI0508

深度卷积神经网络演化历史及结构改进脉络-40页长文全面解读

【获取码】SIGAI0515

循环神经网络综述—语音识别与自然语言处理的利器

【获取码】SIGAI0625

卷积神经网络的压缩与加速

【获取码】SIGAI0709

生成式对抗网络模型综述

【获取码】SIGAI0718

基于深度负相关学习的人群计数方法

【获取码】SIGAI0723

关于感受野的总结

【获取码】SIGAI0806

反向传播算法推导--卷积神经网络

【获取码】SIGAI0810

理解Spatial Transformer Networks

机器视觉类

【获取码】SIGAI0420

人脸识别算法演化史

【获取码】SIGAI0424

基于深度学习的目标检测算法综述

【获取码】SIGAI0503

人脸检测算法综述

【获取码】SIGAI0525

【SIGAI综述】行人检测算法

【获取码】SIGAI0604

FlowNet到FlowNet2.0：基于卷积神经网络的光流预测算法

【获取码】SIGAI0608

人体骨骼关键点检测综述

【获取码】SIGAI0615

目标检测算法之YOLO

【获取码】SIGAI0622

场景文本检测——CTPN算法介绍

【获取码】SIGAI0629

自然场景文本检测识别技术综述

【获取码】SIGAI0716

人脸检测算法之S3FD

【获取码】SIGAI0727

基于内容的图像检索技术综述——传统经典方法

【获取码】SIGAI0817

基于内容的图像检索技术综述——CNN方法

自然语言处理

【获取码】SIGAI0803

基于深度神经网络的自动问答概述

【获取码】SIGAI0820

文本表示简介

工业应用类

【获取码】SIGAI0529

机器学习在自动驾驶中的应用-以百度阿波罗平台为例【上】