VGGNet：

Very Deep Convolutional Networks for Large-Scale Image Recognition

分享人：王嘉恒

目录

1.写作背景
2.特点
3.网络结构
4.训练过程
5.测试结果

1.写作背景

AlexNet是2012年ImageNet竞赛的冠军，在那之后许多学者对AlexNet的结构进行改进，希望得到更好的准确性。其中有两个方向：分别是缩小卷积核和多尺度。

AlexNet 中当时已经提到神经网络的深度可能会对准确性有影响，作者则是更加直接的证明了这一结论，并用VGGNet参加了2014年ImageNet竞赛，拿到了第二名。

可以说作者是集大成：用更小的卷积核，多尺度训练和测试。再次基础上再增加网络的深度，一篇文章就“水出来了”。

2.特点

1.使用小卷积核和步长 3×3

2.层数更深更宽 16、19层

3.连续使用数个相同的padding为1，形状为3×3的卷积层，然后再接最大池化层

4.网络测试阶段将训练阶段的三个全连接层替换成三个卷积层

3.网络结构

作者一共训练了五种网络：

A：11层 没什么特别的

A-LRN：后证明没什么作用

B：加了两个卷积层

C：又加了3个卷积层，但核大小为1×1

D：和C层数相同，但盒核大小为3×3

E：在D的基础上又加了3个卷积层，即VGG19

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

图1.VGG网络结构

3.网络结构

输入：固定尺寸为224×224带下的RGB图像

预处理：每个像素减去训练集中的RGB均值

Padding：1 for 3×3

激活函数：RELU

池化层：每个池化核2×2，步长为2

输出：经过softmax计算的1000个类的向量，代表每个类的概率

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

图1.VGG网络结构

4.训练过程

带动量的小批量梯度下降，动量设为0.9，batch_size设为256

Learning rate：0.01 每当错误率趋于稳定就缩小十倍

权重衰减：惩罚因子5×10^{-4}

Dropout rate：0.5 （前两个全连接层）

权重初始化策略：先训练网络A，然后利用A的网络参数，给其他的模型进行初始化

5.结果分析

作者用两种方法设置训练图像的尺寸S：

1.固定S: S=256和S=384

2.给定S的一个范围：[256,512]

对于固定的S，设定Q=S，对于浮动的S，$Q = (S_{\max} + S_{\min})/2$

A、B对比；A、E对比；BC对比；CD对比

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

图2.单一尺度上(测试时)VGG的性能结果

5.结果分析

对固定的S，使用接近S的三个测试尺寸Q=S，S-32，S+32来进行评估、对抖动的S，测试尺寸为Q=$S_{\min}$，$\frac{S_{\min}+S_{\max}}{2}$，$S_{\max}$

采用多尺寸的测试图像，可以发现在测试时采用尺寸抖动可以获得更好的结果。

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**

图2. 多尺度上VGG的性能结果

5.结果分析

ConvNet config. (Table 1)	Evaluation method	top-1 val. error (%)	top-5 val. error (%)
D	dense	24.8	7.5
	multi-crop	24.6	7.5
	multi-crop & dense	**24.4**	**7.2**
E	dense	24.8	7.5
	multi-crop	24.6	7.4
	multi-crop & dense	**24.4**	**7.1**

图4. 对密集卷积网络评估和多重裁切评估的比较

Dense：用卷积层代替全连接层的结果

Multi-crop：将测试图片的短边缩放到不同大小的Q，然后在Q*Q的图像上裁剪出多个图像，将这些图像作为神经网络的输入。然后得到每个裁剪图像的分类概率，相加平均后作为分类概率。

结论：视觉表达中深度的重要性。