《Very Deep Convolutional Networks For Large-Scale Image Recognition》

摘要：

1. **本文的贡献：**研究了卷积网络的深度在大规模图像识别中对准确率的影响。
2. **主要结论：**卷积网络深度的增加和小卷积核的使用对网络最终分类识别的效果有很大的提高作用。
3. **主要的改进做法：**使用了3×3尺寸大小和步长$stride = 1$的小卷积核；同时在 AlexNet 的基础上加深了卷积层的数量。
4. **模型的比赛结果：**ImageNet 2014 年目标定位竞赛的第一名，图像分类竞赛第二名。

VGG 模型的网络结构：

根据卷积核的大小和卷积层的层数，VGG 共有 6 种配置，分别为 A、A-LRN、B、C、D、E，其中 D 和 E 两种是最为常用的 VGG16 和 VGG19。

VGG16 VGG19

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					

相关结构介绍：

（1） **conv3-64：**指卷积核大小为3×3，输出通道数为 64 的卷积层，同样地，**conv3-128** 指卷积核大小为3×3，输出通道数为 128 的卷积层；

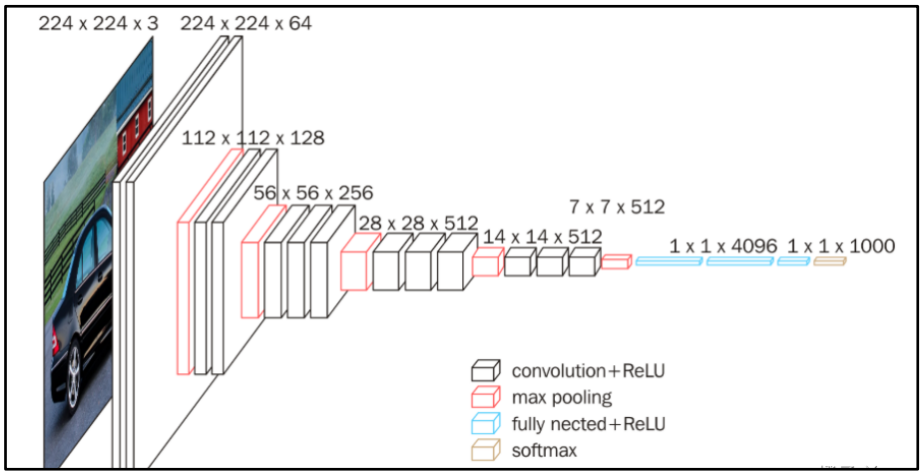
（2） **input（224×224 RGB image）：**指预处理后的输入图片大小为224×244的彩色图像（通道数为 3），即$224 \times 244 \times 3$；

（3） **maxpool：**指最大池化层，采用的是2×2，$stride = 2$的最大池化方法，因此在通过最大池化层后，图像的高度和宽度都相应减半；

（4） **FC-4096：**指全连接层，其中有 4096 个结点。同样地，FC-1000 是指全连接层中有 1000 个结点；

（5） **LRN：**指局部响应标准化（AlexNet 中提出的方法），但本实验证明了该标准化并不会提高网络在相关数据集上的性能，反而会增加内存消耗和计算时间。

VGG-16 的卷积计算图：



卷积-卷积-池化-卷积-卷积-池化-卷积-卷积-卷积-池化-卷积-卷积-卷积-池化-卷积-卷积-卷积-池化-全连接-全连接-全连接

简单说明：以池化层作为分界，VGG-16 共有 6 个块结构，每个块结构中的通道数相同。而由于卷积层和全连接层都有权重系数，其中卷积层 13 层，全连接 3 层，池化层不涉及权重，所以共有$13 + 3 = 16$层。

训练数据集的处理：

1. Multi-scale 训练（多尺度训练）：

用 Multi-Scale 的方法做数据增强——将原始图像缩放到不同尺寸 S，然后再随机裁切为 224×224 的图片。这样不仅能增加数据量，还能对于防止模型过拟合起到不错的效果。

方法 1：在不同的尺度下，训练多个分类器：参数 S 为短边长，训练$S = 256$和$S = 384$两个分类器，其中$S = 384$的分类器用$S = 256$的进行初始化，且将步长调为 10e-3。

方法 2：直接训练一个分类器，每次数据输入的时候，每张图片都会被重新缩放，缩放的短边 S 随机从[256,512]中选择一个。

2. 对于224×224的 RGB 图像，对每一个像素减去其均值。

两种结果预测的方式：

1. **multi-crop：**即对图像进行多样本的随机裁剪，然后通过网络预测每一个样本的结构，最终对所有结果平均；
2. **dense：**利用 FCN 的思想，将原图直接送到网络进行预测，然后将最后的全连接层改为 1x1 的卷积，这样最后可以得出一个预测的 score map，再对结果求平均。(1x1 卷积核的作用：降维，增加数据的非线性性。)

实验结果分析：

（1）单尺度预测：

Table 3: **ConvNet performance at a single test scale.**

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**

由上表可得：1）模型 E（VGG19）的效果最好，即网络越深，效果越好；2）同一种模型，随机裁剪的效果好于固定 S 大小的 256，384 两种尺度，即随机裁

剪的数据增强能更准确的提取图像多尺度的信息。

（2）多尺度预测：

Table 4: **ConvNet performance at multiple test scales.**

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**

由上表可得：1）对比单尺度预测，多尺度综合预测，能够提升预测的精度；2）同单尺度预测一样，多尺度预测也证明了随机裁剪的作用。

（3）多尺度裁剪：

Table 5: **ConvNet evaluation techniques comparison.** In all experiments the training scale S was sampled from $[256; 512]$, and three test scales Q were considered: $\{256, 384, 512\}$.

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**

由上表可得：1）数据生成方式 multi-crop 效果略优于 dense，但精度的提高不足以弥补计算上的损失；2）multi-crop 和 dense 方法结合的效果最优，说明了 multi-crop 和 dense 两种方法能够互为补充。

此外，本文作者还指出：VGG 网络不仅能在 ILSVRC 的分类和检测任务中取得 the state-of-the-art 的精度效果，在其他的数据集上也具有很好的推广能力。

对于本文的感悟：

VGG 神经网络的成功有效证明了神经网络深度的提高能够帮助相关模型获得更好的表现和结果；同时还侧面反映了简单的网络结构，即仅采用3×3的卷积核也能有效搭建卷积神经网络，继而为之后的研究者们提供了更好的设计思路。