Very Deep Convolutional Networks for Large-Scale Image Recognition

汇报人：兰冬雷

2020 年 11 月 29 日

背景

- 卷积网络（ConvNets）最近在大规模图像和视频识别方面取得了巨大的成功。
- ImageNet 大规模视觉识别挑战赛（ILSVRC）在推进深度视觉识别架构方面发挥了重要作用。
- ConvNets 目前大多数人已经尝试的改进方式：

1. 利用了较小的卷积核和较小的第一卷积层步幅（smaller receptive window size and smaller stride of the first convolutional layer.）

2. 在整个图像和多个尺度上密集地训练和测试网络（training and testing the networks densely over the whole image and over multiple scales.）

ABSTRACT

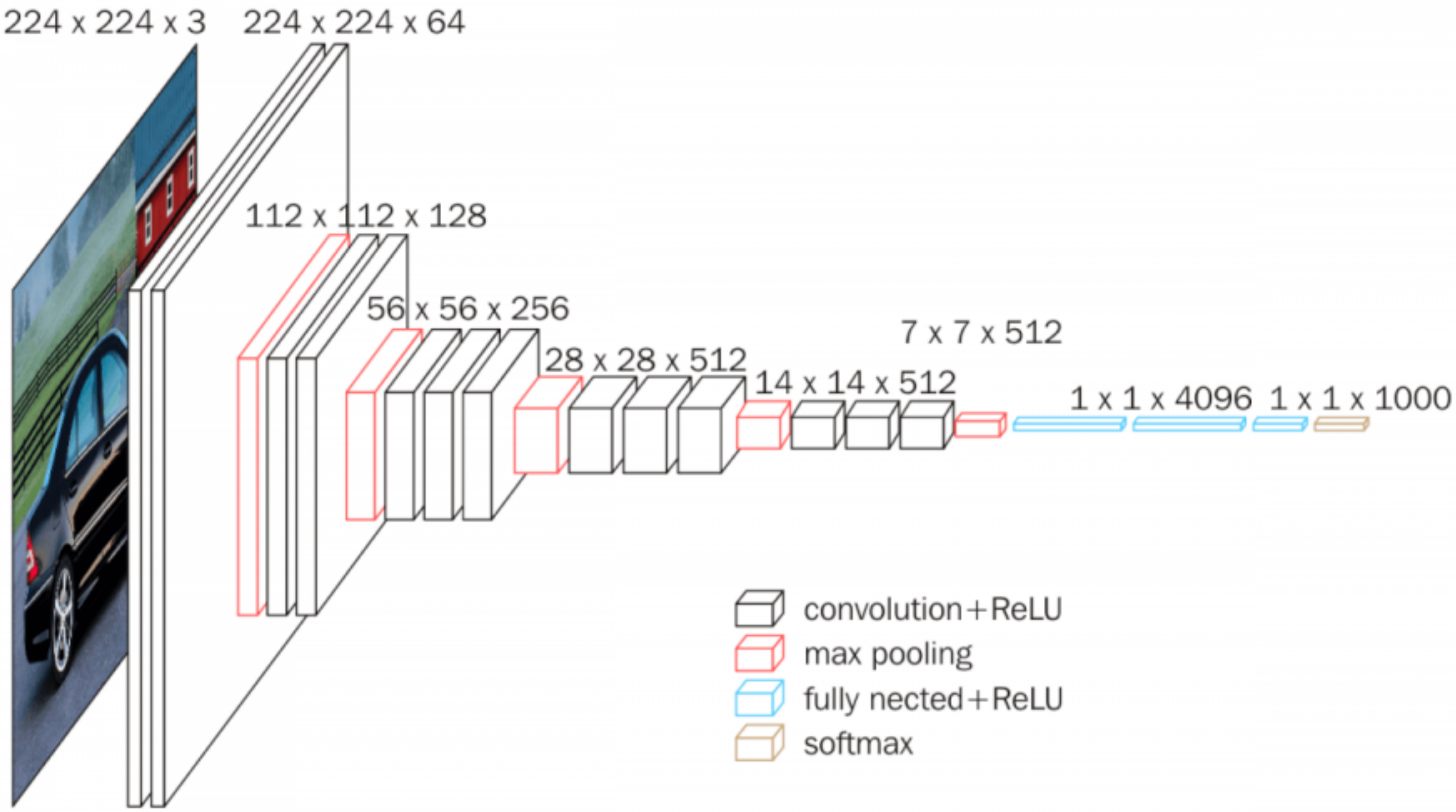
- 论文研究了在大规模图像识别中卷积网络的深度对其精度的影响。
- 使用 3×3 的卷积核。
- 网络深度提升到了 16-19 层。（AlexNet 8 层）
- 获得了 2014 ImageNet Challenge 的 Localistion 第一名和 Classification 第二名。（Classification 的第一名是 GoogLeNet）
- VGG 名称的来源：Visual Geometry Group, Department of Engineering Science, University of Oxford.

VGG-16 网络结构

- 网络输入的图片尺寸是 3ch×224×224
- 卷积核大小为 3×3，（Table 1: 1×1 可以看作是输入通道的线性变换）
- 卷积的步幅（Stride）固定为 1 像素（pixel）
- 有 5 个最大池化层，并非每个卷积层后面都接着一个最大池化层。最大池化的窗口为 2×2，步幅为 2
- 卷积层最后接着 3 个全连接层，前两个有 4096 个通道，最后是一个 1000 的 ILSVRC 分类（Softmax）。
- 所有的隐藏层都使用 ReLU 激活函数
- 摒弃了 LRN， Section 4 实验证实了 LRN 不会提高 ILSVRC 数据集的性能

Feature Map: 224 ➜ 112➜ 56 ➜ 28 ➜14 ➜ 7

Channels: 64 ➜ 128 ➜ 256 ➜ 512



VGG-16 (13 Conv.& 3 FC.)

https://neurohive.io/en/popular-networks/vgg16/

6 个网络的结构

1. A、A-LRN、B、C、D、E 这个 6 网络只在深度上有所不同，其他的都采用通用的设计。

2. 网络参数

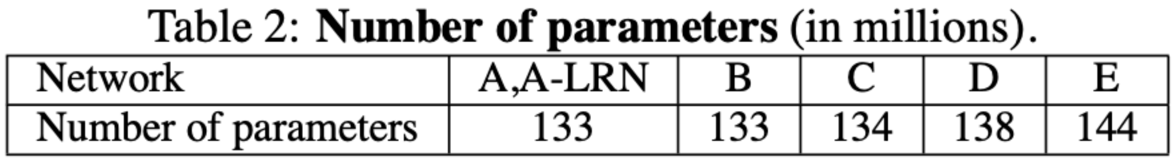
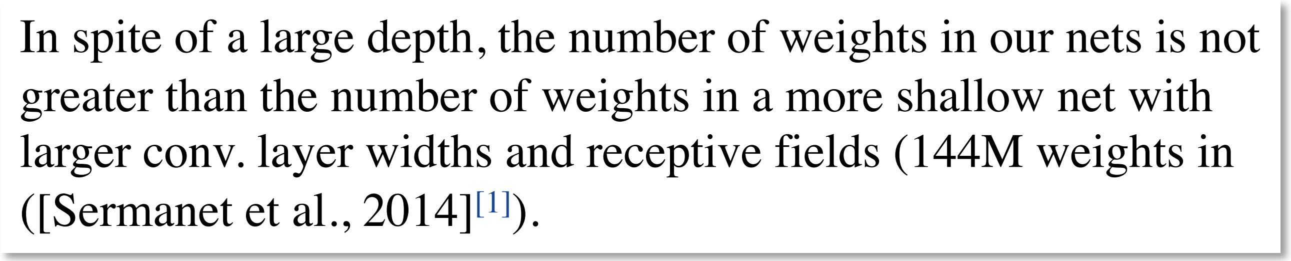


Table 2: **Number of parameters** (in millions).

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



In spite of a large depth, the number of weights in our nets is not greater than the number of weights in a more shallow net with larger conv. layer widths and receptive fields (144M weights in ([Sermanet et al., 2014][1]).

(VGG-16 ≈ 138.36 M、AlexNet ≈ 60.97 M、GoogLeNet ≈ 7M)

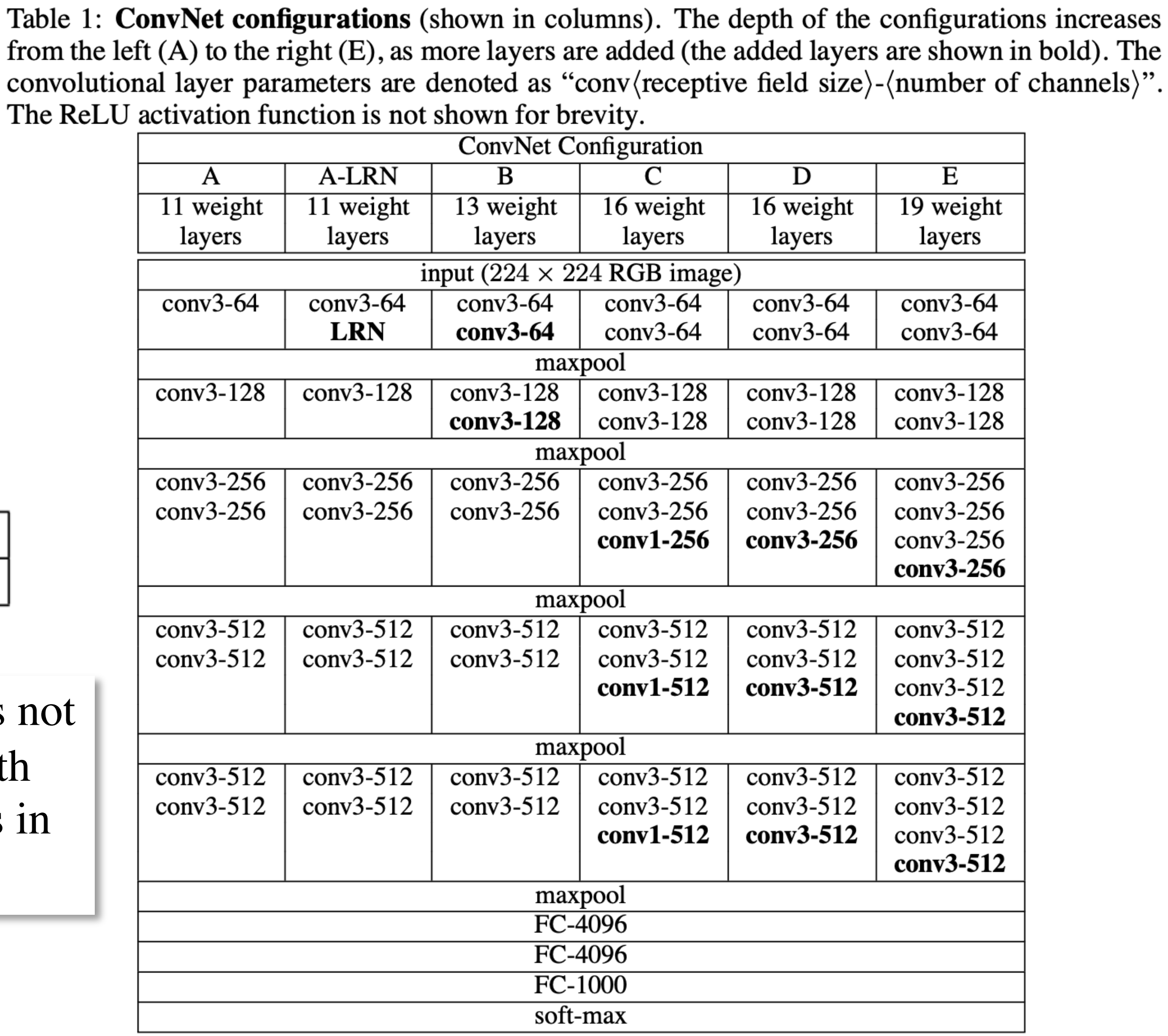


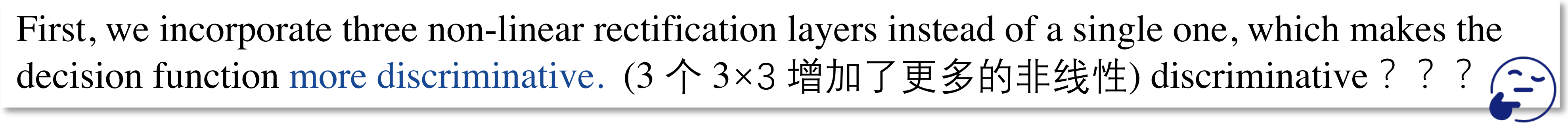
Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The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

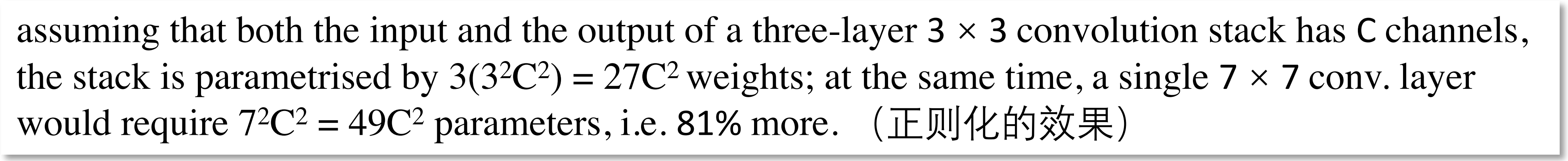
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	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					

[1] OverFeat: 基于 AlexNet，实现了识别、定位、检测共用同一个网络框架；获得了 2013 年 ILSVRC 定位比赛的冠军。

网络的其他讨论

- 感受野（Receptive Fields）：3 个 (3×3) == 1 个 (7×7).
- 使用三个 3×3 conv. 层的堆叠而不是只使用一个 7×7 层？

1. 
First, we incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative. (3 个 3×3 增加了更多的非线性) discriminative？？？

2. 
assuming that both the input and the output of a three-layer 3 × 3 convolution stack has C channels, the stack is parametrised by $3(3^2C^2) = 27C^2$ weights; at the same time, a single 7 × 7 conv. layer would require $7^2C^2 = 49C^2$ parameters, i.e. 81% more. （正则化的效果）

- GoogLeNet 是 ILSVRC-2014 分类任务的第一名，但 VGG 在单网分类精度方面优于 GoogLeNet。

训练 TRAINING/测试 TESTING

- Momentum Mini-batch 梯度下降
- batch size=256
- momentum 参数为 0.9
- 权重衰减（L2 正则化 5×10^{-4}）和 Dropout 正则化（丢弃率为 0.5）
- 学习率初始为 0.01。学习率衰减：精度停止提升时，学习率减少 10 倍，总共减少 3 次。
- 74 个 epochs

- 训练 Training image size ：S （we also refer to S as the training scale）

0. S = 224，那么裁剪到的是整张图像；如果 S ≫ 224，裁剪到的将是图像的一小部分.

1. S = 256 广泛使用，如（AlexNet，GoogLeNet）

2.训练 S=384 时，用 S = 256 预先训练的权重进行初始化.

3.从 $[\mathrm{Smin}, \mathrm{Smax}]$ 随机抽取（$\mathrm{Smin} = 256, \mathrm{Smax} = 512$）

- 测试：Q，不一定等于 S。对每个 S 使用不同的 Q 可以提升性能。

实验结果 ILSVRC-2012 数据集

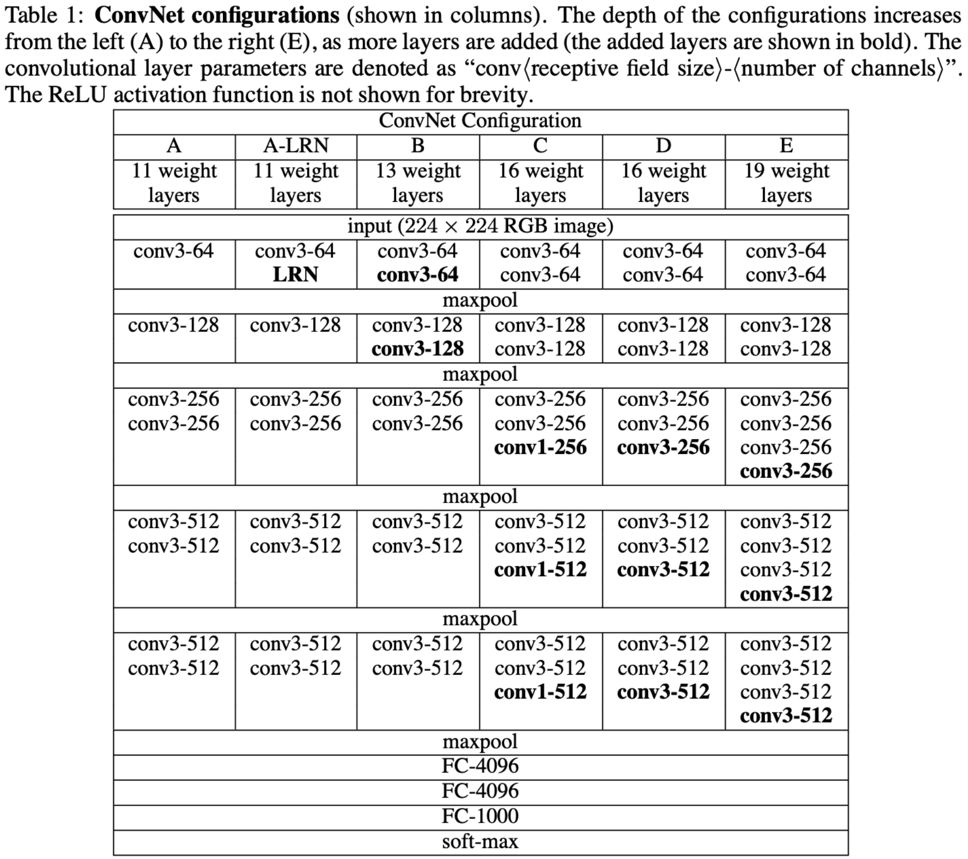


Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The 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					

CovNet Config.

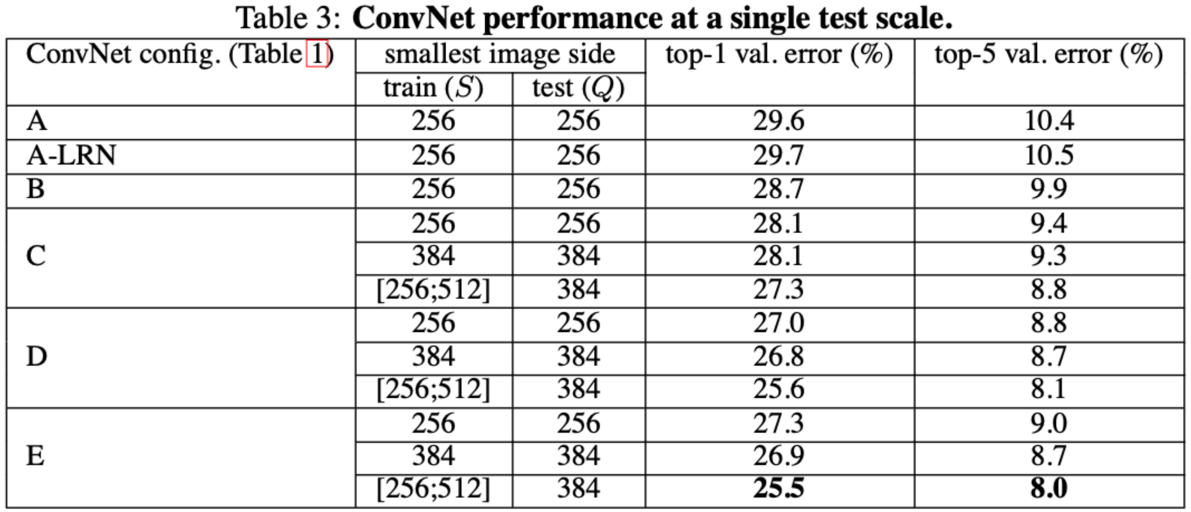


Table 3: **ConvNet performance at a single test scale.**

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**

固定：S，$Q = S$ & 随机：$S \in [Smin，Smax]$，$Q = 0.5(Smin + Smax)$

1.A/A-LRN 2.A~E（深度） 3.B/C（增加了非线性） 4.C/D（3×3 能捕捉更多的上下文信息）5.随机 S

"两个 3×3 与一个 5×5 具有相同的感受野。将 B 网络中的每对 3×3 替换为单个 5×5，替换后的网络的 top-1 误差比 B 的 top-1 误差高 7%."

实验结果 ILSVRC-2012 数据集

相比于固定 Q 的 scale，规模抖动（scale jittering ）会带来更好的性能。

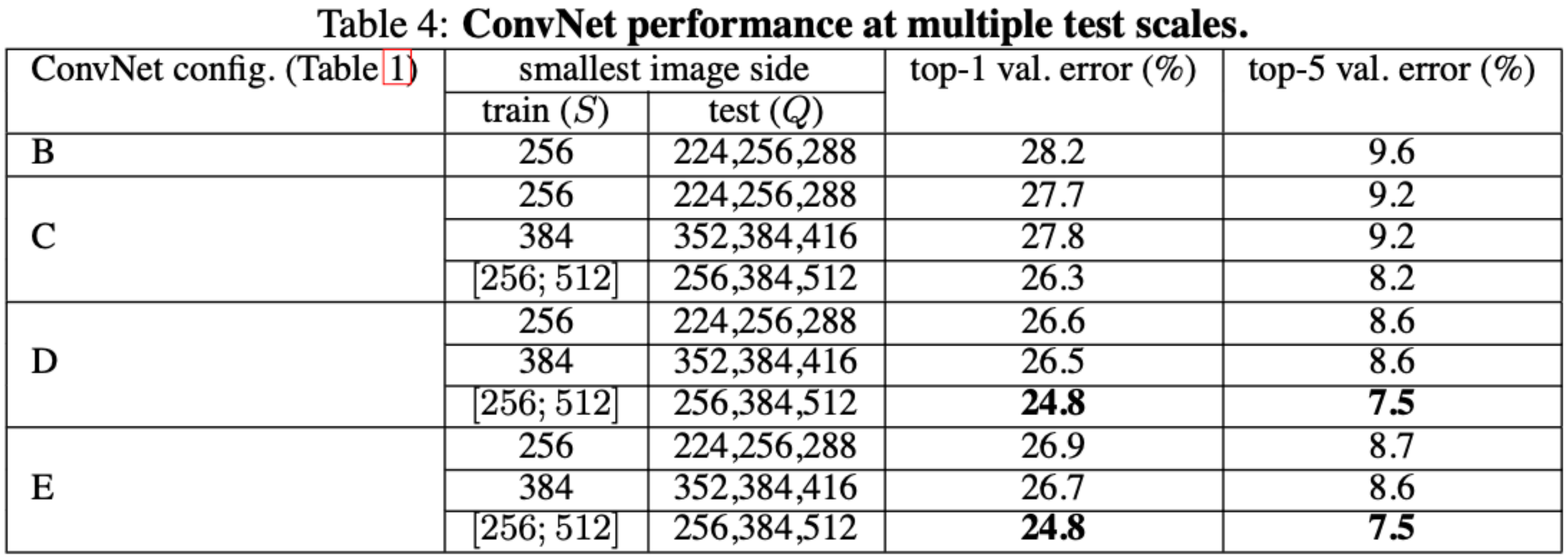


Table 4: **ConvNet performance at multiple test scales.**

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**

固定 Q=384, top-5 error 为 8.0%

dense/multi-crop

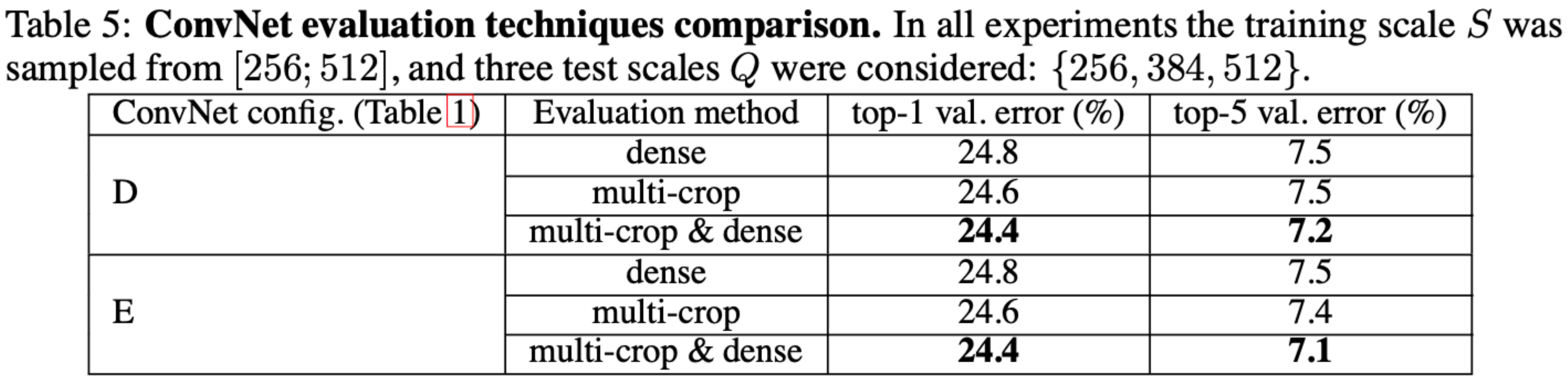


Table 5: **ConvNet evaluation techniques comparison.** In all experiments the training scale S was sampled from $[256; 512]$, and three test scales Q were considered: $\{256, 384, 512\}$.

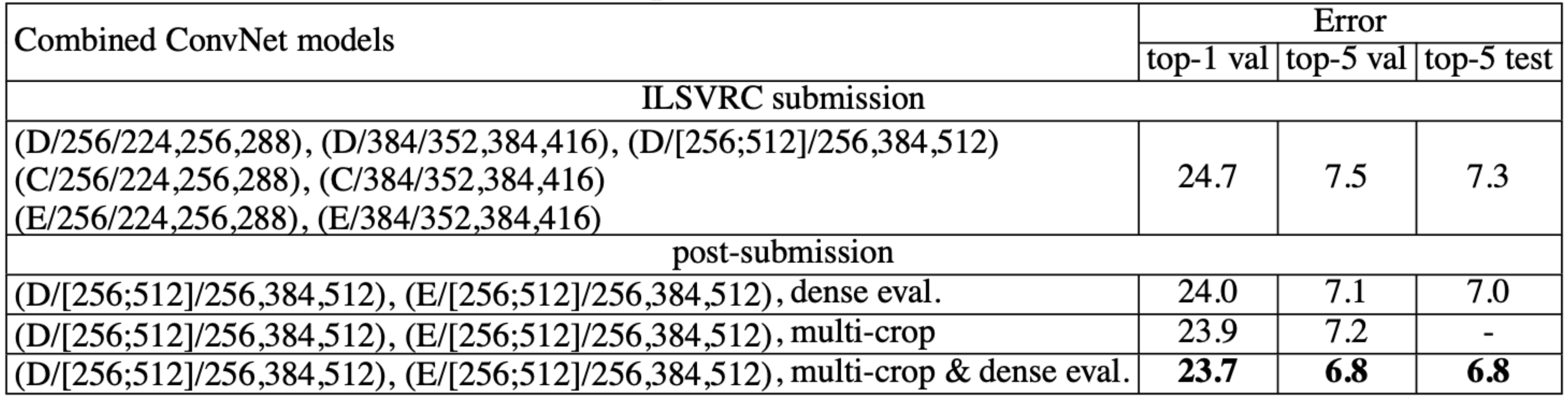
ConvNet config. (Table 1)	Evaluation method	top-1 val. error (%)	top-5 val. error (%)
D	dense	24.8	7.5
	multi-crop	24.6	7.5
	multi-crop & dense	**24.4**	**7.2**
E	dense	24.8	7.5
	multi-crop	24.6	7.4
	multi-crop & dense	**24.4**	**7.1**

困惑：Dense ??? 🤔

Table 5 中，表现最好的单一模型实现了 7.1% 的 top-5 误差（模型E）.Table 6 结合 D 和 E 这个两个网络，测试的 top-5 误差下降了 0.1% .

多个 ConvNet 结合

Table 6: **Multiple ConvNet fusion results.**



Combined ConvNet models	Error		
	top-1 val	top-5 val	top-5 test
ILSVRC submission			
(D/256/224,256,288), (D/384/352,384,416), (D/[256;512]/256,384,512) (C/256/224,256,288), (C/384/352,384,416) (E/256/224,256,288), (E/384/352,384,416)	24.7	7.5	7.3
post-submission			
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), dense eval.	24.0	7.1	7.0
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop	23.9	7.2	-
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop & dense eval.	**23.7**	**6.8**	**6.8**

总结 top-5 error. :

1. ConvNet E, $\mathrm{S} \in [\mathrm{Smin}，\ \mathrm{Smax}]$， $\mathrm{Q} = 0.5(\mathrm{Smin} + \mathrm{Smax})$ ===> 8.0 %
2. ConvNet E, $\mathrm{Q} = \{256, 384, 512\}$ ===> 7.5 %
3. ConvNet E, dense ===> 7.5 %
4. ConvNet E, multi-crop ===> 7.4 %
5. ConvNet E, dense&multi-crop ===> 7.1 %
6. ConvNet D&E, 结合 1~5 ===> 6.8 %

与其他方法进行比较

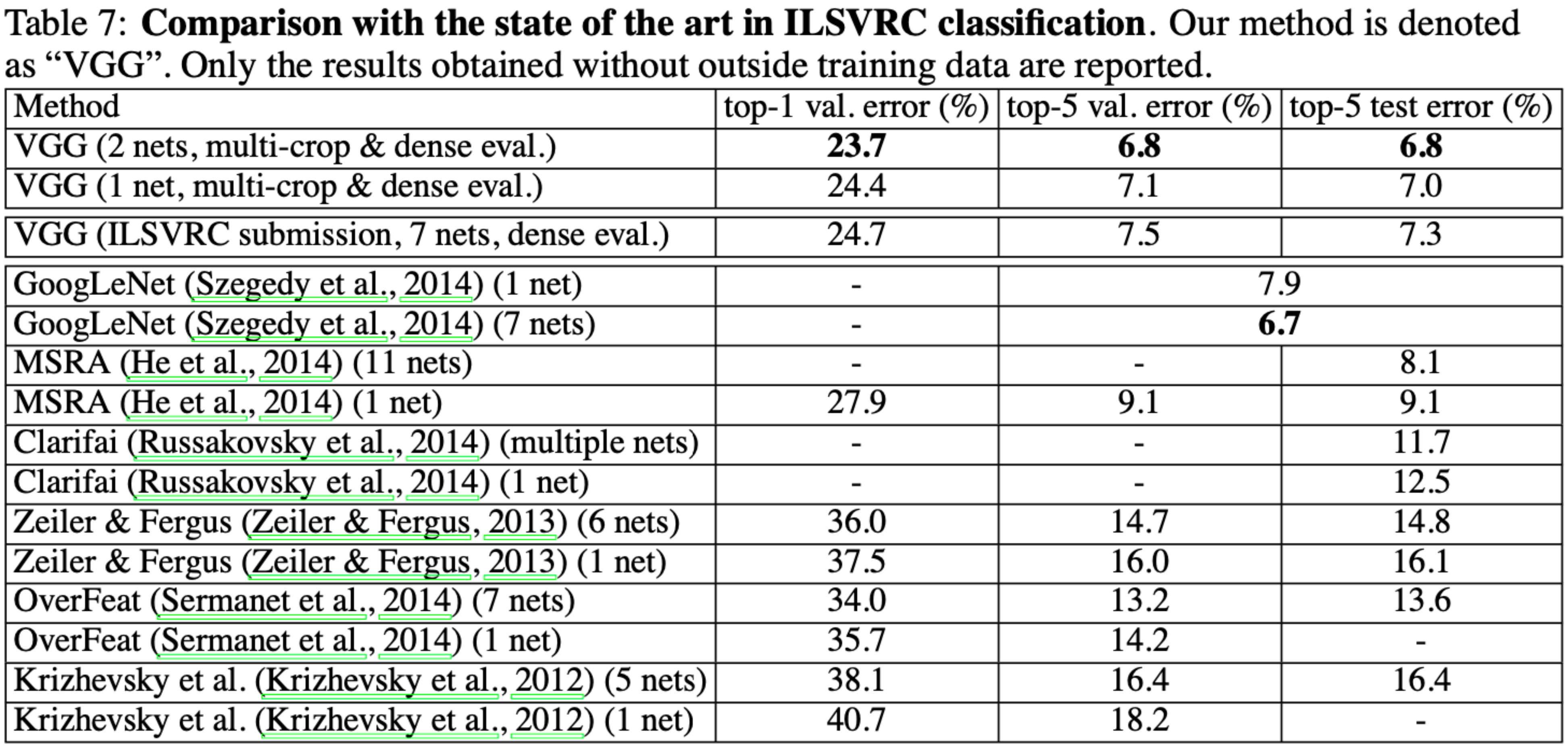


Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

在单网性能方面，VGG 做到了最好（7.0% 的测试误差），GoogLeNet 是 7.9%。

补充： LOCALISATION（定位）

- VGG 团队以 25.3% 的测试误差赢得了 ILSVRC-2014 挑战赛的 Localisation 任务的冠军。比 ILSVRC-2013 的冠军 OverFeat 的结果要好得多，而且 VGG 还有没采用分辨率增强等技术，VGG 还有提升空间。

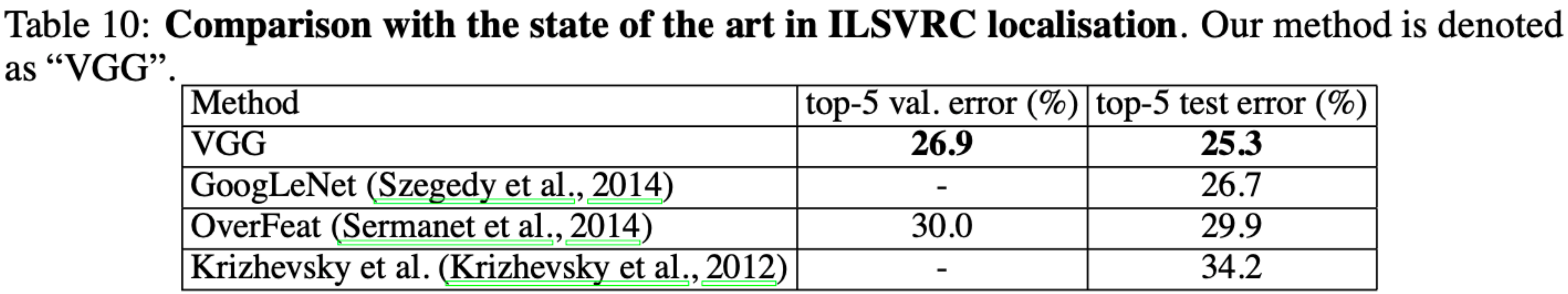


Table 10: **Comparison with the state of the art in ILSVRC localisation**. Our method is denoted as “VGG”.

Method	top-5 val. error (%)	top-5 test error (%)
VGG	**26.9**	**25.3**
GoogLeNet (Szegedy et al., 2014)	-	26.7
OverFeat (Sermanet et al., 2014)	30.0	29.9
Krizhevsky et al. (Krizhevsky et al., 2012)	-	34.2

VGG 在其他数据集上的表现

- 去掉最后一层（1000 类的 Softmax）；
- 使用倒数第二层的 4096 维的激活值作为图像的特征；
- 经过 L2 归一化，并与线性 SVM 分类器相结合，在目标数据集上进行训练。
- VOC-2007 图像数据集包含 10K 张图像，VOC-2012 包含 22.5K 张。每张图像都被标注了一个或几个标签，对应 20 个对象类别。识别性能采用各类平均精度（mAP）来衡量。 Caltech-101 数据集包含 9K 图像，有 102 个类（101 个对象类别和一个背景类）。Caltech-256 有 31K 图像，257 个类别。

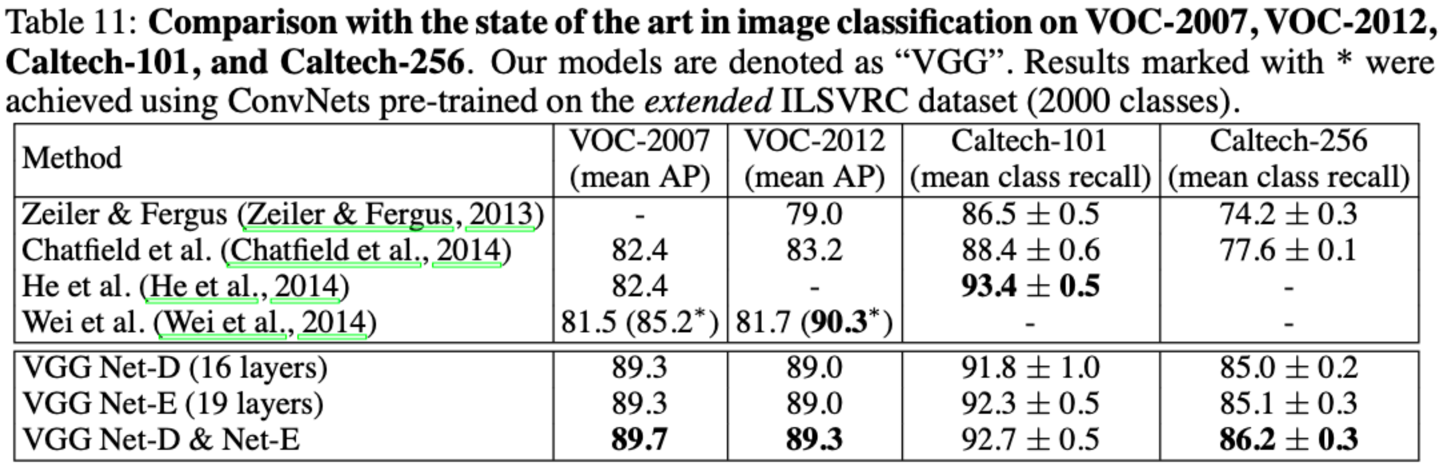


Table 11: **Comparison with the state of the art in image classification on VOC-2007, VOC-2012, Caltech-101, and Caltech-256**. Our models are denoted as "VGG". Results marked with * were achieved using ConvNets pre-trained on the *extended* ILSVRC dataset (2000 classes).

Method	VOC-2007 (mean AP)	VOC-2012 (mean AP)	Caltech-101 (mean class recall)	Caltech-256 (mean class recall)
Zeiler & Fergus (Zeiler & Fergus, 2013)	-	79.0	86.5 ± 0.5	74.2 ± 0.3
Chatfield et al. (Chatfield et al., 2014)	82.4	83.2	88.4 ± 0.6	77.6 ± 0.1
He et al. (He et al., 2014)	82.4	-	**93.4 ± 0.5**	-
Wei et al. (Wei et al., 2014)	81.5 (85.2*)	81.7 (**90.3***)	-	-
VGG Net-D (16 layers)	89.3	89.0	91.8 ± 1.0	85.0 ± 0.2
VGG Net-E (19 layers)	89.3	89.0	92.3 ± 0.5	85.1 ± 0.3
VGG Net-D & Net-E	**89.7**	**89.3**	92.7 ± 0.5	**86.2 ± 0.3**

1. VGG 在 Caltech-101 数据集上相比何恺明等人的方法稍差一些，但是 VGG 在 VOC-2007 上明显优于何恺明等人的方法。
2. (Wei et al., 2014) 进行了预训练和对象检测辅助分类流水线（bject detection-assisted classification pipeline.），因此比 VGG 高 1%.

PASCAL VOC-2012 动作分类任务

- VGG 仅仅依靠非常深的卷积特征的表示能力，就取得了第一名的成绩。

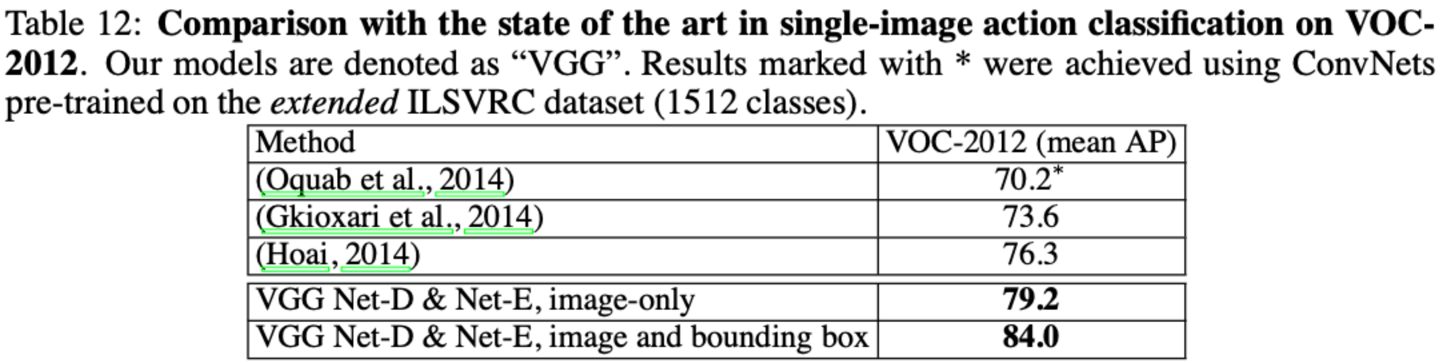


Table 12: **Comparison with the state of the art in single-image action classification on VOC-2012**. Our models are denoted as “VGG”. Results marked with * were achieved using ConvNets pre-trained on the *extended* ILSVRC dataset (1512 classes).

Method	VOC-2012 (mean AP)
(Oquab et al., 2014)	70.2*
(Gkioxari et al., 2014)	73.6
(Hoai, 2014)	76.3
VGG Net-D & Net-E, image-only	**79.2**
VGG Net-D & Net-E, image and bounding box	**84.0**

我的疑问

1. More discriminative ?

2. dense? 🤔

谢谢

汇报人：兰冬雷

2020 年 11 月 29 日