# 用于大规模图像识别的深层卷积网络

## 摘要

在本篇论文中，作者研究了不同深度的卷积网络对于大规模图像识别准确性的影响。作者提出了一种新的卷积神经网络——3×3的卷积核、16或19个权重层（后被称为VGG-16或VGG-19）。除了ImageNet，该网络也可以被泛化迁移于其他数据集，并取得不错的成绩。

|  |
| --- |
| 相比以往的卷积网络，VGG用小卷积核取代了以往的大卷积核，网络深度加深，宽度逐层加宽，没有输入的限制，但是VGG的参数量大。 |

## 导论

对于图像分类问题，从早期的高维浅层神经网络到现在主流的卷积神经网络，随着数据集（如大型公共图像库ImageNet）、算力的提升，分类效率和精度逐步提高。对于卷积神经网络，改进的方向有缩小卷积核和减小第一个卷积层的步长、在整个图像和多个不同缩放比例的图像上密集地测试网络。本文的主要检测不同深度的卷积神经网络的效果并由此提出了VGG网络，该网络可以方便地进行迁移学习。

|  |
| --- |
| 所谓密集地测试网络即是：将整个图片输入网络而不是对图像进行裁剪。 |

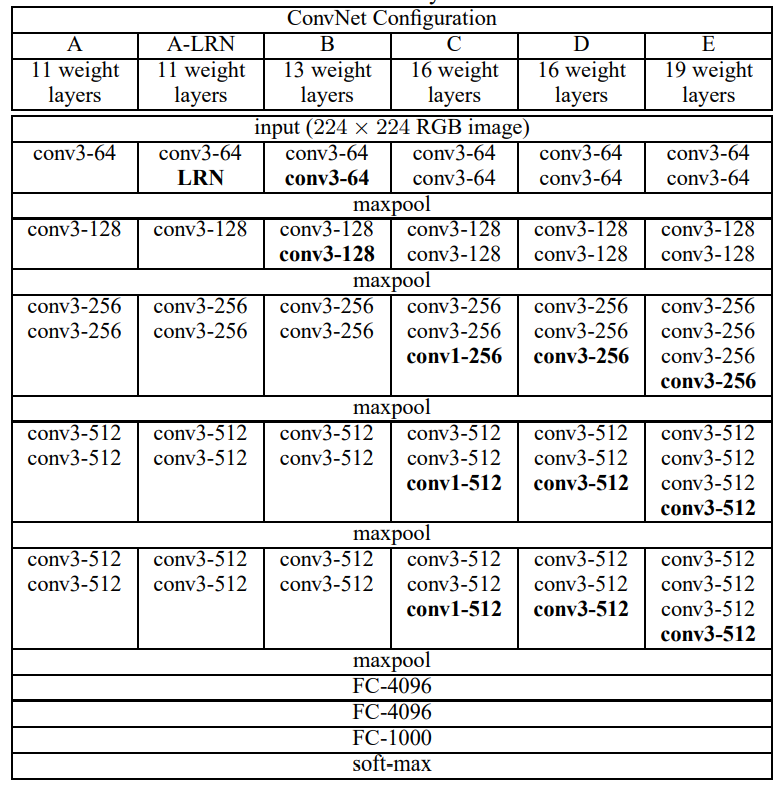
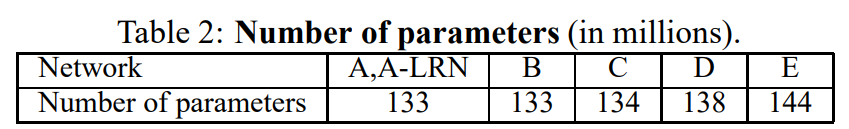
## 卷积神经网络的配置

所有网络的共同配置

在训练阶段，卷积神经网络的输入是一个固定大小的224×224的RGB图像，所做的唯一预处理是从每个像素中减去在训练集上计算的平均RGB值。然后图像通过一个卷积层，该卷积层的卷积核大小是3×3（是捕获上下左右数据的最小尺寸）；其中一个使用了1×1的卷积核，可以被看作是输入的线性变换。卷积步长固定为1个像素。在输入的图像周围填充1像素使得输入的大小和输出的大小一致。采用了5个最大池化层，每个池化层的大小为2×2，步长为2，用来实现输出减少一半的目的。所有的卷积层之后是3个全连接层，第一、二层有4096个节点，第三层有1000个节点，将第三层的输出经过softmax处理得到最后结果。所有隐藏层的激活函数都是ReLU函数。实验中不再使用LRN标准化层，因为作者证明它对图像分类问题没有提升，反而增加了内存消耗和计算时间。

|  |
| --- |
| 感受野也就是卷积核；池化层也就是下采样；3×3大小的卷积核是捕获上下左右数据的最小尺寸。  LRN全称为Local Response Normalization，即局部响应归一化层，LRN函数类似dropout，作为Relu激励函数之后防止数据过拟合而提出的一种处理方法。这个函数很少使用，基本上被类似的dropout这样的方法取代，最早出于AlexNet。 |

网络的不同配置

不同网络的深度不同，从A网络的11层到E网络的19层；作者提出的VGG网络的参数不大于浅层、卷积核更大的卷积网络。

讨论

相较以前使用较大的卷积核和步幅的用于图像识别的神经网络，作者提出的VGG网络使用了非常小的卷积核（3×3）和步幅为1的卷积。两个3×3卷积核等效于一个5×5卷积核，三个3×3卷积核等效于一个7×7卷积核。

使用三个3×3卷积核而不是一个7×7卷积核的理由：

* 由于每一层都有一个非线性激活函数，使得网络的非线性表达能力更好。
* 参数量减少：假设输入和输出都是C通道数，3个3×3的卷积核的参数量为27C^2，而一个7×7卷积核的参数量为49C^2。
* 这可以被看作将一个7×7网络强制施加正则化，迫使将其分解为3个3×3卷积核，其间包含非线性的激活函数。

C网络的1×1卷积核可以被看作是一种在不影响层的大小的条件下增加网络非线性程度的方法。

|  |
| --- |
| 所谓层的规模大小不变是输入和输出通道数量相同 |

之前也有人使用过小卷积核的卷积神经网络，但它们的深度没有作者提出的深。在2014年的ImageNet图像识别大赛上，获得第一名的GoogLeNet也是基于非常深和小的卷积核的卷积神经网络，但是它的网络结构复杂，并且在第一层中大量减少了输入的维度来减小计算参数，这样也就丢失了一部分信息。事实上，在单个网络分类精度方面，作者提出的模型优于GoogLeNet。

## 训练和评估网络的细节

训练

卷积神经网络的训练通常遵循AlexNet的训练过程。使用小批量梯度下降法。批量大小被设置为256，动量达到0.9。权重衰减（L2正则化的因子）被设置成5×10^-4。在前两个全连接层使用丢弃率为0.5的丢弃层。学习率最初被设置为10^-2，当验证集上的准确率停止提高时就将其除以10。总共学习率下降了三次，在370千次迭代后（74周期）学习停止。作者推测，尽管他提出的网络比AlexNet有更多的参数和更大的深度，但是他提出的网络收敛更快，原因如下：

* 由于更大的深度和更小的卷积核而产生的隐式正则化。
* 某些层的预初始化。

不好的初始化可能使网络的学习停止，因为深层网络梯度的不稳定性。为了解决这个问题，作者先训练了A模型，由于A模型较浅，故使用了随机初始化方法。随后在训练更深层次的网络时，作者使用了A网络已经训练好的模型参数来初始化前4个卷积层和全连接层，其他加的层是随机初始化的。在预训练好的层中，没有降低学习率，也允许它们发生改变。对于随机初始化，采用均值为0，方差为0.01的正态分布随机初始化。偏置项（即b）开始被设为0。在作者提交论文之后，发现在没有进行预训练的情况下可以使用Glorot&Bengio的随机初始化程序也能达到相应的效果。

|  |
| --- |
| 深层网络中梯度不稳定：如梯度消失和梯度爆炸 |

为了获得固定大小224×224的输入图像，作者从重新缩放的图片中对每张图片进行裁剪。为了进一步扩充训练集，作者使用了随机水平翻转和随机RGB颜色偏移（这也是AlexNet中使用的方法）。

训练图像重新缩放的说明如下：设S是图像等比例缩放后最小边的长度，由于输入被限定为224×224，所以S可以是任何不小于224的值，对于S = 224的情况，将输入整张图像，对于S > 224的情况，将裁剪图像的一部分。

作者提出了两种设置S的方法：

第一种是固定S，在本次实验中，作者评估了在两个固定尺度上的训练模型：S = 256，（在AlexNet和overflow中被使用）和S = 384。为了加速S = 384 网络的训练，它使用了S = 256的模型参数来进行初始化，并且使用了更小的初始学习率10^-3。

第二种方法是多尺度训练。每个训练图像都被单独缩放，S值介于[Smin, Smax]，本实验中使用了Smin = 256和Smax = 512。由于图像中的物体的大小是不同的，使用这种方法有助于网络训练，可以被视为通过图像缩放来增强训练集，即使用单个模型来识别更大范围尺度上的物体。为了提高速度，使用了S = 384的模型参数来初始化该模型。

测试

首先，输入图像被等比例地缩放，设它的最小边的宽度为Q，此处，Q不一定要等于S（对每个S使用不同的Q值将会对网络的性能有所提高）。然后将网络在测试图像集上密集地应用。首先对完全连接层转换为卷积层：前一个全连接层转换为7×7卷积层，后两个全连接层转换为1×1卷积层。然后将网络应用于未裁剪的图像集。结果得到的是通道数等于分类数、大小取决于输入图片的大小的结果。最后，对每个通道进行求和池化，得到每个类别的分数。

还通过图像的水平翻转来增强测试集，将两种结果的softmax进行平均来得到最终分数。

将整个图像应用于全卷积网络，不需要对图像进行裁剪，因为对每个裁剪进行计算效率低。但是，对图片进行裁剪可以提高网络精度，因为对输入图像进行了更精细的采样。裁剪图片和全图输入其实是互补的，全图输入增加对整个图片的全局感知，有更多的上下文信息，而图像裁剪维度不足时需要用0填充。

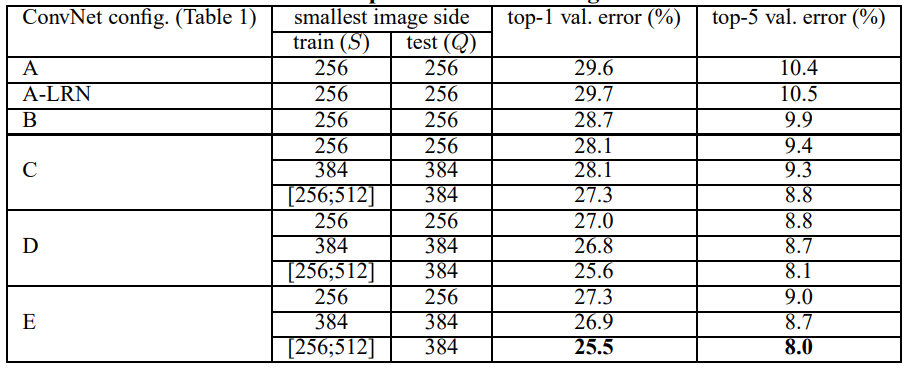
作者认为，在实践中使用图像裁剪并不能显著提高图像识别的准确率。但是为了作比较，还是使用了每张图片进行50个裁剪（25个来自原图，25个来自水平翻转版本），在三种缩放上共进行了150个裁剪，这与GoogLeNet的4个缩放大小上的144个裁剪是可比较的。

实现细节

作者基于C++的Caffe框架实现了代码，并对框架做了修改，以便可以在多块GPU上可以对多种缩放图像进行全图输入。通过把每个批量的数据分给不同的GPU来实现数据并行。在每块GPU的梯度计算完毕后，它们被平均来获取整个批量的数据的梯度。虽然最近提出了先进的方法来加速网络训练，对网络的每个层进行模型和数据的并行处理。但是作者认为他的方法已经足够使用了。在配备有4块NVIDIA Titan Black GPU的系统上，训练网络花费2~3周时间。

## 实验

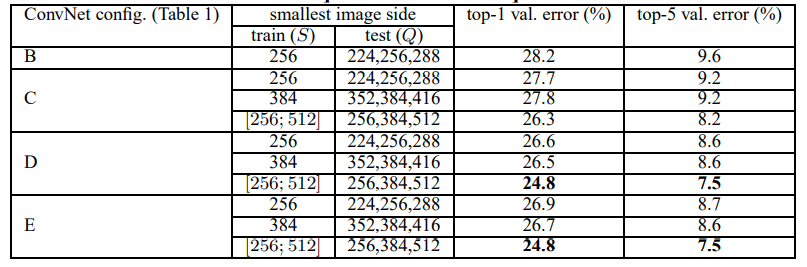
对于同一个缩放程度（即Q为定值）的评估

如果S为定值，则Q = S；如果S∈[Smin, Smax]，则Q = 0.5（Smin + Smax）。结果如下：

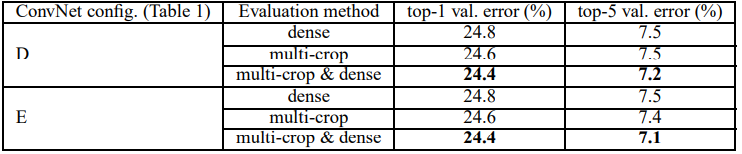
分析如下：

* 使用归一化LRN正则化并不能提升模型的精度，所以作者没有在更深层次上采用B-E架构。
* 分类误差随着网络的深度的增加而减少，C（包含3个1×1卷积层）虽然和D（使用3×3卷积层）深度相同，但是其性能较差，原因是使用1×1卷积核并没有捕获上下文信息。C比B好表明额外的非线性层确实对网络精度提升有帮助。
* 当深度达到19层时，模型的正确率达到了饱和，或许更深层的网络对于更大的数据集有帮助。
* 作者还比较了将B的两个3×3卷积层替换为一个5×5卷积层的大卷积核、浅层的卷积网络，错误率提高了7%，表明了小卷积核深层卷积网络比大卷积核浅层的卷积网络的网络性能更好。
* 在训练时缩放图片不同比例（S处于一定范围）比具有固定最小边的网络有更好的性能，这证实了通过尺度抖动来增强训练集确实有助于获取多尺度图像信息。

对于不同缩放程度（即Q处于一定范围）的评估

将模型在不同的Q上获得的结果取平均值作为后验概率。如果S是定值，Q∈[S - 32, 32，S + 32]；如果S是一个范围，Q∈[Smin, 0.5(Smin + Smax), Smax]，结果如下：

全图输入和图像裁剪

使用图像裁剪的性能略好于全图输入，如果将两种方法进行组合，则结果优于任何的单一一种，结果如下：

多模型融合

作者的最佳结果是通过仅组合两个模型来实现的——明显少于大多数模型，结果达到了6.8%，而同年的GoogleNet使用了7个模型，达到了6.7%。就单网性能而言，作者的体系结构达到了最佳。作者的网络没有偏离经典的卷积神经网络，仍然是传统的逐层串行堆叠网络结构。

## 总结

卷积网络深度的提升有利于提高大规模图像识别分类的准确性。在附录中，作者还展示了VGG网络模型可以很好地推广到广泛的任务和数据集。

# 问题回答

Q1 论文试图解决什么问题？

不同深度的网络对图像识别的影响。

Q2 这是否是一个新的问题？

之前也有论文使用了小卷积核和深度较深的卷积网络，但是都没有作者提出VGG的深。

Q3 这篇文章要验证一个什么科学假设？

深度较深、卷积核较小的卷积神经网络比以往的浅层大卷积核网络的性能更好。

Q4 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

对于卷积神经网络，一种改进方式是改进数据的处理方式，如密集地测试图形或裁剪图像、对图像进行水平翻转、对图像进行标准化等；另一种改进方式是使用小卷积核、较深的网络，牛津视觉小组的成员在这一领域内值得关注。

Q5 论文中提到的解决方案之关键是什么？

使用小卷积核、使用较深的卷积神经网络。

Q6 论文中的实验是如何设计的？

通过测试不同深度的网络来证明更深的网络有更好的效果；同时数据处理方式也参照AlexNet：训练阶段分为S为定值和S处于一定范围两种训练方式；测试阶段分为Q为定值和Q处于一定范围内两种训练方式。

Q7 用于定量评估的数据集是什么？代码有没有开源？

ImageNet等数据集；代码开源。

Q8 论文中的实验及结果有没有很好地支持需要验证的科学假设？

通过比较不同深度的网络效果，可以看出较深的网络有更好效果。

Q9 这篇论文到底有什么贡献？

证明了LRN没有效果；提出用连续的3X3的卷积层代替5X5以及7X7等更加大的卷积层，进而提升网络的深度（在2015年十几层的网络是非常深的），同时网络的效果也提高了。

Q10 下一步呢？有什么工作可以继续深入？

网络的参数量大，难以训练；当深度加深继续加深时，网络会出现梯度消失爆炸、网络退化等问题。