InceptionNet V2/V3:2015

Rethinking the Inception Architechture for Computer Vision

Google & College London

摘要

作者指出卷积神经网络已经在计算机视觉中的很多任务上表现优良，同时低参数量和网络的效率成为了卷积神经网络的发展方向，之后介绍了自己的工作，对比他们在ILSVRC2012提出的模型比较，通过集成新的Inception模块在数据集上面达到了3.5%的top5错误率和17.3%的top1错误率

论文出发点或背景

虽然VGG提取特征的效果很好、结构简单，但是在计算时的消耗也非常大；GoogleNet就是为了在算力受限的平台上运行而发明的一种新的卷积神经网络架构。对比AlexNet，GoogLeNet只有十二分之一的参数量，而且VGG的参数时AlexNet的三倍。Inception模块的使用使得在内存或容量受限的设备中使用卷积神经网络进行视觉识别成为了可能。然而Inception模块的复杂性也导致了对网络进行修改是件很难的事情。同时GoogLeNe文章中也没有说清楚各种设计为什么是有效的，如果单纯增加网络的深度，单纯堆叠模块的话，有时候反倒会导致损失函数值上升。

在本文中首先阐述了几条一般性原则和优化思想，这些方法和原则被证明是扩展卷积网络的有效方法。

GoogLeNet在一大优点就是在计算的时候进行了降维，这可以被看作是以一种计算效率高的方式对卷积进行分解的一种特殊情况。

论文创新思路

文章阐述的四个一般性原则：

1.在网络早期要避免网络的表征瓶颈，特征图的表示大小应该是从输入到输出逐渐减小的，应该避免极端压缩

2.高维表示在网络中更容易进行局部处理

3.低纬度嵌入聚合之后，对网络不会产生严重的不利影响，甚至还会促进更快的学习

4.好的网络应该能平衡网络每层上的深度和宽度，计算预算应该在网络的深度和宽度之间保持平衡的分布

卷积分解：

1.两个3×3卷积核的感受野可以和一个5×5的卷积核感受野对等，三个3×3可以和一个7×7的对等，通过这种将大卷积核分解为小卷积核可以减少网络的参数。网络的表达效果没有变差。

2.空间不对称卷积分解：小的卷积核也可以分解为更小的卷积核，比如3×3可以分解为2×2的，但是通过计算发现，如果将3×3卷积核不对称分解为1×3和3×1的卷积核会减少更多的参数。同时通过实验发现，这种分解在网络浅层的表现不是很好，但是在中等大小的特征图上面可以有很好的结果。

辅助分类器的使用：

1.发现辅助分类器在训练的早期并没有提高收敛性

2.去除辅助分类器之后发现对网络的精度没有太大的影响，辅助分类器可以被看作是一个正则化方式。

高效的空间降维方式

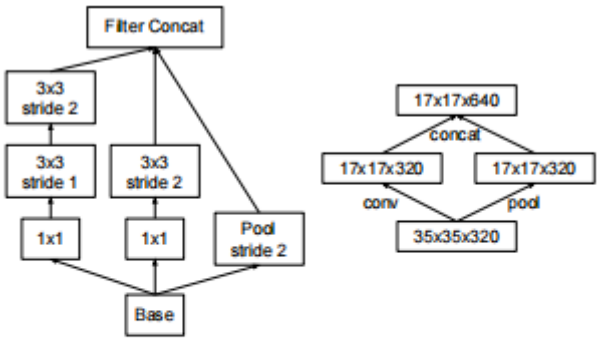

Figure 10. Inception module that reduces the grid-size while expands the filter banks. It is both cheap and avoids the representational bottleneck as is suggested by principle 1. The diagram on the right represents the same solution but from the perspective of grid sizes rather than the operations.

标签平滑LSR：一种新的正则化方式

论文方法的大概介绍

小卷积核等效为大卷积核

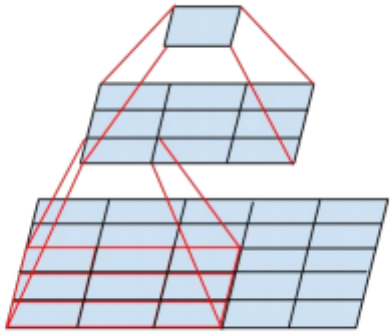

Figure 1. Mini-network replacing the 5 × 5 convolutions.

非对称分解卷积核

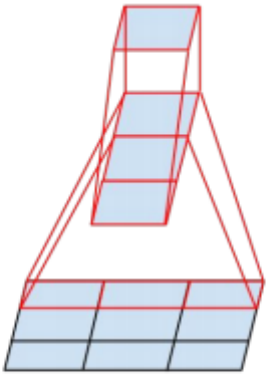

Figure 3. Mini-network replacing the 3 × 3 convolutions. The lower layer of this network consists of a 3 × 1 convolution with 3 output units.

替换大卷积核后的模块

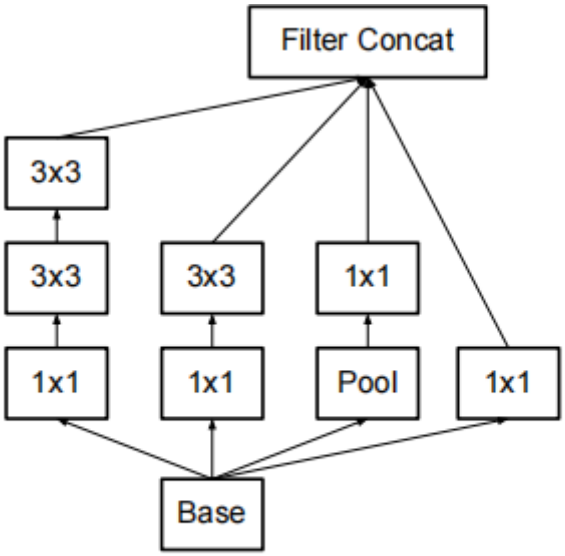

Figure 5. Inception modules where each 5×5 convolution is replaced by two 3×3 convolution, as suggested by principle 3 of Section 2

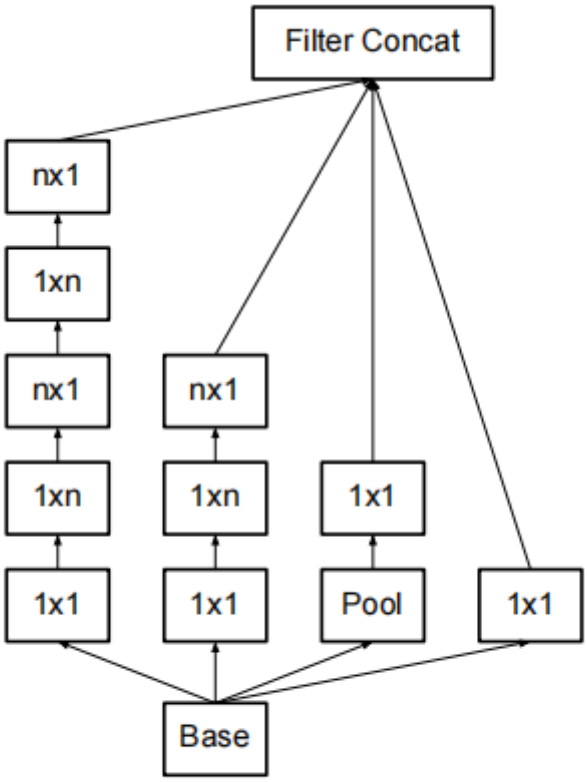

Figure 6. Inception modules after the factorization of the $n \times n$ convolutions. In our proposed architecture, we chose $n = 7$ for the 17×17 grid. (The filter sizes are picked using principle 3)

辅助分类器：

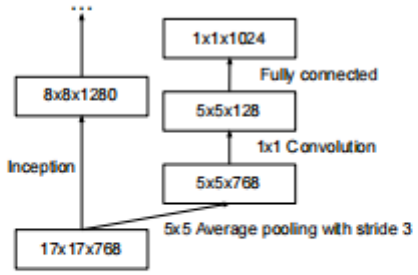

Figure 8. Auxiliary classifier on top of the last 17×17 layer. Batch normalization[7] of the layers in the side head results in a 0.4% absolute gain in top-1 accuracy. The lower axis shows the number of itertions performed, each with batch size 32.

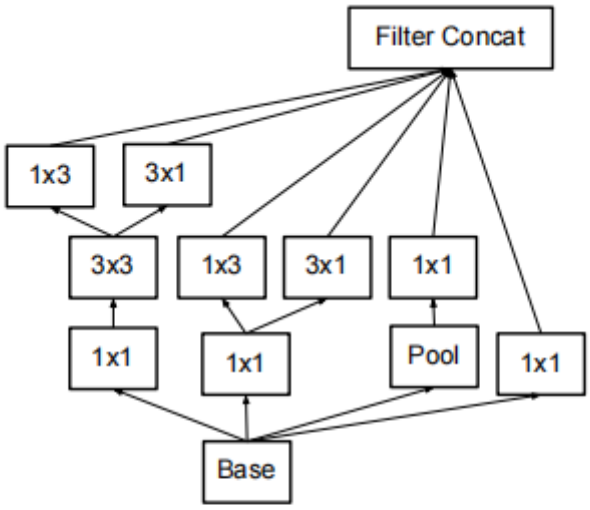

Figure 7. Inception modules with expanded the filter bank outputs. This architecture is used on the coarsest (8×8) grids to promote high dimensional representations, as suggested by principle 2 of Section 2. We are using this solution only on the coarsest grid, since that is the place where producing high dimensional sparse representation is the most critical as the ratio of local processing (by 1×1 convolutions) is increased compared to the spatial aggregation.

Inception V2 ： 虽然有42层之深，但是网络的消耗只是GoogLeNet的2.5倍，并且比VGGNet更高效

部分参数配置：

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	As in figure 5	35×35×288
5×Inception	As in figure 6	17×17×768
2×Inception	As in figure 7	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

Table 1. The outline of the proposed network architecture. The output size of each module is the input size of the next one. We are using variations of reduction technique depicted Figure 10 to reduce the grid sizes between the Inception blocks whenever applicable. We have marked the convolution with 0-padding, which is used to maintain the grid size. 0-padding is also used inside those Inception modules that do not reduce the grid size. All other layers do not use padding. The various filter bank sizes are chosen to observe principle 4 from Section 2.

实际效果

低分辨率输入情况下的模型性能：

Receptive Field Size	Top-1 Accuracy (single frame)
79 × 79	75.2%
151 × 151	76.4%
299 × 299	76.6%

Table 2. Comparison of recognition performance when the size of the receptive field varies, but the computational cost is constant.

结果：虽然低分辨率的网络需要更长的训练时间，但是其最终结果的指标核高分辨率相当接近

Network	Top-1 Error	Top-5 Error	Cost Bn Ops
GoogLeNet [20]	29%	9.2%	1.5
BN-GoogLeNet	26.8%	-	**1.5**
BN-Inception [7]	25.2%	7.8	2.0
Inception-v2	23.4%	-	3.8
Inception-v2 RMSProp	23.1%	6.3	3.8
Inception-v2 Label Smoothing	22.8%	6.1	3.8
Inception-v2 Factorized 7 × 7	21.6%	5.8	4.8
Inception-v2 BN-auxiliary	**21.2%**	**5.6%**	4.8

Network	Crops Evaluated	Top-5 Error	Top-1 Error
GoogLeNet [20]	10	-	9.15%
GoogLeNet [20]	144	-	7.89%
VGG [18]	-	24.4%	6.8%
BN-Inception [7]	144	22%	5.82%
PReLU [6]	10	24.27%	7.38%
PReLU [6]	-	21.59%	5.71%
Inception-v3	12	19.47%	4.48%
Inception-v3	144	**18.77%**	**4.2%**

Table 4. Single-model, multi-crop experimental results comparing the cumulative effects on the various contributing factors. We compare our numbers with the best published single-model inference results on the ILSVRC 2012 classification benchmark.

Network	Models Evaluated	Crops Evaluated	Top-1 Error	Top-5 Error
VGGNet [18]	2	-	23.7%	6.8%
GoogLeNet [20]	7	144	-	6.67%
PReLU [6]	-	-	-	4.94%
BN-Inception [7]	6	144	20.1%	4.9%
Inception-v3	4	144	**17.2%**	**3.58%***

Table 5. Ensemble evaluation results comparing multi-model, multi-crop reported results. Our numbers are compared with the best published ensemble inference results on the ILSVRC 2012 classification benchmark. *All results, but the top-5 ensemble result reported are on the validation set. The ensemble yielded 3.46% top-5 error on the validation set.

个人理解

1.提出了四点设计网络结构时的原则

- 在网络早期要避免网络的表征瓶颈，特征图的表示大小应该是从输入到输出逐渐减小的，应该避免极端压缩
- 高维表示在网络中更容易进行局部处理
- 低纬度嵌入聚合之后，对网络不会产生严重的不利影响，甚至还会促进更快的学习
- 好的网络应该能平衡网络每层上的深度和宽度，计算预算应该在网络的深度和宽度之间保持平衡的分布

2.对卷积核进行分解，通过低秩进行近似，大大减少参数量

3.n×1和1×n卷积核可以看作是分别从宽度和长度上提取特征，就好像视频理解中P3D中对卷积进行时空上的拆解

4.提出了新的正则化方法：标签平滑