Inception_V1:2014

Going Deeper with Convolutions

摘要：

我们提出了一种叫做Inception的卷积神经网络架构，并且在ILSVRC14竞赛的分类和检测任务中取得了SOTA。这种网络的主要特点是提升计算资源的利用。通过设计，我们在保持计算预算不变的情况下，提升了网络的宽度和深度。网络的设计符合$the\ Hebbian\ principle$和多尺度视觉处理原则。在ILSVRC14中我们提交的是一个叫做GoogLeNet的神经网络网络，由22层组成，并在分类和检测任务中进行了性能评估。

论文出发点或背景：

三年以来计算机视觉的巨大发展，不仅仅是硬件上的提升和数据集的增大，发现了很多有用的网络结构也是促进其发展的重要原因。我们的GoogLeNet比两年前比赛的获胜网络参数小了12倍的同时达到了更好的准确率。在目标检测中我们摒弃了使用更大更深的网络的思路，而是将**网络的结构和经典的计算机视觉协同起来。**

另外一点是**移动设备和嵌入式设备的流行**，算法的效率，尤其是算法的耗能和内存的使用显得十分重要。本文的模型考虑并不只是考虑模型的准确率，将其实际使用也考虑了进去。

文章的名字来源于Lin在 internet meme 上著名的“we need to go deeper”,在我们的网络中deep有着两层含义，一是我们的Inception 模块在直观上是更深的，另一方面是我们的工作基于Arora的理论工作。

从LeNet-5开始，卷积神经网络就进入到了**多个卷积层（包含卷积、归一化、最大池化）后接全连接层的范式。**由此范式出现的变体在各大数据集上面获得了不错的效果，但是在更大的数据集上面，人们的思路逐渐趋向于加大卷积层的层数和卷积尺寸，同时加上Drop out解决过拟合问题。**单纯加大网络的尺寸存在两个缺点：1.有更多的参数，网络容易过拟合；2.计算资源的使用明显增加**

解决上述两个问题的一个思路就是引入稀疏性，用稀疏的层代替全连接层（卷积层也是一种模仿生物系统得到的稀疏层），理论支持：Arora:Hebbian priciple.然而目前的计算设施对于非均匀稀疏数据的数值运算中没有很好的支持。

受到灵长类动物视觉皮层的神经科学模型的启发，Serre等人通过**使用不大小固定的Gabor滤波器来处理多尺度**，我们的工作也使用了相同的策略。不同的是，我们的滤波器是由学习得到的，22层inception module的堆叠得到了GoogLeNet

在Network in Network 中**使用1×1卷积核以提高卷积神经网络的表征能力，增加网络的深度。**我们在我们的网络中也大量使用了这种方法，但是我们的**主要目的是通过1×1卷积核降维进而消除计算瓶颈，否则将会限制我们网络的深度。**

目前，目标检测的SOTA是R-CNN，它将目标检测任务分解为

利用低层次的特征例如颜色和纹理，以类别未知的方式对物体生成建议的区域，然后通过使用CNN进行识别。我们在目标检测赛道也进行了类似的两阶段方法并进行了改进。例如使用多盒预测对更好的对象边界盒召回以及更好地对边界盒进行分类的集成方法。

论文创新思路：

Inception结构的主要思想是考虑如何使用卷积神经网络的最优局部稀疏结构被现有的密集量所替换和近似，我们所需要的就是找到并在空间上重复它。

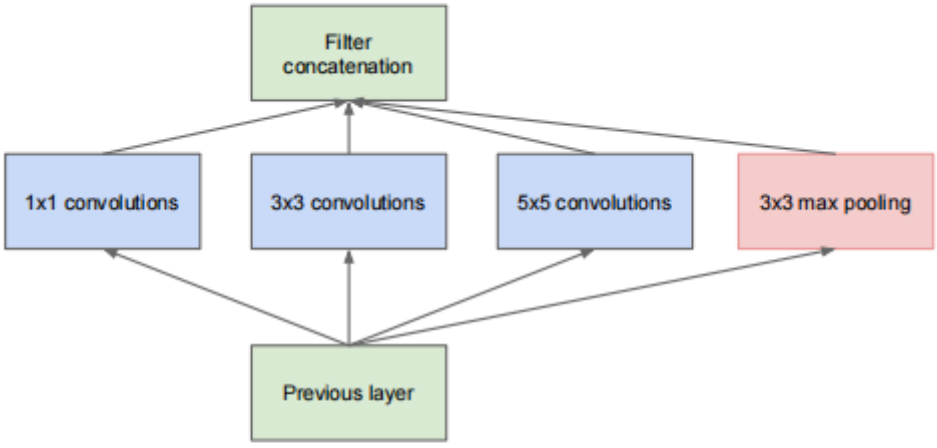
通过1×1卷积核降维进而消除计算瓶颈，除了被用作减少量外，它们还包括使用ReLU激活函数，使它们具有双重用途。

使用不同大小的卷积核来多尺度处理

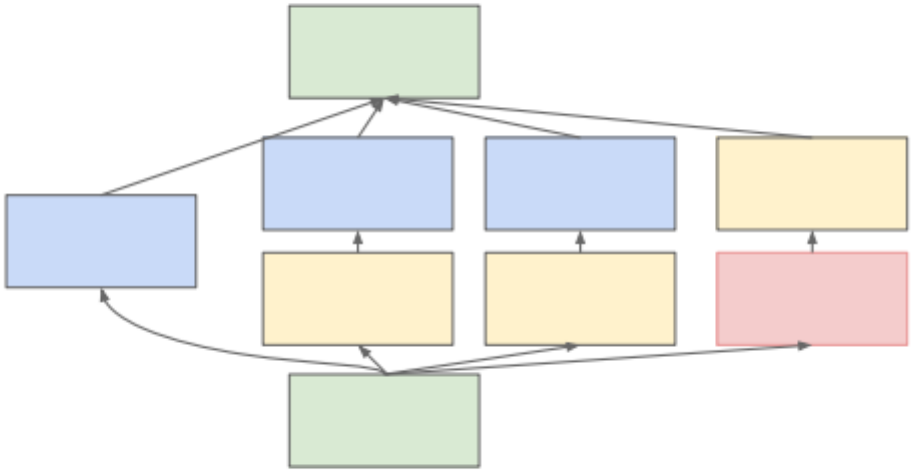
当前版本的初始架构被限制在过滤器大小1×1,3×3和×5×5；这个决定更多的是基于方便而不是必要。这也意味着，建议的架构是所有这些层的组合，**它们的输出滤波器组连接成一个单一的输出向量，形成下一阶段的输入。**此外，**由于池化操作对于当前卷积网络的成功至关重要，因此建议在每个这样的阶段添加一个替代的并行池化路径也应该产生额外的有益效果**

该设计遵循了实践直觉，即视觉信息应该在不同的尺度上进行处理，然后进行聚合，以便下一阶段能够同时从不同的尺度上提取特征。

论文方法的大概介绍：



(a) Inception module, naïve version



(b) Inception module with dimensionality reduction

Figure 2: Inception module

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Table 1: GoogLeNet incarnation of the Inception architecture.

所有的卷积，包括那些在初始模块内的卷积，都使用了ReLU

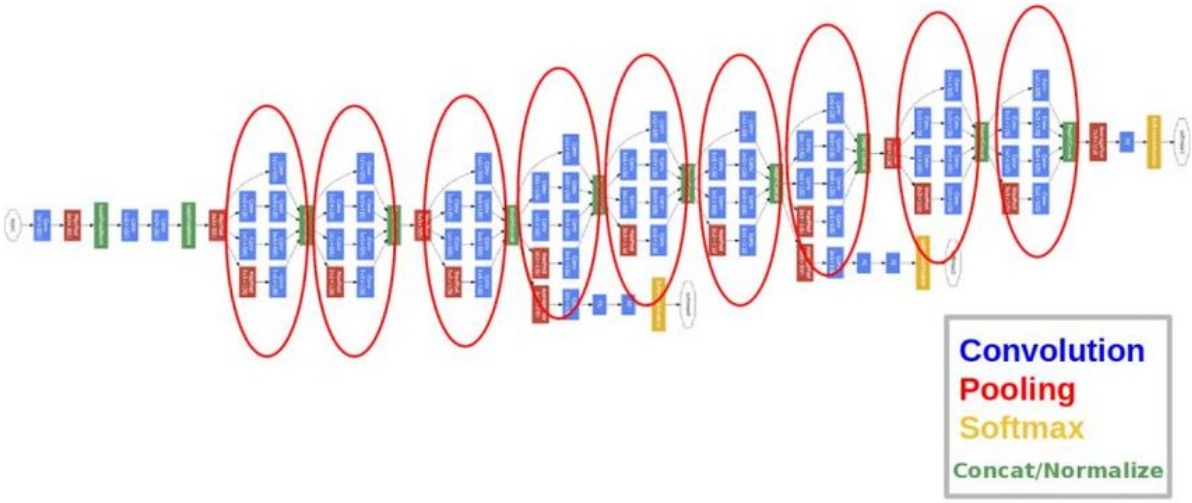
“#3×3reduce”和“#5×5reduce”表示在3×3和5×5卷积之前的减少层中使用的1×1卷积核的数量

辅助分类器的参数配置：

- An average pooling layer with 5×5 filter size and stride 3, resulting in an $4\times4\times512$ output for the (4a), and $4\times4\times528$ for the (4d) stage.
- A 1×1 convolution with 128 filters for dimension reduction and rectified linear activation.
- A fully connected layer with 1024 units and rectified linear activation.
- A dropout layer with 70% ratio of dropped outputs.
- A linear layer with softmax loss as the classifier (predicting the same 1000 classes as the main classifier, but removed at inference time).

A schematic view of the resulting network is depicted in Figure 3.

训练策略：动量为0.9的随机梯度下降策略 + 每8个epoch学习率降低4%，使用Polyak平均来创建在推理时使用的最终模型。



实际效果：

该网络的设计考虑到了计算效率和实用性，因此推理可以在单个设备上运行，甚至包括那些计算资源有限的设备，特别是低内存占用的设备。

我们发现，将全连接层替换为平均池使top1的精度提高了约0.6%，但是即使去除完全连接层，使用dropout仍然是必要的

实验表明，辅助分类器（损失以一个折扣权重加到网络的总损失中（辅助分类符的损失被加权为0.3））的影响相对较小（约0.5%），只需要其中一个就能达到相同的效果。

分类任务：

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

Table 2: Classification performance.

Number of models	Number of Crops	Cost	Top-5 error	compared to base
1	1	1	10.07%	base
1	10	10	9.15%	-0.92%
1	144	144	7.89%	-2.18%
7	1	7	8.09%	-1.98%
7	10	70	7.62%	-2.45%
7	144	1008	6.67%	-3.45%

Table 3: GoogLeNet classification performance break down.

目标检测任务：

Team	Year	Place	mAP	external data	ensemble	approach
UvA-Euvision	2013	1st	22.6%	none	?	Fisher vectors
Deep Insight	2014	3rd	40.5%	ImageNet 1k	3	CNN
CUHK DeepID-Net	2014	2nd	40.7%	ImageNet 1k	?	CNN
GoogLeNet	2014	1st	43.9%	ImageNet 1k	6	CNN

Table 4: Comparison of detection performances. Unreported values are noted with question marks.

Team	mAP	Contextual model	Bounding box regression
Trimps-Soushen	31.6%	no	?
Berkeley Vision	34.5%	no	yes
UvA-Euvision	35.4%	?	?
CUHK DeepID-Net2	37.7%	no	?
GoogLeNet	38.02%	no	no
Deep Insight	40.2%	yes	yes

Table 5: Single model performance for detection.

个人对这篇论文的理解：

1.文章的很多思路都来源于生物学，比如稀疏结构和多尺度提取信息。保持网络的稀疏性可以避免过拟合，比如ReLU激活函数的使用（AlexNet提出），还有通过卷积层来代替全连接层，也是一种稀疏结构。多尺度提取信息可以更大程度结合学习到的特征，就好像之前看到关于transformer切分patch的操作，有研究表明，单张图片切分为更多的patch网络的识别效果更好，以及关于卷积神经网络中的感受野的相关问题，将不同尺度的特征进行融合，网络可以学习更好

2.Inception Net 通过将网络变宽变深来解决问题，增大了网络的表达能力，神经网络的结构设计开始朝着多支路方向演化。同时引入1×1卷积核作为数据降维模块，减少了运算时的算力要求。之后常见于各种网络设计之中。