AlexNet:2012

ImageNet Classification with Deep Convolutional Neural [paper]

University of Toronto

摘要：

在ImageNet LSVRC-2010比赛中，训练了一个卷积神经网络对1000种类别的数据进行分类，top-1和top5的error rate 分别为37.5%和17.0%,优于之前的所有方法。使用的网络包含五个卷积层（其中一部分带有最大池化层），有6千万参数和六十五万个神经元，三层全连接层，最后一层输出1000个维度，后接softmax得到每个类别的置信度。为了加快训练，使用了non-saturating neurons(非饱和神经元)并且在GPU上进行训练（提升卷积运算的效率），为了降低发生过拟合的风险，采用了最近出现的一种技术-drop out，效果显著。在ILSVRC-2012比赛中，使用该模型的变体，在测试集的top5 error_rate达到了15.3%，远超比赛第二名的26.2%

论文出发点或背景

出发点和背景

现有的物体识别方法大都是通过使用机器学习算法实现的，为了提升模型的表现，我们可以采用更大的数据集，可以使用效果更好的模型，也可以使用更好的方法来防止过拟合。在小的数据集上，机器的表现已经达到了和人近似的水平，（即使通过数据增强来扩充小数据集）但是对于一些真实场景下的依旧会表现出相当大的变数，所以使用更大的训练集是十分有必要的。同时，小数据集的缺点已经被所有人认识到了，似乎当前情况下采集数量为百万级别的标注数据集成了目前唯一的解决方法。新的数据集中的代表有LableMe和ImageNet.

要学习大数据集的特征，我们的网络就需要足够的"学习容量",**然而，对象识别任务的巨大复杂性意味着，这个问题即使是由像ImageNe这样大的数据集也不能指定**，所以我们的模型需要有一些先验知识对我们没有的数据进行补偿，**卷积神经网络就是这样一类网络，它的网络容量可以被深度和宽度指定，对于图像也有很多正确的假设，对比标准的前馈神经网络，卷积神经网络有着更少的连接和参数，这也使得他们更为容易训练。**理论上的最佳表现可能稍微差一点（没理解这句话在做什么对比，但是我感觉都是在阐述卷积神经网络针对图像识别任务的优越性）

论文创新思路

创新思路：

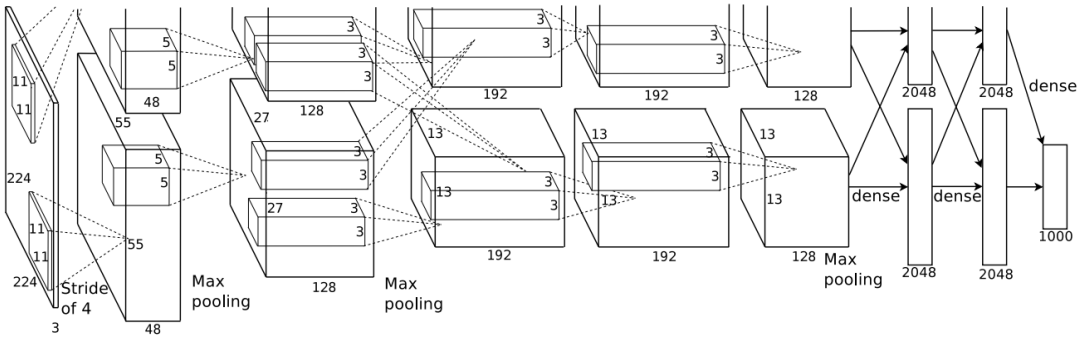

Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

1.引入非线性激活函数ReLU

A four-layer convolutional neural network with ReLUs **(solid line)** reaches a 25% training error rate on CIFAR-10 six times faster

than an equivalent network with tanh neurons **(dashed line)**. The learning rates for each network were chosen independently to make training as fast as possible. No regularization of any kind was employed. The magnitude of the effect demonstrated here varies with network architecture, but networks with ReLUs consistently learn several times faster than equivalents with saturating neurons.

在CIFAR-10数据集上，分别使用tanh(x)和ReLU(x)函数进行一个四层卷积神经网络的训练，发现ReLU比tanh到达25% error_rate(出现收敛)快了大约6倍

2.在多个GPU上面进行运算

The parallelization scheme that we employ essentially puts half of the kernels (or neurons) on each GPU, with one additional trick: the GPUs communicate only in certain layers.

一张GTX 580 只有3GB的显存，限制了训练网络的大小，所以AlexNet将一般的kernel放在了A支路，另一半放在了B支路，在固定的层之间进行信息交互。

3.局部相应归一化（LRN）（目前很少使用）

ReLU的优点：不需要对输入进行归一化处理而防止饱和

$$b_{x,y}^{i} = a_{x,y}^{i} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^{j})^{2} \right)^{\beta}$$

论文中使用了k=2，n=5，α=10−4和β = 0.75。

4.overlapping pooling

在普通的CNN中池化层的步长和核的大小相等，实验中令步长小于核的大小，发现top1和top5错误率分别下降了0.4%和0.3%，同时发现这种重叠池化可以让网络过拟合风险变小

5.数据增强

采用了两种方法：

1.The fifirst form of data augmentation consists of generating image translations and horizontal reflections.

We do this by extracting random 224 × 224 patches (and their horizontal reflections) from the 256×256 images and training our network on these extracted patches。

在256×256的图片上截取5张224×224的大小的图片，然后再水平上进行翻转，可以得到10张符合网络输入的图片，最后对这十张的预测进行平均输出

2.The second form of data augmentation consists of altering the intensities of the RGB channels in training images.

改变RGB通道上的强度

通过PCA，得到每个通道上的主成分，然后乘以一个高斯随机变量（均值为0，方差为0.1）

a Gaussian with mean zero and standard deviation 0.1. Therefore to each RGB image pixel $I_{xy} = [I_{xy}^R, I_{xy}^G, I_{xy}^B]^T$ we add the following quantity:

$$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

pi和λi分别为RGB像素值的3×3协方差矩阵的特征向量和特征值，αi为上述随机变量。

6.Drop out

有一种集成学习的思想，每次有一个新的输入，就会对应一个新的网络架构

每个神经元以0.5的概率被drop out，在test阶段，带有drop out的层中的每个神经元都会被使用，他们的输出会被乘以0.5。

At test time, we use all the neurons but multiply their outputs by 0.5, which is areasonable approximation to taking the geometric mean of the predictive distributions produced by the exponentially-many dropout networks.

论文方法的大概介绍

整体架构：

五个卷积层，三个全连接层，最后一层输出经过softmax函数作为预测的概率值输出

第二、四、五层卷积层仅仅和同GPU的上一个卷积层存在映射关系，第三个卷积层与第二个卷积层所有的卷积核存在映射关系，全连接层与上一层所有神经元存在映射关系，局部相应归一化应用在第一二卷积层。ReLU激活函数在每个卷积层和全连接层中都有使用。

输入224×224×3，经过第一个卷积层（c_in = 3 ,c_out = 96,size = (11,11),stride=4），得到两个55×55×48的特征图，分别在两个GPU上（55×55×48），第二层卷积层，先对上一层的进行了LRN，然后用256个(5×5×48）大小的卷积核进行卷积，之后经过最大池化层，每个GPU上的特征图为27×27×128，经过LRN和最大池化后接第三层卷积层（c_in=256,c_out=384,size=(3,3)）得到13×13×192的特征图，第四个卷积层（c_in = 192，c_out = 384,size(3,3)），之后第五层的卷积层为(c_in=192,c_out=256,size=(3,3)),经过最大池化层以后，展平进入全连接层，对应的维度为4096.

训练细节：

一个batch有128个样本

momentum = 0.9

weight_decay = 0.0005 weight decay here is not merely a regularizer:it reduces the model's training error.

$$v_{i+1} := 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot w_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w}\Big|_{w_i} \right\rangle_{D_i}$$

$$w_{i+1} := w_i + v_{i+1}$$

每一层的权重随机初始化复合均值为0，标准差为0.01的高斯分布

用常数1初始化了第二第四第五卷积层以及全连接层中的bias,其余用0进行初始化

学习率初始化为0.01，当错误率随着训练增大时学习率就会减小十倍，所有的层采用一样的学习率。训练90轮，花费六天

实际效果

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

个人对这篇论文的理解

1.证明了网络深度对于网络性能的影响（在discussion中提出了这个观点）

2.提出了LRN，但是后面好像被证明效果不是很好，后期出现的BN等归一化效果更好

3.在效果上超越了传统方法，但是受限于当时的硬件

4.通过使用重叠的窗口进行池化，pvt-v2在划分patch的时候也使用了这种方式，感觉能更大程度上让不同窗口间进行信息交互，保留一些局部连续性

单词

单词	释义
high-resolution	高分辨率的
state-of-art	最高水平，最高水准

单词	释义
variant	变种
label-preserving transformation	标签保留转换（可以理解为数据增强以后标签不变）
exhibit	表现出
considerable	相当大的
tens of thoudands of	成千上万的
shortcoming	缺点
specified	规定的
breadth	宽度，幅度，广泛性
slight	轻微的
prohibitively	禁止地；过高地；过分地
facilitate	使更容易，使便利；促进，推动
inherent	内在的，固有的，固定属于（某人）的，相同的
inferior	次的，较差的，低等的，自卑的
in terms of	依据；按照；在……方面；以……措词
saturating	饱和的
scheme	机制，计划，体制，密谋
columnar	柱形的，圆柱的
desirable	令人满意的
property	属性，特性
adjacent	邻近的，毗邻的
resemblance	外表，相似，相像
distinct	不同的
substantial	严重的，牢固的
intensity	强度
capable	有能力的，升任的
corresponding	相应的，相关的，通信
approximately	大约，近似
illumination	照明
conjunction	结合，同时发生

单词	释义
interchangeably	可交换地
densely-sampled	密集抽样的