ImageNet Classification with Deep Convolutional Neural Networks

简介

对2010年ImageNet上120万高分辨率图像进行1000种分类，比当时最先进的结果表现都要好

神经网络有6000万个参数和65万个神经元，包含五个卷积层，有些后面还有最大池化层

为加速训练，使用非饱和神经元和对卷积操作非常有效的GPU实现

为减少全连接层的过拟合，使用dropout正则化

目录

Introduction

The dataset

The architecture

Reducing overfitting

Details of learning

Results

Discussion

Introduction

* 目标识别已经大量使用机器学习的方法
* 有标签的图片数据还很少，小规模数据对于数字识别来说足够，并已经接近人类表现。
* 现实世界的目标和数字有很大不同，故需要更大的数据集。如LabelMe和ImageNet
* 卷积神经网络性能表现良好，连接数和参数都较少，易于训练
* 本文使用ImageNet数据集训练了一个大型卷积神经网络，我们提升了网络的表现，减少其训练时间，我们用了许多有效的技巧来防止过拟合。最终的网络包含五个卷积层和三个全连接层。网络的深度十分重要
* GPU显存和训练时间限制了网络的规模，更快的GPU和更大的训练集会使结果更好

The dataset

* 神经网络需要固定的输入维度，故将图片下采样至256×256。
* 输入矩形图片，先将短边重新调整为256，再裁剪中心的256×256块
* 在训练集上对像素减去平均活跃度，不对图片进行其他预处理，故本模型是在原生RGB像素值上训练出的

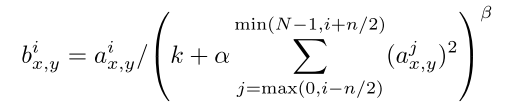
The architecture

使用ReLU（Rectified Linear Units）非线性函数代替tanh和sigmoid函数，基于梯度下降法的训练时间大大减少。

在多块GPU上训练

局部响应归一化（Local Response Normalization）

a表示第i个核在位置（x,y）运用激活函数ReLU后的输出，n是同一位置上临近的kernal map的数目，N是kernal的总数。参数K,n,alpha，belta都是超参数

b是归一化后的值。对局部神经元的活动创建竞争机制，使得其中响应比较大的值变得相对更大，并抑制其他反馈较小的神经元，增强了模型的泛化能力。

重叠池化Overlapping Pooling

步长为s，池化filter大小为z×z，如果s = z，就是普通的局部池化，如果s > z，就是重叠池化

相邻池化窗口之间会有重叠，可以提高网络性能，并且有助于缓解过拟合

整体结构

网络含有8个权重层，其中五个卷积层和3个全连接层，最后一个全连接层连接到1000类的softmax分类器上，我们的网络最大化多项逻辑回归目标，这相当于最大化预测分布下正确标签的对数概率的训练案例的平均值。

在每个卷积层和全连接层之后都会应用ReLU非线性函数

输入图片尺寸为224×224×3（实际上应为227×227），第一个卷积层有96个输入为11×11×3的过滤器（filter），卷积步长（stride）为4个像素，第一层得到的输出为55×55×96。最大池化后为27×27×96。第二层卷积层使用的filter为5×5×96，有256个。第三层有384个尺寸为3×3×256的filter，第四层384个尺寸为3×3×384的filter，第五层256个尺寸为3×3×384的filter。每个全连接层有4096个神经元

减少过拟合reducing overfitting

Data Augmentation

第一种形式的数据增强包括生成图像平移（image translations）和水平反射（horizontal reﬂections）。

通过在256×256的图像中随机提取224×224大小的块（以及其水平反射），并在这些提取的块上训练神经网络，这将我们的训练集扩大了2048倍。

测试的时候，通过提取10个224×224的块并对其进行预测，以及对网络的softmax层对10个块的预测求平均值。

第二种方式包括改变训练图像的RGB通道的强度。

我们通过ImageNet训练集在RGB像素值上进行PCA（principal Component Analysis)），对于每一个训练图片，我们增加其多倍的主部分，其大小与相应的特征值乘以一个随机数成比例，该随机数服从高斯分布N(0, 0.01)

该方案近似抓取自然图像的重要特性，即，目标身份对于照明的强度和颜色的变化是不变的。

Dropout（随机失活）

让每个神经元的输出为0，其概率为0.5。

使用dropout的神经元对前向传递没有贡献，并且也不参与反向传播。减少了神经元复杂的共同适应，因为神经元不能依赖于其他特定神经元的存在，可以让神经元学习更有力的特征。在测试时，我们使用所有神经元但是将它们的输出乘以0.5。

Dropout大约使收敛所需的迭代次数加倍。

Details of learning

使用随机梯度下降法来训练模型。

我们使用随机梯度下降训练我们的模型，批量大小为128个示例，动量为0.9，权重衰减为0.0005。 我们发现，这种少量的权重衰减对于模型学习很重要。换句话说，这里的重量衰减不仅仅是一个正则化器：它减少了模型的训练误差。

我们从高斯分布N(0, 0.0001)初始化每层中的权重。 我们用常数1初始化了第二，第四和第五卷积层以及全连接层中的神经元偏差。这种初始化通过为ReLU提供正输入来加速学习的早期阶段。我们用常数0初始化剩余层中的神经元偏差。