一介绍

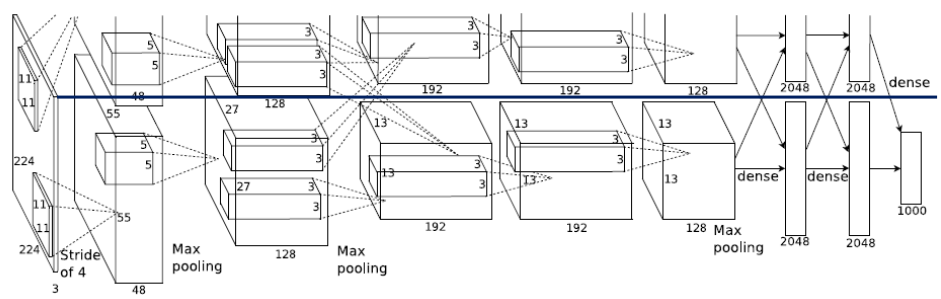
在图像识别方面，机器学习方法往往是提取图像特征，如图像的SIFT，HoG特征，再使用机器学习方法学习这些特征，如SVM，来进行图像中物体的识别。与机器学习相比深度学习不需要提取图像的特征，可以直接将图像作为输入来学习。随着计算机硬件的发展，计算能力逐渐加强，而手机和互联网的发展使得可获得的数据越来越大，这两个原因使得深度学习发挥的空间越来越大。

在图像识别方面，深度学习主要是使用卷积神经网络进行学习。卷积神经网络的主要优势在于它可以通过网络结构提取图像的特征并形成层级结构。通过卷积和池化学习图像的invariant features。相比于传统神经网络（多层感知器），使用卷积核可以大大减少参数量。这使得我们可以训练更深的网络或使用数量更大的数据。卷积神经网络的结构中包括卷积操作（Activation function as ReLU），和Pooling multi-responses (max-pooling, mean-pooling)。训练过程中主要包括：使用Feedforward计算误差，通过Gradient descent and stochastic gradient descent (SGD)最小化误差，Backpropagation来更新参数。通过Dropout和Data augmentation可以减少过拟合，并且通过Mini-batches 和epoch方法来增加计算速度。输出一般使用softmax，softmax is a multinomial logistic regression，also known as Cross Entropy。

本文分析了近些年提出的主要卷积神经网络的特点，并分析了他们的结构。在实验中，使用了MNIST数据集训练VGG, ResNet, and GoogleNet networks，来进行图像分类的任务。通过实验数据比较他们之间的优势与缺陷。

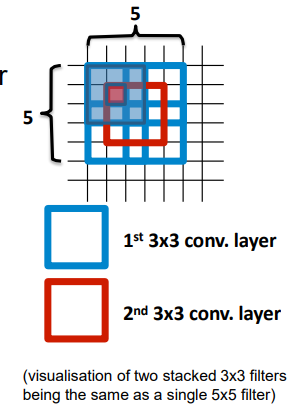
二相关工作

与1998年提出的LeNet5是早期的CNN结构，他是一个包括Max Pooling的一个五层的结构。2012提出的AlexNet是CNN的resurrection，因为它在Large Scale Visual Recognition Challenge (ILSVRC-2012) ImageNet Competition中使得错误率降低了百分之十，这是一个巨大的飞跃。它是一个7 layers network excluding Max Pooling，获得如此巨大提升的原因是1使用dropout降低了过拟合2使用ReLU作为激活函数加快了收敛速度3使用Two GPUs implementation降低学习时间，解决了sigmoid函数收敛速度慢并且远处的函数值与很远处的函数值差不多的问题，而且比tanh函数误差降低快六倍。

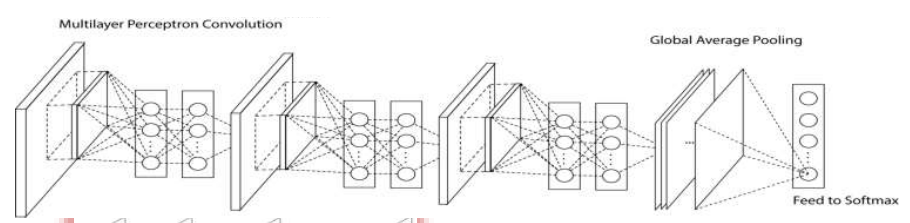


通过AlexNet也发现了深度学习的几个问题：1深度学习往往有巨大的参数量，参数的数量越多，如果没有足够大量的数据的话，就越容易过拟合。2大量的零权重浪费了非常多的计算量3随着层数的加深，在训练中反向传播参数的过程中出现value diminishing，也就是说很难将最优化误差的反馈传播到最前面的层中的参数。

从AlexNet开始，研究人员的主要突破方向放在了1如何减少参数量同时又能保证精度。2如何解决value diminishing问题。处理好这两个问题，就可以加深网络层数进而提升效果。2015提出的VGG使用了Cascaded layers of filters – Narrower & deeper, less parameters with the same effectiveness of a wider & shallower filter。通过小的卷积核的叠加代替大的卷积核，减少了参数量，进而增加深度，提升效果。



2014年提出的Network in Network在卷积中加入了多层感知器，从而得到More Local Abstraction。它使用全局平均池化代替了全连接层，从而大大减少了参数数量。



2015年提出的GoogLeNet采用了Inception modules。在每个Inception module 中使用了大小不同的卷积核，这更能体现出自然图像的性质。在GoogLeNet中，卷积核的数目随着深度不断增加，并且设置了三个输出位置，这有效减缓了value diminishing的影响。

2016提出的ResNet中，将后面层与前面层建立传播关系，从而overcome vanishing gradient。如果一个模块不需要学习，那么就直接吧它的权重设为零，这样即使加深网络的深度准确率也不会下降。

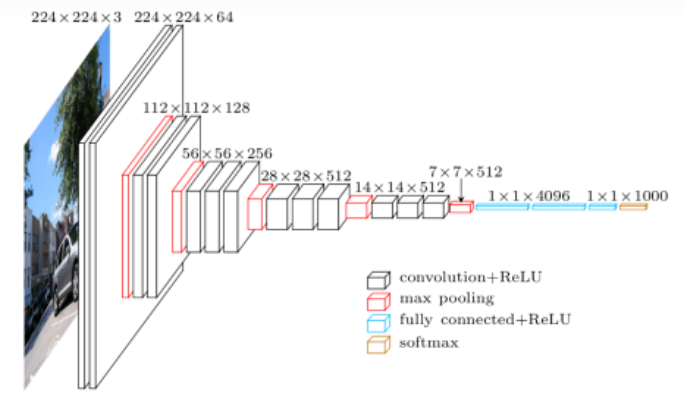
三模型

3.1 Model Architecture

(I) VGG-16

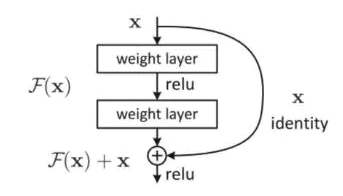
VGG的输入为fixed size image inputs (224x224) and mean subtraction。它使用了Use stacks of small receptive filters (3x3) and (1x1) with 1 pixel convolutional strides，大大减少了参数量。卷积后使用Spatial preserving padding，并且使用5 max-pooling layers carried out at 2x2 windows with stride of 2，Max-pooling only applied to some conv layers。

网络最后，接着 3 个全连接层（Fully-Connected layers）。神经元数目分别为：4096，4096，1000，最后 1000 用于 1000 分类。最后用一个 softmax 层，用于计算类别的概率（约束在 0~1 之间，和为 1）。整个网络使用 ReLU 作为非线性激活函数。由于 Local Response Normalisation (LRN) 对网络无益，所以网络未使用。



(II) ResNet

ResNet 的核心思想是 引入一个 identity shortcut connection的结构,直接跳过一个或多个层

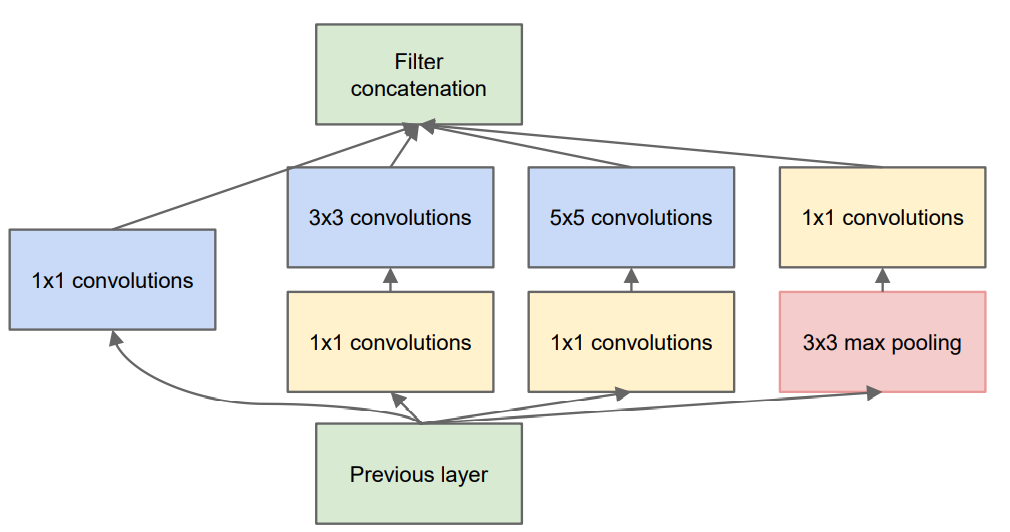


对于一个堆积层结构（几层堆积而成）当输入为 x 时其学习到的特征记为 H(x) ，现在我们希望其可以学习到残差 F(x)=H(x)-x ，这样其实原始的学习特征是 F(x)+x 。之所以这样是因为残差学习相比原始特征直接学习更容易。当残差为0时，此时堆积层仅仅做了恒等映射，至少网络性能不会下降，实际上残差不会为0，这也会使得堆积层在输入特征基础上学习到新的特征，从而拥有更好的性能。

(III) GoogLeNet

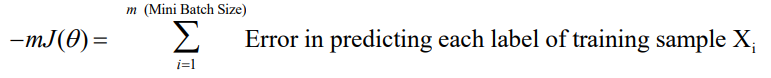
Inception是个网中网，网中总共包含4个子网，该结构将CNN 中常用的卷积（1x1，3x3， 5x5）、池化操作（3x3）堆叠在一起（卷积、池化后的尺寸相同，将通道相加），一方面增加了网络的宽度，另一方面也增加了网络对尺寸的适应性。

在 inception 的基础上，还可以加上降维功能的结构，如下图所示，在原始 inception 结构的基础上，在分支2，3，4上加入了**卷积核大小为1x1的卷积层**，目的是为了降维（减小深度），减少模型训练参数，减少计算量。



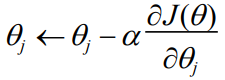
训练

A model is trained by a subset m (mini-batch) of the training set, then random repeat over many mini-batches in one epoch. Repeat over many epochs. Per training sample, Feedforward computes layer-outputs propagated to the softmax. Per mini-batch, minimise softmax loss function J (Cross-Entropy).



Model learning by minimise a loss function from reducing its gradient: Minimise the negative log likelihood, i.e. Maximum Likelihood Estimation (MLE) or Maximum A Posteriori probability (MAP) assuming uniform prior

Gradient descent: Estimate Optimal weights & biases q and a feature representation (e.g. the final FC layer) from many iterations (epochs) on a training dataset by calculating the gradient of error / loss function (a hyper-surface in a parameter space) and reducing it iteratively (gradient descent) controlled by α learning rate



Backpropagation (the Chain Rule) calculates the gradient on the loss function and update all the parameters (weights & biases) of all the layers propagated back to the input.

四实验

4.1 Datasets

The MNIST database [1] of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size- normalized and centred in a fixed-size image.

训练集（training set）由来自250个不同人手写的数字构成，其中50%是高中学生，50%来自人口普查局（the Census Bureau）的工作人员。测试集（test set）也是同样比例的手写数字数据，但保证了测试集和训练集的作者集不相交。

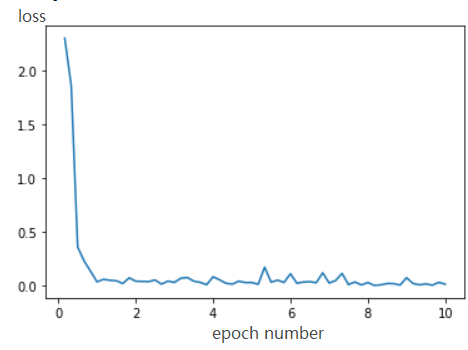
 MNIST数据集一共有7万张图片，其中6万张是训练集，1万张是测试集。每张图片是28 × 28的0-9的手写数字图片组成。每个图片是黑底白字的形式，黑底用0表示，白字用0-1之间的浮点数表示，越接近1，颜色越白。

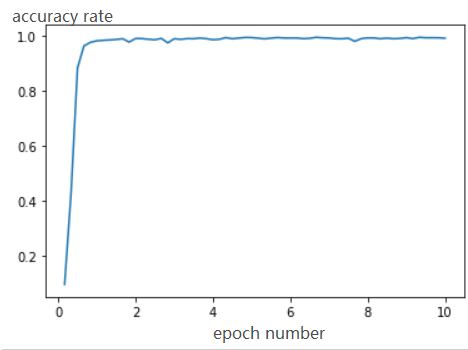
4.2 Testing Results

训练时，batchsize等于200，学习率等于0.001，在每个epoch中，每50个batch打印一下loss 并计算 accuracy rate，每一个epoch记录一次时间

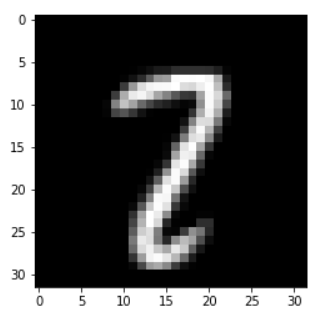
(I) VGG-16

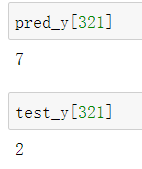
在训练vgg时，mnist数据集的图像大小需要放大，否则会随着池化使得图像越来越小，提前变为一个点，使得训练无法进行。



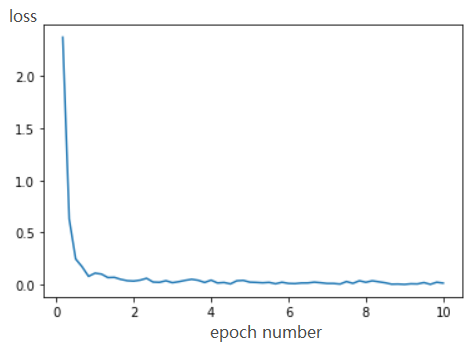


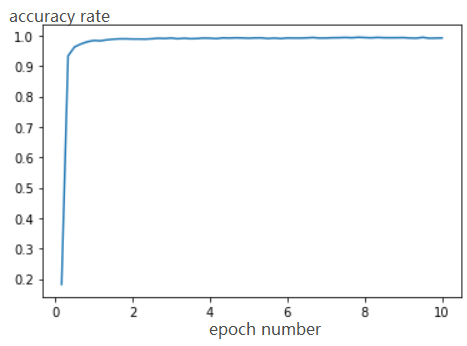
287.233229136467



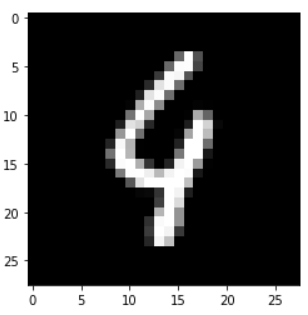


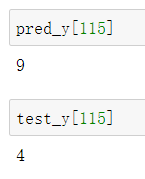
(II) ResNet



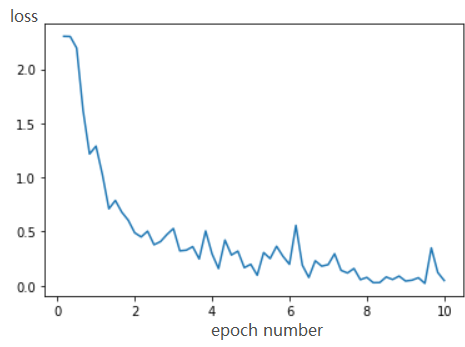


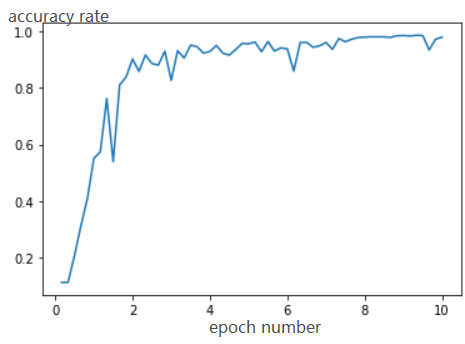
55.179085445404056



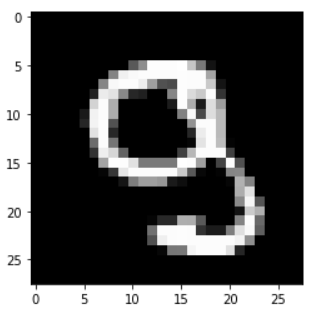


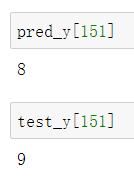
(III) GoogLeNet





127.78432106971741





五结论

通过实验结果，可以知道resnet在手写体识别的任务中表现最好，收敛快，运算速度快，准确率高。Vgg虽然同样收敛快，准确率高，但是运算速度三者最慢。GoogLeNet 运算速度中等，但是准确率不如其他两者，收敛慢，波动大。

六参考文献