2、神经网络简介

人工神经网络ANN（Artificial Neural Network）是一种有监督学习的机器学习模型。神经网络由人工神经单元连接成网络，每个神经单元接收与其相连的各个单元的输入信号，乘以对应的连接权重并累加。这个值经过激活函数（通常是非线性函数，如S型Sigmoid函数、ReLU函数等）之后成为该单元的输出信号。神经元之间的连接权重即是需要拟合的参数。目前的神经网络从小到大，有几千至几百万之众，连接参数可以打到上千万甚至上亿的量级。

通过这种线性变换（权重与输入点积）与非线性变换（激活函数）结合的方式，高维空间中线性不可分的输入特征被变换到了易于分割的空间中，从而完成对输入的分类、识别等信息提取过程。

应用中的人工神经网络会把人工神经单元组织成层，每层之间互相连接，信号从（一般经过预处理之后的）输入层开始，逐层地传递到输出层，得到有用的信息。最简单的神经网络可由三层构成：第一层为输入层，第二层为隐层，第三层为输出层。数学上已经得到证明的是【来源请求】，随着隐层神经元数量的增多，一个三层的神经网络理论上可以任意逼近一个未知的连续函数。但神经元增多同时会带来训练上的难度，故目前大部分高精度的CNN模型选择增多网络层数（即“深度学习”概念中的“深度”）来提高拟合的逼近程度。如ILSVRC的ImageNet图片分类物体探测挑战【来源请求】中，2017年的精度冠军得主MSRA使用了一个多达128层的神经网络。

【神经元长什么样】

为了解决特定的问题，神经网络需要根据可供拟合的已有数据集对其大量参数调整，即“训练”过程。根据实际问题，我们规定一定的损失函数。这些损失函数是参数的函数，表征模型的性能如何，训练的目的就是以训练数据集为根据，最小化损失函数。类比最小二乘法线性拟合的过程，各个样本即训练数据集，线性函数的两个系数即需要训练的参数，损失函数即误差的平方和。

和最小二乘等简单的模型不同，神经网络的损失函数最值通常无法表达为参数和训练集的解析表达式形式。故损失函数的最小化要通过数值方法，在训练集上反复迭代近似到其极小值。常见的方式是梯度下降法调整参数，逐步迭代求得逼近极小值时的参数。当激活函数可导时，易证【来源待查】数值梯度恰好可以写成由输出向输入端反向传播的形式，故训练过程为反向传播。

模型训练效果达标后，参数固定下来，网络即可使用。显然，反向传播训练网络的过程会在模型得到应用之前在提供商的服务器上完成，使用时网络只需根据确定的参数对输入做前向传播即可。故，本文侧重于优化Caffe的前向计算过程。