**基于BP神经网络的数字识别**

**江南大学数字媒体学院**



**姓 名： 宋昕**

**专 业： 软件工程**

**学 号： 6181611008**

摘 要

手写数字识别是利用机器或计算机自动辨认手写体阿拉伯数字的一种技术，是光学字符识别技术的一个分支。由于阿拉伯数字通用，并且数字识别和处理也常是一些自动化系统的核心和关键，所以对手写体数字识别的研究通用性强，且意义重大。本文通过对MINST库进行数据预处理与构建基于BP神经网络的分类器来实现手写数字识别的任务。在此基础上，探索能够优化基于BP神经网络分类器的不同方案。将不同的实验方案对实验结果精度的影响作比较分析。

**关键词：**手写数字；预处理；BP神经网络；分类器优化方案

ABSTRACT

Handwritten digit recognition is a technique that automatically recognizes handwritten Arabic numerals using a machine or computer and is a branch of optical character recognition technology. Because Arabic numerals are common, and digital recognition and processing are often the core and key of some automated systems, the research on handwritten digit recognition is versatile and significant.

This paper realizes the task of handwritten digit recognition by preprocessing data in MINST library and constructing classifier based on BP neural network. On this basis, explore different schemes that can optimize the BP neural network classifier. The effects of different experimental protocols on the accuracy of experimental results were compared and analyzed.

**Keywords**: handwritten digits; preprocessing; BP neural network; classifier optimization scheme

目 录

摘 要 I

ABSTRACT II

目 录 i

第1章 引言 1

1.1 研究手写数字识别的意义 1

1.2 手写数字识别的应用 1

第2章 数据预处理 2

2.1 灰度处理 2

2.2 one-hot编码 2

第3章 神经网络 3

3.1 感知机神经元 3

3.2 神经网络组成 3

3.2.1 结构 3

3.2.2 激励函数 4

3.2.3 学习规则 5

第4章 基于BP神经网络分类 7

4.1 损失函数 7

4.2 梯度下降 7

4.3 BP算法 8

4.4 基本步骤 8

4.5 原理推导 9

4.5.1 前向传播 9

4.5.2 误差传递 9

4.5.3 参数更新 9

4.6 算法优化策略 10

4.6.1 减轻神经元饱和 10

4.6.2 减轻过度拟合现象 11

4.6.3 加速训练过程 12

4.6.4 优化超参数 12

第5章 实验及结果分析 14

5.1 实验过程 14

5.2 实验结果 14

5.3 实验结果分析 15

第6章 结论与展望 17

6.1 总结 17

6.2 展望 17

参考文献 19

致 谢 21

1. 引言

手写数字识别是一个经典的模式识别领域的问题，因其典型性和广泛应用性而具有重大的理论研究价值和应用前景。手写数字识别研究的目标是：利用计算机通过某些算法准确高效地辨认手写体阿拉伯数字。它是手写字符识别的一个热门方向，是光学字符识别技术的一个重要分支。

## 研究手写数字识别的意义

脱机情况下手写数字识别不但具有很高的应用价值，同时也具有很高的理论价值。首先，由于经济的发展，金融市场化进程的日益加快，票据业务发展很快，票据数量也与日俱增。其中个人凭证，支票，发票，进账单等等票据均需要处理大量的信息。而目前，票据录入仍然依赖人工处理方式，因而使得票据管理工作也相对落后。如果能通过手写体字符识别技术来实现信息的自动录入，无疑会大大有利于解决传统人工处理方式中存在的工作量大、成本高、效率低、时效性差等问题。因此，手写数字的识别研究有着重大的现实意义和十分广阔的应用前景，一旦研究成功并投入应用将产生巨大的社会和经济效益。其次，由于数字是全世界通用的符号，识别种类较小，有助于作深入分析及验证一些新的理论。由于手写数字识别的方法很容易推广到其它一些相关问题上，例如英文字母、汉语拼音等文字的识别，所以它逐步成为模式识别的一个热点的研究领域。从50年代开始，许多的研究者就在这一研究领域开展了广泛的探索并为模式识别的发展产生了积极的影响。可以说，手写数字识别的研究将有助于模式识别、机器理解、机器人技术的发展，具有很大的理论价值。

## 手写数字识别的应用

目前手写数字识别主要的应用领域有：

（1）手写数字识别被应用在大规模数据统计中。例如，人口普查、成绩单录入、行业年检、财务报表录入等应用中，需要输入大量的数据，以前完全靠手工输入，耗费大量的人力和物力。若使用手写字符识别技术将大大提高这方面的工作效率。

（2）手写数字识别被应用在财务、税务、金融领域应用。随着我国经济的快速发展，每天会有大量的财务、税务、支票等需要处理。如果使用计算机进行自动化处理，无疑会节省大量的时间和金钱。该领域对识别的精度要求很高，因此对预处理和识别的算法要求就很高。

（3）手写数字识别被应用到邮件分拣系统中。传统的邮件分拣主要是由人工来实现的。随着经济的发展，各种私人和商务信件会越来越多，一些大城市每天处理的邮件高达几百万件。因此邮件的自动分拣成为大势所趋。

1. 数据预处理

当用神经网络作为手写数字的识别分类器时，输入端为彩色或灰度的手写数字图像，需要对其进行的数据预处理是灰度变换。而在输出端，0~9的数字为识别的结果，然而在实际的使用中，通常用one-hot编码对分类结果进行变换。在用神经网络训练模型时，采用的数据也要进行数据集的划分。

## 灰度处理

RGB色彩模式是工业界的一种颜色标准，是通过对红(R)、绿(G)、蓝(B)三个颜色通道的变化以及它们相互之间的叠加来得到各式各样的颜色的，RGB即是代表红、绿、蓝三个通道的颜色，这个标准几乎包括了人类视力所能感知的所有颜色，是目前运用最广的颜色系统之一。灰度图是指只含亮度信息，不含色彩信息的图象，就象黑白照片：亮度由暗到明，变化是连续的，这种亮度称为灰度，灰度分为256阶。RGB图像有3层即3个通道，而灰度图只有1层。将3通道的RGB图像转为单通道的灰度图像即为灰度转换。RGB转灰度的公式是Gray = R\*0.299 + G\*0.587 + B\*0.114。

## one-hot编码

One-Hot编码，又称为一位有效编码，主要是采用N位状态寄存器来对N个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候只有一位有效。One-Hot编码是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后，每个整数值被表示为二进制向量，除了整数的索引之外，它都是零值，它被标记为1。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数字 | one-hot编码 | | | | | | | | | |
| 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |

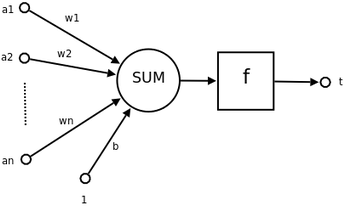
1. 神经网络

神经网络在机器学习和认知科学领域，是一种模仿生物神经网络的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统，通俗的讲就是具备学习功能。现代神经网络是一种非线性统计性数据建模工具。神经网络已经被用于解决各种各样的问题，例如机器视觉和语音识别。这些问题都是很难被传统基于规则的编程所解决的。

神经网络可以看成是一种由神经元“感知机”组成的网络结构。神经网络包含结构，激励函数，学习规则这三个部分。神经网络分成单层神经元网络和多层神经元网络。

## 神感知机神经元

感知器使用特征向量来表示的前馈神经网络，它是一种二元分类器，把矩阵上的输入x（实数值向量）映射到输出值f(x)上（一个二元的值）。w是实数的表示权重的向量，wx是点积。b是偏置，一个不依赖于任何输入值的常数。偏置可以认为是激励函数的偏移量，或者给神经元一个基础活跃等级。



f(x)（0或1）用于对x进行分类，看它是肯定的还是否定的，这属于二元分类问题。如果b是负的，那么加权后的输入必须产生一个肯定的值并且大-b，这样才能令分类神经元大于阈值0。从空间上看，偏置改变了决策边界的位置（虽然不是定向的）。

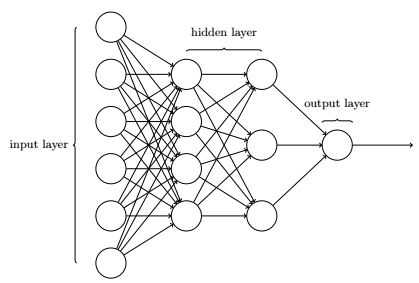


由于输入直接经过权重关系转换为输出，所以感知器可以被视为最简单形式的前馈式人工神经网络。

## 神经网络组成

### 结构

结构指定了网络中的变量和它们的拓扑关系。例如，神经网络中的变量可以是神经元连接的权重ｗ和神经元的激励值a。从结构上，神经网络还可分成单层神经元网络和多层神经元网络。



(1)单层神经元网络

最基本的神经元网络形式，由有限个神经元构成，所有神经元的输入向量都是同一个向量。由于每一个神经元都会产生一个标量结果，所以单层神经元的输出是一个向量，向量的维数等于神经元的数目。

(2)多层神经元网络

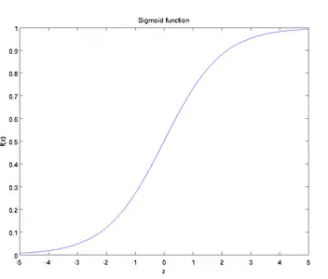
包含多个输入向量与输出向量，中间包含着一个或多个隐藏层。

### 激励函数

大部分神经网络模型具有一个短时间尺度的动力学规则，来定义神经元如何根据其他神经元的活动来改变自己的激励值。一般激励函数依赖于网络中的权重（即该网络的参数）。

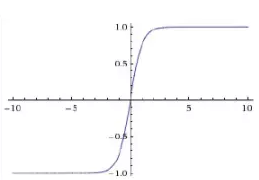
在计算网络中， 一个节点的激励函数定义了该节点在给定的输入或输入的集合下的输出。标准的计算机芯片电路可以看作是根据输入得到开（1）或关（0）输出的数字电路激活函数。这与神经网络中的线性感知机的行为类似。然而，只有非线性激活函数才允许这种网络仅使用少量节点来计算非平凡问题。 在人工神经网络中，这个功能也被称为传递函数。常见的激励函数分为以下几类：

(1)Sigmoid函数



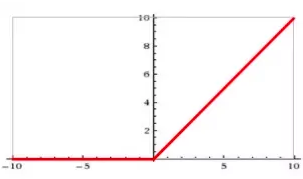
sigmoid函数也称为Logistic 函数，用于隐层神经元输出取值范围为(0,1)。它可以将一个实数映射到(0,1)的区间，可以用来做二分类。在特征相差比较复杂或是相差不是特别大时效果比较好。sigmoid缺点的是激活函数计算量大，反向传播求误差梯度时，求导涉及除法反向传播时，很容易就会出现梯度消失的情况，从而无法完成深层网络的训练。

(2)Tanh函数



Tanh函数也称为双曲正切函数，取值范围为[-1,1]。tanh在特征相差明显时的效果会很好，在循环过程中会不断扩大特征效果。与 sigmoid 的区别是，tanh 是 0 均值的，因此实际应用中 tanh 会比 sigmoid 更好。

(3)ReLu函数



ReLu函数也称为整流线性函数。ReLU 的优点是用 ReLU 得到的 SGD 的收敛速度会比 sigmoid/tanh 快很多。ReLU 的缺点是训练的时候很脆弱。

(4)Softmax函数

Softmax函数又称为归一化指数函数。它能将一个含任意实数的K维向量Z “压缩”到另一个K维实向量σ(Z)中，使得每一个元素的范围都在(0,1)之间，并且所有元素的和为1。softmax函数用于多分类神经网络输出。

### 学习规则

学习规则指定了网络中的权重如何随着时间推进而调整。这一般被看做是一种长时间尺度的动力学规则。一般情况下，学习规则依赖于神经元的激励值。它也可能依赖于监督者提供的目标值和当前权重的值。例如，用于手写识别的一个神经网络，有一组输入神经元。输入神经元会被输入图像的数据所激发。在激励值被加权并通过一个函数（由网络的设计者确定）后，这些神经元的激励值被传递到其他神经元。这个过程不断重复，直到输出神经元被激发。最后，输出神经元的激励值决定了识别出来的是哪个字母。

具体的学习方法则因网络结构和模型不同而不同，常用反向传播算法即BP算法来验证。

1. 基于BP神经网络分类

手写数字识别用的是基于BP神经网络的分类器。神经网络的分类原理是希望网络中的参数找到合适的值以至于网络的输出值能够拟合所有的输入值。一个神经网络中的参数就包括了权值w和偏置b，以及给定的一些超参数如学习效率η。为了能够描述输出拟合输入的过程，引入了损失函数进行描述。其中用到的拟合规则，即参数更新的规则是在损失函数上运用梯度下降。本章将给出基本的概念，BP算法以及基于BP网络的分类器的优化方案。

## 损失函数

损失函数是用来估量模型的预测值与真实值的不一致程度，是一个非负实值函数，通常使用*C(y,，f(x))*来表示。*y*为实际值，*f(x)*为预测值。损失函数越小，模型的鲁棒性就越好。损失函数是经验风险函数的核心部分，也是结构风险函数重要组成部分。常见的损失函数分为以下几类：

(1)平方损失函数(最小二乘法)

最小二乘法是线性回归的一种，最小二乘法将问题转化成了一个凸优化问题。最小二乘的基本原则是：最优拟合直线应该是使各点到回归直线的距离和最小的直线，即平方和最小，即最小二乘法是基于距离的，而这个距离就是欧几里得距离。

(2) log对数损失函数（逻辑回归）

在逻辑回归的推导中，它假设样本服从伯努利分布（0-1分布），然后求得满足该分布的似然函数，接着取对数求极值等。而逻辑回归并没有求似然函数的极值，而是把极大化当做是一种思想。

(3)交叉熵代价函数

在逻辑回归的推导中，它假设样本服从伯努利分布（0-1分布），然后求得满足该分布的似然函数，接着取对数求极值等。而逻辑回归并没有求似然函数的极值，而是把极大化当做是一种思想。

## 梯度下降

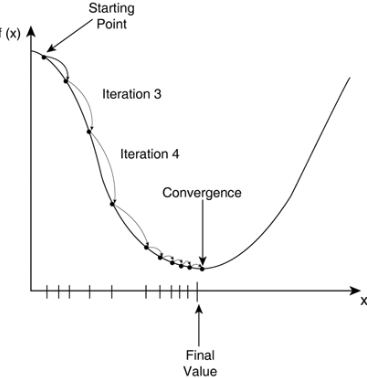
梯度下降法是一个一阶最优化算法，通常也称为最速下降法。 要使用梯度下降法找到一个函数的局部极小值，必须向函数上当前点对应梯度（或者是近似梯度）的反方向的规定步长距离点进行迭代搜索。

损失函数定义了一种预测值与真实值的差别，因此目标便是要使得损失函数取值最小，才能保证预测值越来越接近真实值，越来越能够拟合输出。使用梯度下降来求解损失函数最小便是一个策略。要注意的是，损失函数要确保是一个凸函数，这要才能保证在使用梯度下降的过程中，局部最小便是全局最小。梯度下降的更新法则是：

***w = w –ηw***

***b = b –ηb***

当权值*w*或偏置值*b*大于Final Value时，*Δw*或*Δb*大于0，即*w*或*b*的更新是减去正值，是向左趋向Final Value的；当权值*w*或偏置值*b*小于Final Value时，*Δw*或*Δb*小于0，即*w*或*b*的更新是减去负值，是向右趋向Final Value的。因此从理论上，用” – ”表示就一定能到达损失函数的最低点。



这样梯度就可以表示权值*w*和偏置*b*更新的步长。在参数更新的过程中，便可以用梯度下降来更新权值与偏置。

## BP算法

基本BP算法由正向传播和误差的反向传播两个过程组成。

正向传播时，输入样本从输入层进入网络，经隐层逐层传递至输出层，如果输出层的实际输出与期望输出(导师信号)不同，则转至误差反向传播；如果输出层的实际输出与期望输出(导师信号)相同，结束学习算法。

反向传播时，将输出误差(期望输出与实际输出之差)按原通路反传计算，通过隐层反向，直至输入层，在反传过程中将误差分摊给各层的各个单元，获得各层各单元的误差信号，并将其作为修正各单元权值的根据。这一计算过程使用梯度下降法完成，在不停地调整各层神经元的权值和阈值后，使误差信号减小到最低限度。

权值和阈值不断调整的过程，就是网络的学习与训练过程，经过信号正向传播与误差反向传播，权值和阈值的调整反复进行，一直进行到预先设定的学习训练次数，或输出误差减小到允许的程度。

## 基本步骤

BP神经网络算法主要步骤如下：

(1)前向传播。组成输入模式由输入层经过隐含层向输出层的“模式顺传播”过程。

(2)误差传递。网络的期望输出与实际输出之差的误差信号由输出层经过隐含层逐层休整连接权的“误差逆传播”过程。

(3)权值更新。由“模式顺传播”与“误差逆传播”的反复进行的网络“记忆训练”过程。

(4)停止训练。网络趋向收敛即网络的总体误差趋向极小值的“学习收敛”过程。

## 原理推导

### 前向传播

BP神经网络是一种这一层的输入均来源于上一层的输出的网络结构，即前馈型神经网络。从输入层开始，基于带权值和偏置的网络逐层更新直到最后的输出层。

…

其中*Z2*代表第二层神经元的带权输入，*w2*代表第二层神经元与第一层神经元之间连接的权值，*b2*代表第二层神经元与第一层神经元之间连接的偏置值，*a2*代表第二层的神经元经过激励函数后的输出值，*f(·)*代表激励函数。对于第二层神经元来说，输入就是输入层X。对于最后一层神经元L层即输出层来说，输入就是上一层L-1层的输出值*aL-1*。

### 误差传递

定义误差为，L层即为最后一层输出层，输出层误差方程为：

将公式用Hadamard乘积符号⊙向量化表示为：

误差从最后一层前向更新。在神经网络中每一层神经元都有带权输入，每一层带权输入都和当层的权值*wL*与偏置*bL*有关。要从损失函数入手更新每一层的权值*wL*与偏置*bL*，就要更新每一层的带权输入，即更新每一层的误差即可。

使用下一层误差来表示当前层的误差的公式为：

### 参数更新

损失函数*C*对每一层神经元的权值*wL*与偏置*bL*的梯度为：

权值*wL*与偏置*bL*的误差公式为：

权值*wL*与偏置*bL*的更新公式为：

## 算法优化策略

应用当前的BP神经网络架构，选择不同的激励函数与损失函数组合，已经能够完成较为出色的识别任务了，然而对BP神经网络的性能提升还有许多优化策略与技巧。

### 减轻神经元饱和

当神经网络的学习速率开始变得缓慢时，就要考虑是否发生了神经元饱和的现象。以下是减轻神经元饱和现象的常用方法。

(1)权值的初始化

一般情况下，权值*wL*的初始化方法通常是归一化的高斯分布，即均值为0，标准差为1的分布值。而带权输入*Z=wX+b*的值在有可能>>1或<< -1，在Z经过激励函数*f()*后，很有可能激活值*a=f(Z)*十分接近0或1，也就表示隐藏神经元会饱和。所以当出现这样的情况时，在权重中进行微小的调整仅仅会给隐藏神经元的激活值带来极其微弱的改变。而这种微弱的改变也会影响网络中剩下的神经元，然后会带来相应的代价函数的改变。结果就是，这些权重在进行梯度下降算法时会学习得非常缓慢。为了避免这种类型的饱和导致最终学习速度的下降，一个方法是使用均值为0，标准差为1/的具有更尖锐峰值的高斯分布，以使神经元更不可能饱和。

(2)重新选择损失函数与激励函数的选择

当损失函数选为平方损失函数，激励函数选为sigmoid函数时，将会面临神经元的饱和的情况。在这种选择下，权值*w*与偏置*b*的梯度公式为：

可以知道，的值在函数接近0或1的时候很小，这样小的数会导致权值*w*与偏置*b*的更新步长很小，表现为学习速率下架。导致此结果的原因便是神经元饱和。为了避免神经元饱和现象的出现，重新选择损失函数与激励函数的组合使得权值*w*与偏置*b*的更新步长中不出现值可能非常小的导数即可。当损失函数依然为平方损失函数，而激励函数*f()*是关于*ZL*的线性函数时，权值*w*与偏置*b*的梯度公式为：

当损失函数选为交叉熵代价函数，而激励函数仍然是*sigmoid*函数时，权值*w*与偏置*b*的梯度公式为：

在这种组合下，便消除了导数值过小导致的神经元饱和现象。同样的组合还有损失函数选为log似然函数，激励函数选为softmax函数时，权值*w*与偏置*b*的梯度公式为：

因此，当选用最小二乘法+线性激励函数，交叉熵代价函数+sigmoid函数或log似然函数+softmax函数的组合，均可以避免导数值过小导致的神经元饱和现象。又由于log似然函数+softmax函数的组合表达的是一种输出值是概率的思想，因此更适合这种类别不重叠的手写数字分类问题。

### 减轻过度拟合现象

在机器学习的模型中，如果模型的参数太多，而训练样本又太少，训练出来的模型很容易产生过拟合的现象。在训练神经网络的时候经常会遇到过拟合的问题，过拟合具体表现在：模型在训练数据上损失函数较小，预测准确率较高；但是在测试数据上损失函数比较大，预测准确率较低。

当在训练集上表现良好而在测试集上表现得失败时，神经网络可能发生了过度拟合现象。解决过度拟合的方法是用验证集判断训练集是否有必要继续学习，从而砍掉不重要的学习过程，以增强训练模型的泛化能力。以下是解决神经网路过拟合的常用方法：

(1)规范化

正则化的神经网络常常能够比非正则化的泛化能力更强，这是一实验事实。规范化的想方法是增加一个额外的项到代价函数上，此项叫做规范化项。

规范化项是所有权重的平方的和，称为规范化参数。加入规范项后的权值w与偏置b的偏导数变成：

而权值w与偏置b的更新规则变成了：

即w将有更小的权重，所以规范化也称为权值衰减。

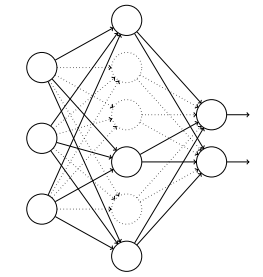
(2)弃权dropout

dropout是指在深度学习网络的训练过程中，对于神经网络单元，按照一定的概率将其暂时从网络中随机丢弃。使用dropout的流程是：

首先随机（临时）删掉网络中一半的隐藏神经元，输入输出神经元保持不变（图中虚线为部分临时被删除的神经元）。

然后把输入x通过修改后的网络前向传播，然后把得到的损失结果通过修改的网络反向传播。一小批训练样本执行完这个过程后，在没有被删除的神经元上按照随机梯度下降法更新对应的参数w和b。

然后重复这一过程：恢复被删掉的神经元（此时被删除的神经元保持原样，而没有被删除的神经元已经有所更新）。从隐藏层神经元中随机选择一个一半大小的子集临时删除掉（备份被删除神经元的参数）。对一小批训练样本，先前向传播然后反向传播损失并根据随机梯度下降法更新参数w和b（没有被删除的那一部分参数得到更新，删除的神经元参数保持被删除前的结果）。

**

### 加速训练过程

影响神经网络的问题的问题主要有两个，一个是过拟合，一个是训练速度太慢。加速训练过程可以采用mini-Batch训练的方法，即小批量训练。在不采用批训练的策略下，每一个训练样本在经过前向传播与反向传播后，都要更新一次网络中的参数，对于MINST训练集来说，50,000个训练样本意味着50,000次网络参数的更新，这样会导致网络的训练速度变缓。因此，采用批量训练的策略是很有必要的。假设mini-Batch的个数为m，则意味着每经过m个样本才更新一次网络参数，这样训练的速度就是之前不采用批量训练的m倍了。采用mini-Batch后，参数的更新策略为：

即对m个参数的梯度求和后再求均值即得到新的参数更新规则。

### 优化超参数

神经网路中的超参数主要包括：学习率η，正则化参数，神经网络的层数L，每一个隐层中神经元的个数N，学习的迭代次数epoch，小批量数据batch size，输出神经元的编码方式，代价函数的选择，权重初始化的方法，神经元激活函数的种类，参加训练模型数据的规模这十一类超参数。

这些都是可以影响神经网络学习速度和最后分类结果，其中神经网络的学习速度主要根据训练集上代价函数下降的快慢有关，而最后的分类的结果主要跟在验证集上的分类正确率有关。因此可以根据该参数主要影响代价函数还是影响分类正确率进行分类。以下便是常用的超参数调整策略：

(1)调整学习速率

目前深度学习使用的都是非常简单的一阶收敛算法，梯度下降法，不管有多少自适应的优化算法，本质上都是对梯度下降法的各种变形，所以初始学习速率对深层网络的收敛起着决定性的作用。如果学习速率太小，会导致网络误差函数值cost下降非常慢；如果学习速率太大，那么参数更新的幅度就非常大，就会导致网络收敛到局部最优点，或者loss直接开始增加。

学习速率的选择策略在网络的训练过程中应该是不断在变化的。在刚开始的时候应该选择相对较大的学习速率，这样cost下降得更快；当训练一段时间之后，参数的更新就应该有更小的幅度以避免越过全局最优点，所以学习速率一般会做衰减。衰减的方式包括到一定的步数时将学习速率乘上0.1或是呈指数衰减。用验证集绘制cost图像，当cost的收敛变慢或者还增加的时候，便可以让学习速率开始下降。

(2)选择batch size

如果batch size过小，训练数据就会非常难收敛。增大batch size，相对处理速度加快，但是所需内存容量增加，而且不能够再频繁更新权重。batch size 的正确选择是为了在内存效率和内存容量之间寻找最佳平衡。选择最好的小批量数据大小是一种折衷思想。选择合适的batch size的方法是在验证集上画出准确率随时间变化的曲线图，选择得到最快性能的提升的小批量数据大小即可。

(3)选择迭代次数

提前停止表示在每次迭代后，都要计算验证集上的分类准确率。当准确率不再提升时，就终止训练，这也就确定了迭代次数。提前停止也能够帮助避免过度拟合。一种不错的解决方案是如果分类准确率在一段时间内不再提升的时候终止。

1. 实验及结果分析

本文利用MINST手写数字库进行手写数字识别实验 。MNIST 数据集来自美国国家标准与技术研究所，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% 来自人口普查局的工作人员。测试集(test set) 也是同样比例的手写数字数据。手写数字图片为28×28的灰度图像，其中训练集包含60,000张手写数字图片与60,000个标签值，测试集包含10,000张手写数字与10,000个标签值。

而在真实训练的时候，通常会将训练集再次划分为真正的训练集(包含50,000张手写数字图片与50,000个标签值)与验证集(包含10,000张手写数字图片与10,000个标签值)。验证集的作用是进行超参数的选择。

## 实验过程

(1)选取训练集作为训练样本，测试集作为测试样本。

(2)设置网络训练要用到的各种参数：学习速率η，迭代周期，mini-batch的大小。权值初始化。

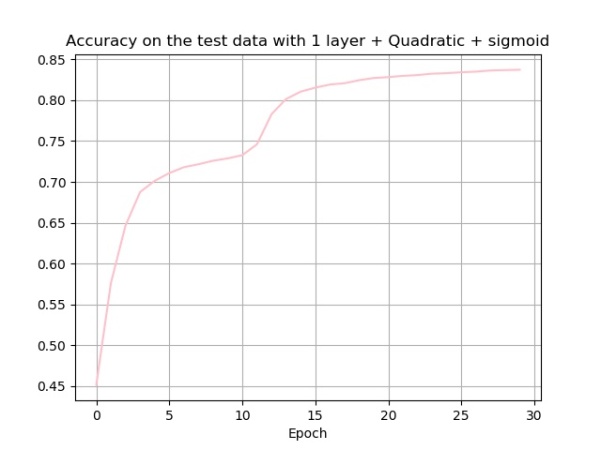
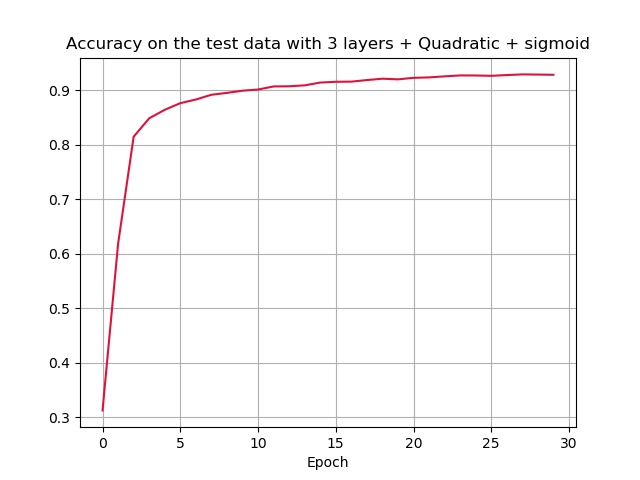
(3)构建BP神经网络。

(4)优化BP网络，将优化后的网络与初始网络的识别效果做比较。权值初始化的方法*(Z~N(0,))*，规范化。

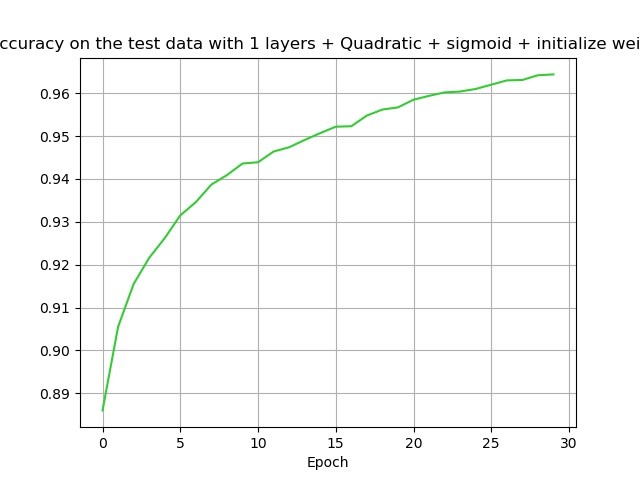
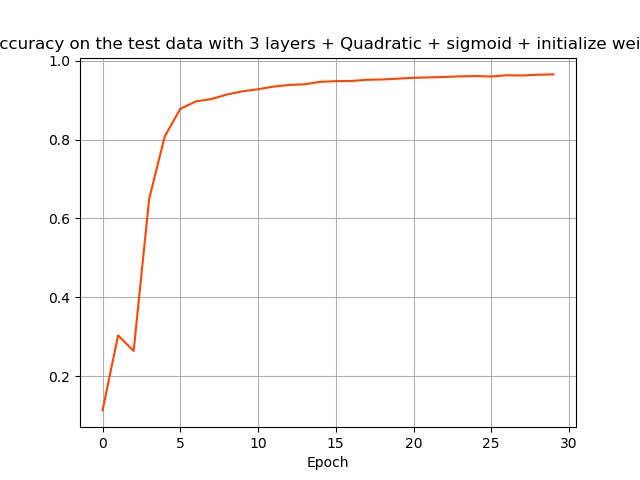
(5)设置不同的超参数，探究网路对手写数字识别效果的影响。探究的超参数有：损失函数，激活函数，网络层数(隐含层层数layer)。

## 实验结果

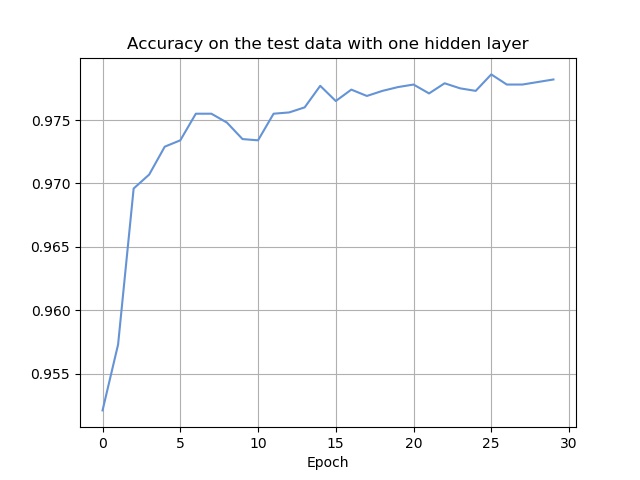
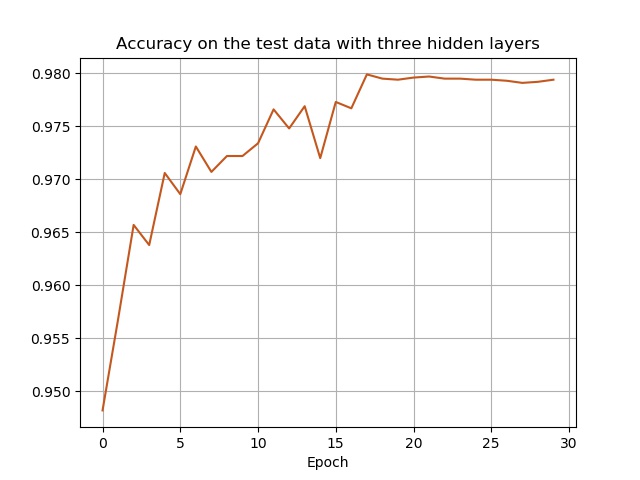
(1)当激活函数与损失函数是sigmoid+均方误差时，探究隐含层层数对正确率的影响：

(2)当激活函数与损失函数是sigmoid+均方误差，且加入权值初始化后，探究隐含层层数对正确率的影响：

(3)当激活函数在非最后一层是tanh，输出层为softmax，损失函数为log似然函数，且加入权值初始化方法，探究隐含层层数对正确率的影响：

## 实验结果分析

从实验结果来看，达到最高正确率的方法是使用tanh+softmax+log似然函数，且使用权值初始化方法，再配合3层隐含层，正确率可以达到98%以上。但在这种组合下，正确率的曲线在最开始十分陡峭，且几乎在17代的时候就不学习了，所以在这种情况下，最初的学习速率可能过大而导致曲线在最初来回震荡，而在17代后产生了过拟合的情况而导致不学习，所以这个模型其实是不好的。

从正确率图像上看，可以发现用1层隐含层配合的sigmoid+均方误差+权值初始化的方法虽然在正确率上只有96.44%，但它的曲线是平稳光滑且逐步提升的，因此有理由相信，继续优化这一组参数后，它的正确率应该还会有继续上升的空间。

从总体来说，使用了权值初始化方法后，在其他参数不变的情况下，测试集上的正确率都有了很大的提升，所以可以得出权值初始化是提升网络性能的大杀器的结论。

在隐含层增加，而其他参数不变的情况下，测试集上的正确率并没有出现特别大的进步，反而可能更不平缓，表现得性能更糟糕。因此，并不是网络结构越深，识别的效果就越好。

1. 结论与展望

## 总结

随着信息化时代的逐步发展，效率的提升对于人们的日常生活来说十分重要。手写数字的自动化识别对于提升人们的日常生活的效率有巨大帮助，不仅是邮政系统的自动识别数字，还是银行系统自动处理票务业务。本文针对手写数字识别展开研究，通过应用基于BP神经网络的分类器，与一些优化分类器的设计，成功将识别率达到了98%以上。本文通过基于BP神经网络的分类器实现了手写数字的识别，并通过优化BP网络逐步提高了识别的正确率。论文的总结工作围绕着以下三方面进行：

（1）神经网络：神经网络的主要结构是由输入层，隐含层，输出层，以及连接各层的网络权重组成。

（2）BP神经网络：BP神经网络包括前向传播与反向传播两个过程。反向传播应用梯度下降算法，反向传播的主要步骤是：计算输出层误差，逐步反向传播误差，计算网络权重梯度，更新网络权重。

（3）优化方案：对神经网络的优化主要是针对超参数的优化与网络结构的优化，目的是加速训练，提高正确率，避免网络神经元饱和，减轻网络的过拟合现象。超参数的选择对于优化网络很有影响，主要包含学习率，正则化参数，神经网络的层数，每一个隐层中神经元的个数，学习的迭代次数，小批量数据，代价函数的选择，权重初始化的方法，神经元激活函数的种类，参加训练模型数据的规模等。优化网络的结构主要使用的方法是弃权Dropout。Dropout的原理是每一次迭代时随机暂时失效一半的神经元，降低神经元的耦合性，以提高神经网络的泛化能力。

## 展望

随着该领域研究的不断深入，各种各样的问题不断涌现：(1)梯度消失/梯度爆炸。神经网络随着深度的加深训练过程越来越难，面临着梯度爆炸或梯度消失与参数繁重的问题。(2)复杂网络中的不稳定梯度。

未来基于神经网络的类脑智能的研究还有许多亟待解决的问题与挑战：（1）认知神经网络 尽管深度神经网络在语音识别和图像/视频识 别等任务中显示出很大的成功，现有的人工神经网络结构还远远不及生物神经网络结构复杂，仍然是对生物神经系统信息处理的初级模拟，这是制约神经网络智能化水平的一个重要瓶颈。目前深层神经网络仅能完成一些简单的语音与视觉理解任务，在理论上还存在很多局限，训练网络的学习算法目前也十分有限。另一方面，神经认知计算科学对视觉注意力、推理、抉择、学习等认知功能的研究方兴未艾。如何从脑科学和神经认知科学寻找借鉴，从理论上发展出功能更加强大的类脑计算模型如认知神经网络，来解决人工智能面临的局限，将有可望实现更 高层次的类脑智能。（2）主动神经网络生物个体在于环境接触过程中，智能水平会得到提高。人脑可以在没有监督信息时主动地从周围环境中学习，实现对客观世界中物体的区分。因此，如果要实现更加高级的智能行为，现有神经网络的发展需要突破利用神经元与网络结构的结构模拟 思路，从结构模拟向功能模拟乃至行为模拟转换， 借鉴人与环境之间的交互过程，主动且自动的完成 增强学习，以摆脱对监督信息的依赖，在更严苛的环境下完成学习任务，这也是实现高级类脑智能的 可能途径。（3）感知-理解-决策神经网络类脑智能行为可以大概归结为“感知”、“理解”与“决策”三个方面。目前的神经网络模型的功能大都局限在对数据的理解层面，而事实上一个高级的智能机器应该具有环境感知与推理决策的功能。如何发展具有环境感知、数据理解以及推理决策能力的网络模型，也是实现高级类脑智能的必然要求。（4）复杂神经网络实现机器计算能力的提高曾经将神经网络重新拉 回大众关注的视野。对于许多互联网公司来说，如何实现对海量大数据的快速高效训练是深层神经 网络走向实用化的重要标志。现有的 Hadoop 平台不适合迭代运算、SGD 又不能依并行方式工作、而GPU在训练DNN 时仍然显得比较吃力。同时，平台的能耗问题也成为制约其进一步发展的主要因素。为迎接未来深度学习在产业化过程中的挑战，高性能并行加速计算平台的开发成为当务之急。另一方面，生物神经元之间的连接带有随机和动态性，而不是如人工神经网络那样确定和一成不变，如何用计算机硬件或者算法来模拟这一过程虽极具挑战但意义重大。（5）深度神经网络深层神经网络一个最主要的特点在于其具有大量可调的自由参数，这使得其构建起的模型具有较高的灵活性。但另一方面却缺乏有力的理论指导和支撑，大多数情况下仍过分依赖于经验，带有一定程度的随机性。如此复杂的模型很容易在特定数据集上得到近乎理想的拟合效果，然而在推广泛化性能上却往往很难得到保障。为防止过拟合带来的问题，今后应当在数据的规模、网络的结构以及模型的正则化等方面开展工作使得深度神经网络更好地发挥其功能。

参考文献

1. Gangaputra, Sachin. "Handwritten digit database". Retrieved 17 August 2013.Candes E J. Curvelets, multiresolution representation, and scaling laws[J]. 2007, 4119:1-12.
2. Kussul, Ernst; Tatiana Baidyk (2004). "Improved method of handwritten digit recognition tested on MNIST database". Image and Vision Computing. 22 (12): 971–981. doi:10.1016/j.imavis.2004.03.008.
3. David W. A. Bourne. Steepest Descent Method.
4. 林嘉宇, 刘荧. RBF神经网络的梯度下降训练方法中的学习步长优化[J]. 信号处理, 2002, 18(1):43-48.
5. 丛爽, 赵何. 反向传播网络的不足与改进[J]. 自动化博览, 1999(1):25-26.
6. 赵立强, 张晓华, 高振波,等. 基于BP神经网络的主分量分析人脸识别算法[J]. 计算机工程与应用, 2007, 43(36):226-229.
7. 甘俊英, 张有为. 基于BP神经网络的人脸识别[J]. 系统工程与电子技术, 2003, 25(1):113-115.
8. Paul J. Werbos (1994). The Roots of Backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting. New York, NY: John Wiley & Sons, Inc.
9. Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. Learning representations by back-propagating errors. Nature. 8 October 1986, 323 (6088): 533–536.
10. Lalis, Jeremias; Gerardo, Bobby; Byun, Yung-Cheol. An Adaptive Stopping Criterion for Backpropagation Learning in Feedforward Neural Network.International Journal of Multimedia and Ubiquitous Engineering. 2014, 9 (8): 149–156 [17 March 2015].