**武汉大学计算机学院**

**本科生课程设计报告**

**基于numpy的DNN神经网络新冠肺炎高效CT图像识别分类模型**

专 业 名 称 ：计算机学院 软件工程

课 程 名 称 ：商务智能

指 导 教 师 ：朱卫平

学 生 学 号 ：2019302080117

学 生 姓 名 ：钟孝云

二○二一年十二月

# 郑 重 声 明

本人呈交的设计报告，是在指导老师的指导下，独立进行实验工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计报告不包含他人享有著作权的内容。对本设计报告做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计报告的知识产权归属于培养单位。

本人签名： 钟孝云 日期： 2021.12.28

摘 要

自2019年新冠肺炎在中国武汉爆发以来，截止至2021年12月，全球超过200个国家及地区累计确诊新冠肺炎人数已超2亿8千万人，迫切需要制定有效措施控制这一流行病。本课程设计提出了一种基于numpy库实现的DNN模型，采用线性全连接层进行CT图片数据的拟合并以此进行图像的分类预测，使用简单的模型参数获得不错的效果，在模型大小不超过5MB的同时，训练集和测试集的accuracy分别超过了99%和97.8%,同时不调用API实现模型的拟合过程，达到模型设置可控化、模型训练快速化、模型大小轻便化、模型使用简单化的效果。使用商务智能的预测性分析，通过CT图像预测患者感染新冠肺炎的概率。

**关键词：**DNN；深度学习；新冠肺炎；CT图像；商务智能；预测性分析

目录

[郑 重 声 明 2](#_Toc91728439)

[1. 请说明分类算法衡量时准确率(accuracy)的局限性并举例说明。 6](#_Toc91728440)

[1.1 准确率(accuracy)的局限性： 6](#_Toc91728441)

[1.2 准确率(accuracy)的局限性举例说明： 6](#_Toc91728442)

[2. 请说明卷积神经网络中参数共享和稀疏连接的含义。 6](#_Toc91728443)

[2.1 参数共享 6](#_Toc91728444)

[2.2 稀疏连接 7](#_Toc91728445)

[3. 请针对下述数据画出ROC曲线图。 7](#_Toc91728446)

[3.1 第一步：按照属于‘正样本’的概率将所有样本排序 7](#_Toc91728447)

[3.2 对每个样本计算坐标 8](#_Toc91728448)

[3.3 坐标结果与ROC曲线 11](#_Toc91728449)

[4. 选题背景 12](#_Toc91728450)

[5. 需求分析 12](#_Toc91728451)

[5.1 编写目的 12](#_Toc91728452)

[5.2 目标 13](#_Toc91728453)

[5.3 用户的特点 13](#_Toc91728454)

[5.4 对性能的规定 14](#_Toc91728455)

[5.4.1 精度 14](#_Toc91728456)

[5.4.2 时间特性要求 14](#_Toc91728457)

[5.4.3 灵活性 14](#_Toc91728458)

[5.4.4 输入输出要求 14](#_Toc91728459)

[5.5 运行环境规定 14](#_Toc91728460)

[6. 解决方案概述 15](#_Toc91728461)

[6.1 方案概述 15](#_Toc91728462)

[6.2 优点&创新点1-选题新颖贴合时事 15](#_Toc91728463)

[6.3 优点&创新点2-模型精确度高 16](#_Toc91728464)

[6.4 优点&创新点3-未借助深度学习框架，numpy底层实现 16](#_Toc91728465)

[6.5 优点&创新点4-网络结构简单轻量化 16](#_Toc91728466)

[7. 数据集介绍 16](#_Toc91728467)

[7.1 数据集来源 16](#_Toc91728468)

[7.2 数据集介绍 16](#_Toc91728469)

[7.3 数据集预处理 17](#_Toc91728470)

[7.4 训练集与测试集分割 17](#_Toc91728471)

[8. 模型设置与关键代码实现 18](#_Toc91728472)

[8.1 模型配置说明 18](#_Toc91728473)

[8.2 模型使用说明 18](#_Toc91728474)

[8.3 参数初始化、前向传播与激活函数、反向传播 18](#_Toc91728475)

[8.4梯度下降与结果预测 20](#_Toc91728476)

[8.5 数据预处理、加载数据、开始训练 21](#_Toc91728477)

[9. 模型训练与测试结果 23](#_Toc91728478)

[9.1 多隐藏层训练结果 23](#_Toc91728479)

[9.2 单隐藏层训练结果 27](#_Toc91728480)

[10. 总结与讨论 27](#_Toc91728481)

[10.1 实验总结与验证 27](#_Toc91728482)

[10.2 改进与展望 28](#_Toc91728483)

[参考文献 28](#_Toc91728484)

[小记 28](#_Toc91728485)

[附录 29](#_Toc91728486)

# 1. 请说明分类算法衡量时准确率(accuracy)的局限性并举例说明。

## 1.1 准确率(accuracy)的局限性：

准确率是样本分类问题中最简单也是最直观的评价指标。但存在明显的缺陷：比如负样本占99%时，分类器把所以样本都预测为负样本也可以获得99%的准确率。所以，当不同类别的样本比例非常不均衡时，占比大的类别往往成为影响准确率的最主要因素，此时准确率指标并不足以说明分类器的好坏。

## 1.2 准确率(accuracy)的局限性举例说明：

举例说明：google抓取了argcv 100个页面，而它索引中共有10,000,000个页面,随机抽一个页面，要求分类是否是argcv的页面。如果以accuracy来判断分类效率，一个算法是直接将所有的页面都判断为"不是argcv的页面",因为这样效率非常高,而accuracy已经到了99.999%(9,999,900/10,000,000),领先绝大多数分类器，但事实上这个算法显然不是需求期待的。

# 2. 请说明卷积神经网络中参数共享和稀疏连接的含义。

## 2.1 参数共享

由于卷积神经网络要处理的输入图片数据较大，如果统一采用全连接层进行图像特征提取时，网络中要训练的参数数量太多，比如如果图片数据为100\*100，而第一层隐藏层神经元个数为10000，则需要一共个参数记录连接的权值，太大的参数数量不利于网络的训练和拟合。

卷积神经网络的参数共享就是为了解决这个问题：对一副图像采用统一的卷积核进行卷积操作，对原图像中每个大小为5\*5的窗口使用统一的卷积模板进行卷积，从而提取出卷积后的特征图；假设我们使用100个不一样的卷积核，分别执行上述操作，那么我们总共使用的参数只需要5\*5\*100 + 100（b偏置量数目），其结果远小于全连接层的参数数目，能极大地减小网络参数数量。

## 2.2 稀疏连接

在使用全连接层进行模型训练时，每第m+1层的神经元会接收所有第m层的神经元的输出数据，并将这些数据一同执行计算。

而在卷积神经网络中采用了稀疏连接的方法，即：每个隐藏层m中的每一个神经元具有固定的receptive field接收范围，也称为感受野的大小，即每个神经元只会固定接收一定数目的上一层神经元的数据。表现在空间域中，也就是卷积核的大小；比如一个3\*3大小的卷积核去执行卷积操作，得到的该像素点数据只跟原窗口的9个值相关，而在全连接层中下一层的每个输入都与该层的所有输出相关，故称为稀疏连接。

# 3. 请针对下述数据画出ROC曲线图。

## 3.1 第一步：按照属于‘正样本’的概率将所有样本排序

|  |  |  |
| --- | --- | --- |
| 元组编号 | 类 | 概率 |
| 1  3  6  4  5  7  8  2  10  9 | P  N  N  P  P  N  N  N  N  P | 0.97  0.70  0.63  0.60  0.55  0.53  0.51  0.50  0.40  0.30 |

## 3.2 对每个样本计算坐标

样本1：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 1 | 3 |
| 反例 | 0 | 6 |

X轴坐标（false positive rate）= 0/（0+6）= 0

Y轴坐标（true positive rate）= 1/（1+3）= 0.25

样本3：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 1 | 3 |
| 反例 | 1 | 5 |

X轴坐标（false positive rate）= 1/（1+5）= 1/6

Y轴坐标（true positive rate）= 1/（1+3）= 0.25

样本6：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 1 | 3 |
| 反例 | 2 | 4 |

X轴坐标（false positive rate）= 2/（2+4）= 2/6

Y轴坐标（true positive rate）= 1/（1+3）= 0.25

样本4：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 2 | 2 |
| 反例 | 2 | 4 |

X轴坐标（false positive rate）= 2/（2+4）= 2/6

Y轴坐标（true positive rate）= 2/（2+2）= 0.5

样本5：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 3 | 1 |
| 反例 | 2 | 4 |

X轴坐标（false positive rate）= 2/（2+4）= 2/6

Y轴坐标（true positive rate）= 3/（3+1）= 0.75

样本7：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 3 | 1 |
| 反例 | 3 | 3 |

X轴坐标（false positive rate）= 3/6= 0.5

Y轴坐标（true positive rate）= 3/（3+1）= 0.75

样本8：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 3 | 1 |
| 反例 | 4 | 2 |

X轴坐标（false positive rate）= 4/6= 2/3

Y轴坐标（true positive rate）= 3/（3+1）= 0.75

样本2：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 3 | 1 |
| 反例 | 5 | 1 |

X轴坐标（false positive rate）= 5/6

Y轴坐标（true positive rate）= 3/（3+1）= 0.75

样本10：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 3 | 1 |
| 反例 | 6 | 0 |

X轴坐标（false positive rate）= 1

Y轴坐标（true positive rate）= 3/（3+1）= 0.75

样本9：

|  |  |  |
| --- | --- | --- |
|  | 预测结果 | 预测结果 |
| 真实情况 | 正例 | 反例 |
| 正例 | 4 | 0 |
| 反例 | 6 | 0 |

X轴坐标（false positive rate）= 1

Y轴坐标（true positive rate）= 1

## 3.3 坐标结果与ROC曲线

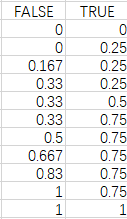


图3.1 坐标结果

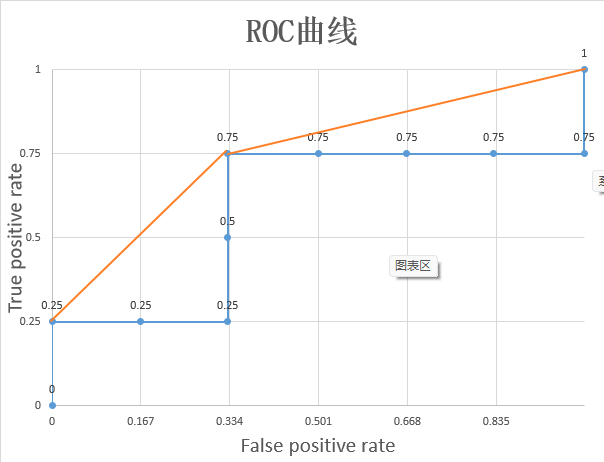


图3.2 ROC曲线

# 4. 选题背景

2019年12月，一场未知的病毒性肺炎疫情严重影响了中国武汉。该病毒很快被识别并命名为SARS-CoV-2。世界卫生组织随后提出将其称为2019年冠状病毒肺炎（COVID-19）肺炎。截至2021年12月底，已有超过200个国家和地区受到影响，累计确诊病例超过2亿8千万例，病例数还在不断增加。这一严峻形势突出表明，迫切需要制定有效措施，控制这一流行病。

对COVID-19肺炎患者进行早期诊断以便及时治疗对于控制疫情至关重要。但受COVID-19影响的许多地区的医疗资源有限和患者人数众多，通常会导致诊断和医疗决策（如隔离或住院）的等待时间很长，这可能会增加交叉感染的机会并导致预后不良。

因此，寻找一种能快速有效识别新冠疫情感染情况的方法就成了一个研究方向。虽然核酸检测能有效检测新冠感染情况，但基于CT图像扫描更为迅速，为快速临床诊断提供条件。

在COVID-19大流行期间，第一手CT图像数据和临床数据集的可用性对于帮助指导临床决策，提供信息以加深对这种病毒感染的理解以及为系统建模提供基础（可能有助于早期诊断以进行及时的医疗干预）将是至关重要和重要的。实现这一目标的一个方法是创建一个开放获取的综合资源，其中包含个体患者的胸部CT图像，通过对CT图像建模高效区分是否感染新冠疫情，能够促进国际共同努力抗击COVID-19肺炎。

# 5. 需求分析

## 5.1 编写目的

为能够在规定时间内开发出符合预期标准的基于numpy的DNN神经网络新冠肺炎高效CT图像识别分类模型，编写此需求说明书。

模型要求能在保证模型微量化快速化的同时，对输入CT图像有高于95%的新冠疫情预测准确率。完成这些目标需要一定的计划与明确的解决方案。

本需求分析说明书适用于该项目的客户方管理人员、需求分析员，用户文档编写者，项目管理员，项目产品开发人员，产品测试人员以及技术支持人员。

## 5.2 目标

设计一个基于numpy的DNN神经网络新冠肺炎高效CT图像识别分类模型，其要保证模型轻量化，训练简单化的同时保证高准确率。对输入肺部CT图像进行识别分析，给出其具体所属的类别，是新冠肺炎阳性、阴性或无信息。

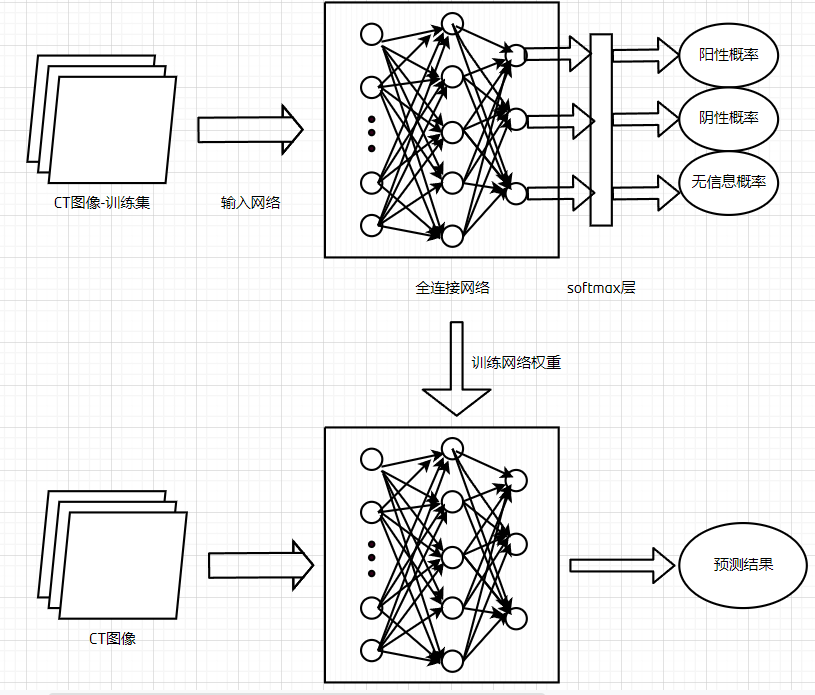


图5.1 对模型的需求流程图

## 5.3 用户的特点

用户在使用该模型时，首先考虑到医疗需求需要尽可能准确分类，所以在模型测试时要保证高精度识别出结果，最好能将准确率控制在95%以上。同时由于医疗设备特殊，需要模型尽可能轻量化便于实装；同时考虑到多种应用途径，一个合适的神经网络模型能拟合多种数据分类，所以还要求网络结构尽可能简单易用，同时易于修改和训练。

## 5.4 对性能的规定

### 5.4.1 精度

模型为了满足医疗诊断和疫情防控对时间和精度的高要求，要求在训练集上准确率accuracy不小于98%、在测试集上准确率accuracy不小于97%。

### 5.4.2 时间特性要求

处理时间：实装搭载模型后要求一张图片的网络处理时间不超过0.1s（不包括读取图片时间），并能在总时间0.5s内完成全部读取、处理和输出结果流程。

### 5.4.3 灵活性

为保证模型使用灵活性，模型总大小应不超过5MB，并能灵活更改模型隐藏层大小、层数与激活函数，便于依据同一网络模型针对不同数据进行拟合。

### 5.4.4 输入输出要求

输入：图片分辨率大小为64\*64，通道数为3的图片数据

输出：经softmax分类器输出的最大可能性的标签值，共3类信息，包含“Ni”（无信息）、“N”（阴性）、“P”（阳性）。

## 5.5 运行环境规定

运行配置要求：

（1）系统    Windows Vista 64 Bit Service Pack 2 以上 。

（2）CPU：Intel Core 2 Quad CPU Q6600（4核,2.40GHz）以上

（3）内存需要4GB。

（4）硬盘需要2GB。

（5）Python 版本不低于2.7.0

# 6. 解决方案概述

## 6.1 方案概述

基于上述背景及需求分析，我设计并实现了一种基于numpy的DNN神经网络新冠肺炎高效CT图像识别分类模型。该模型主体框架由纯numpy实现的神经网络各个步骤组成，包括定义网络的w与b的维度大小与初始化方式、前向传播、激活函数Relu的计算与求导、反向传播与梯度下降进行参数的更新。在实现全连接网络的基础上，还额外实现了数据的预处理、数据读取、accuracy预测和最终test范例的脚本代码。

使用该神经网络模型进行CT图像新冠肺炎诊断，有如下4处优点与创新点：

## 6.2 优点&创新点1-选题新颖贴合时事

在新冠疫情爆发近两年来，无数人内心被疫情牵动。宝贵的新冠病毒数据集被收录，背后是一个又一个患者与他身后的故事。本课程设计运用商务智能的预测性分析知识，结合深度学习与医疗生物专业，在与人们生命安全息息相关的医疗健康领域，想要通过机器学习的知识为驱散疫情贡献出自己的一份力量。

## 6.3 优点&创新点2-模型精确度高

经网络训练测试，在训练集和测试集的accuracy分别超过了99%和97.8%，实现了网络对数据集较为精准的拟合，精确度已达到能够有效帮助诊断的大小，能够在一定程度上给予医疗人员诊断帮助。

## 6.4 优点&创新点3-未借助深度学习框架，numpy底层实现

该神经网络实现方式为numpy库进行底层函数编写，从最基本的神经网络各个步骤组成，包括定义网络的w与b的维度大小与初始化方式、前向传播、激活函数Relu的计算与求导、反向传播与梯度下降进行参数的更新，不借助深度学习框架辅助网络训练。

## 6.5 优点&创新点4-网络结构简单轻量化

该神经网络模型包含三个隐藏层，总参数大小不超过5MB，部署方便轻量快捷。

# 7. 数据集介绍

## 7.1 数据集来源

数据集来源：<http://ictcf.biocuckoo.cn[1>]

## 7.2 数据集介绍

该数据集由华中科技大学医院（HUST-UH/ HUST-LH）收集，将单个CT切片分为三种类型：（1）非信息性CT（NiCT）图像，其中肺实质未被捕获以供任何判断;（2）CT（pCT）阳性图像，其中可以明确识别与COVID-19肺炎相关的成像特征;和（3）阴性CT（nCT）图像，其中两个肺部的成像特征与COVID-19肺炎无关。

其中，nCT共包含9979张图片、pCT共包含4001张图片、NiCT共包含5705张图片。图片分辨率大小均为512\*512，位深度均为8

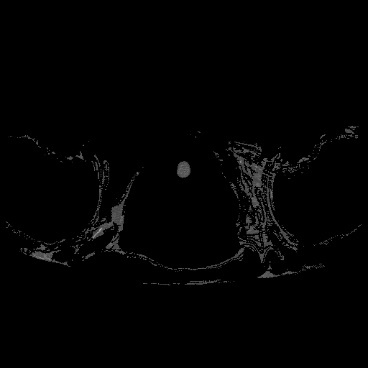
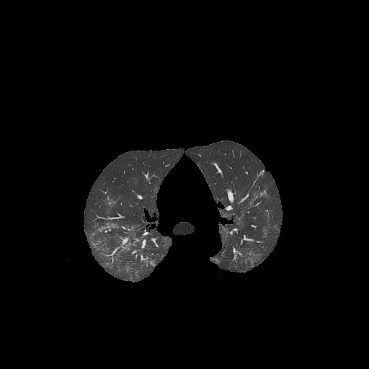
  

图7.1 从左往右依次是nCT0006、NiCT0015、pCT0040

## 7.3 数据集预处理

由于原图像大小过大，不适合全连接网络直接进行训练，于是采用图像下采样的方法，对图像进行了下采样，具体大小设置为64\*64大小，保持通道数和位深度不变，处理后的图片可以直接全部读入内存中进行训练。

图7.1 从左往右依次是下采样后的nCT0006、NiCT0015、pCT0040

## 7.4 训练集与测试集分割

本次实验首先将训练集和测试集按9：1比例划分，然后在保存图片路径和label的txt文件中打乱顺序，形成三种label均匀分布的数据集合。得到train图片17714张，test图片1969张。

# 8. 模型设置与关键代码实现

## 8.1 模型配置说明

该神经网络模型采用三个隐藏层的全连接网络架构，其网络神经元数目分别为：[64\*64\*64,50,30,20,10,config.numClass] （其中config.numClass = 3）



图8.1

激活函数选用Relu,输出层计算概率选用softmax层进行分类，与之对应选择CrossEntropy交叉熵损失函数作为优化的根本依据。

参数初始化方法为随机初始化，优化策略为梯度下降法，将训练集的17714张图片拼接为一个多维向量投入网络进行训练。

学习率采用0.075、0.05、0.01三个梯度，训练总epoch数量为3000，具体训练策略在第9节会详细给出具体数据。

## 8.2 模型使用说明

程序入口在’code/model.py’文件中，可以手动选择是否加载预训练参数以及预训练参数的读取路径；在layers\_dims中定义模型的隐藏层大小数量；在L\_layer\_model函数中定义训练的epoch数量和learning rate，在训练完成后会自动保存参数并输出在测试集上验证结果。

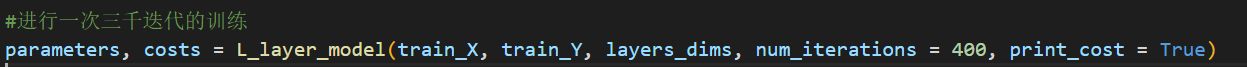


图8.2

## 8.3 参数初始化、前向传播与激活函数、反向传播

参数初始化中，传入一个表示网络层数与大小的 dim 向量，根据 dim 对每一层的参数进行初始化，权重设置为接近 0 的随机数，bias 偏置设置为 0

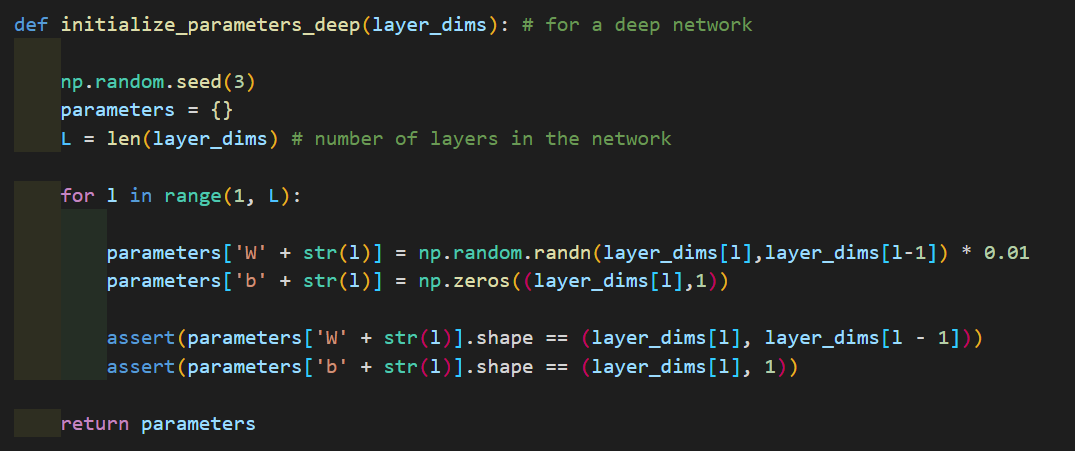


图8.3

线性前向传播，输入上一层的计算结果和权重以及偏置，输出计算的线性结果和保存的缓存便于反向传播。激活函数选择Relu实现。

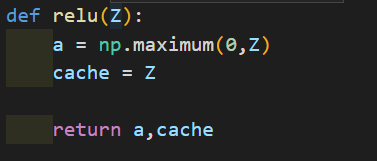
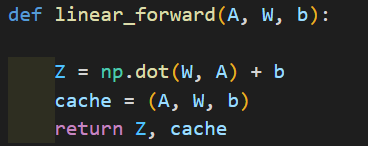


图8.4

前向传播通过输入样本 X 和参数 parameters,输入一次前向计算的结果 AL 与计算过程中存储的缓存 cache；反向传播输入计算结构 AL，样本真实值 Y 和运算中存储的 cache，先对将交叉熵损失对输出层的激活函数 softmax 求偏导得到:

=\*𝐴𝐿 − 𝑌

将这个结果带入到接下来的梯度计算中，得到每一层的梯度，输出梯度。

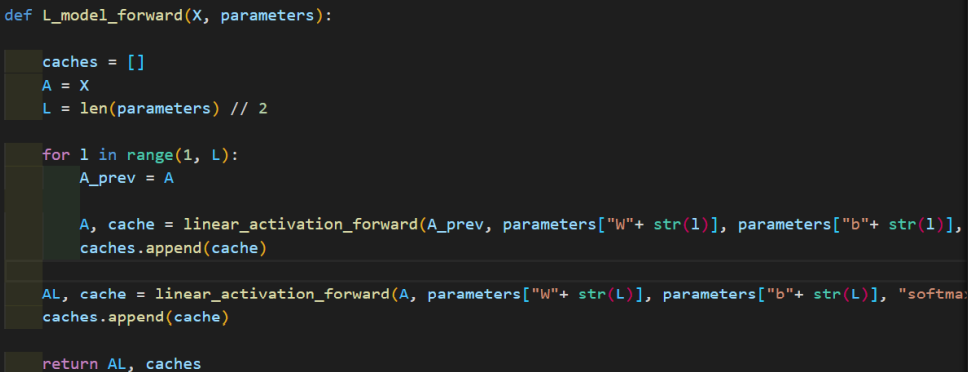


图8.5 前向传播

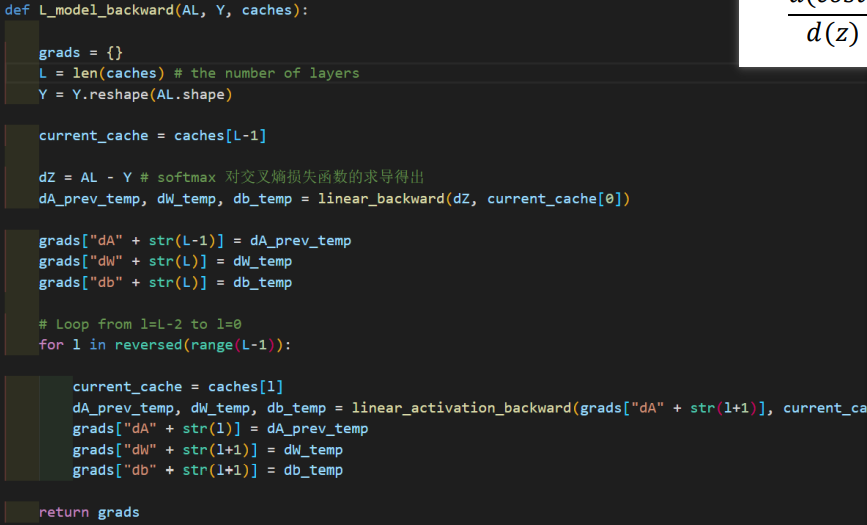


图8.6 反向传播

## 8.4梯度下降与结果预测

梯度下降方法输入模型原参数、梯度和学习率，对每一层的参数进行更新。

结果预测方法输入计算结果和真实标签，通过计算得到成功预测的准确率。

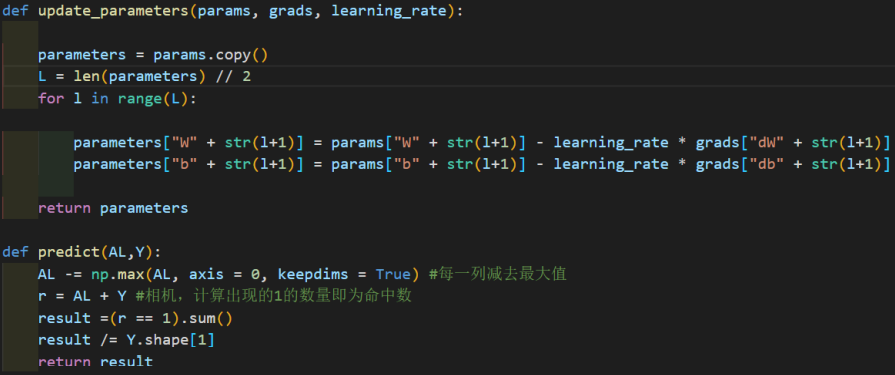


图8.7

## 8.5 数据预处理、加载数据、开始训练

数据预处理脚本首先获取数据所在的目录，根据图片所在的目录名记录为它的label标签值，将其预处理后的图片保存在miniCTscans文件目录中，并将新路径与其label标签值拼接作为一行数据写入txt文件中，目的是方便后续读取数据集同时也读入label。代码如下：

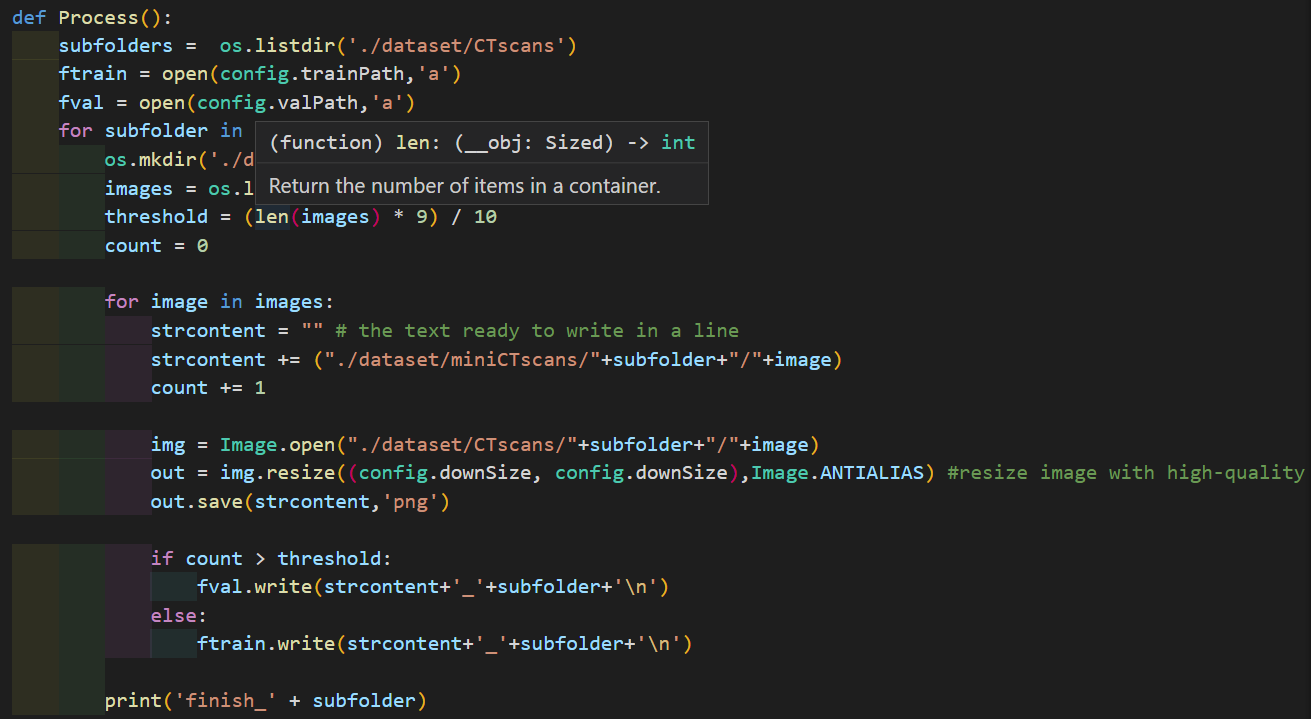


图8.8 数据预处理

加载数据分图像数据与标签数据两部分加载。首先在按行读取txt文件后获得string格式的标签值与图片地址，根据图片地址读取图片数据后将其转为一维长向量并拼接到总训练数据的末端；标签数据则根据config中定义的dictionary转换成一维向量，同时也拼接到总label数据的末端。代码如下：



图8.9 加载数据

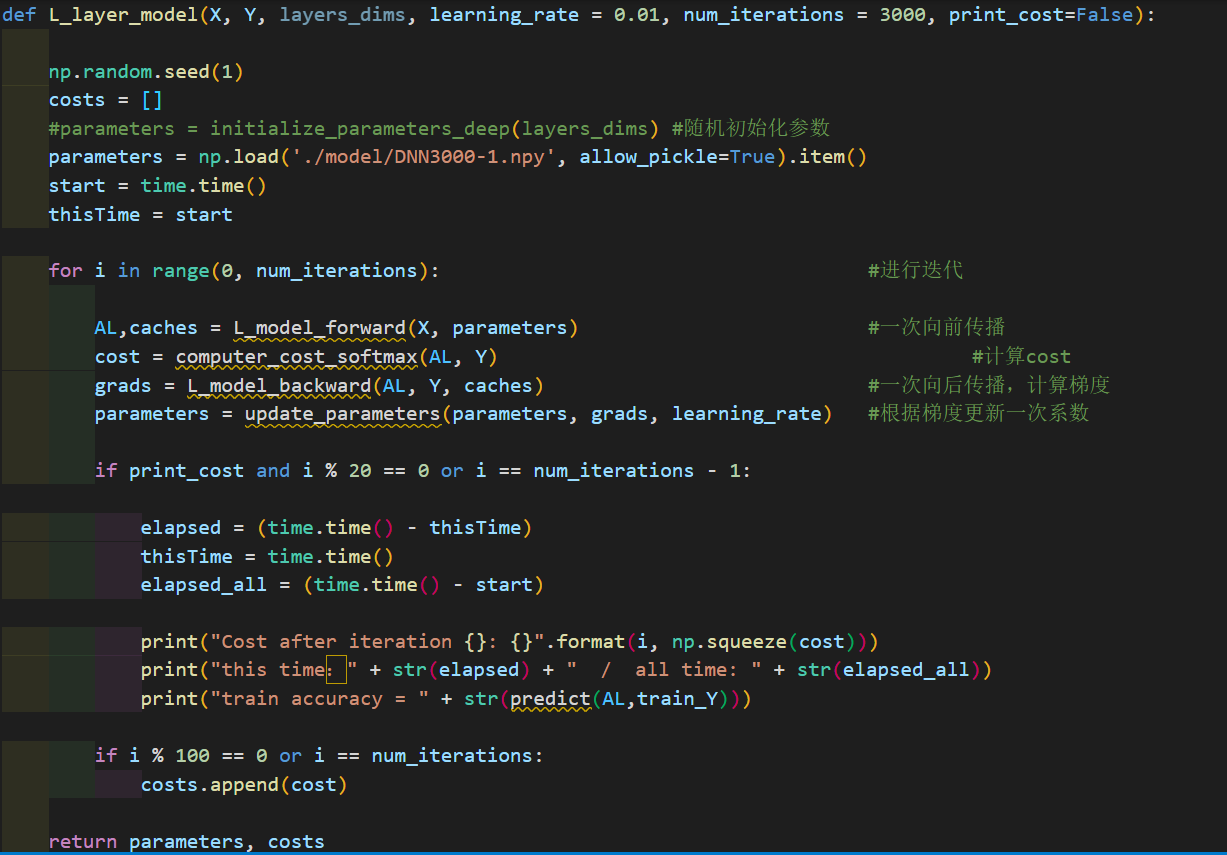


图8.10 开始训练

# 9. 模型训练与测试结果

## 9.1 多隐藏层训练结果

训练策略如下：

首先以0.075的学习率训练200个epoch

再以0.05的学习率训练了80个epoch

最后以0.01的学习率训练了2720个epoch

训练过程日志文件记录如下（部分）：

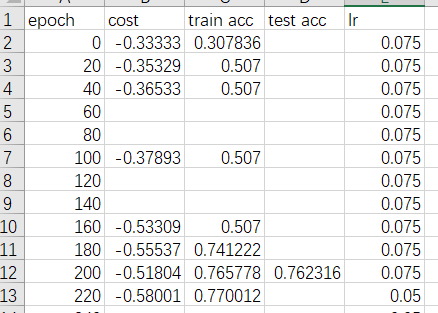


图 9.1

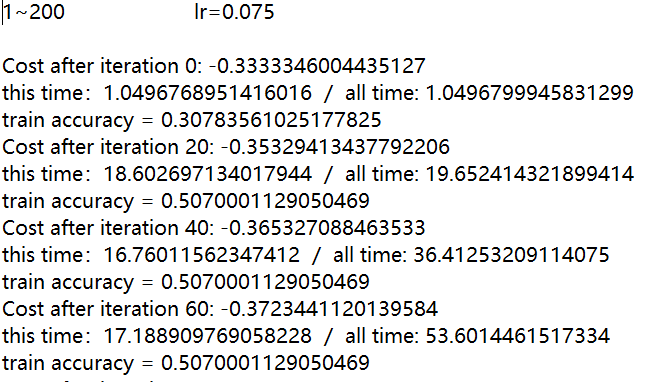


图9.2

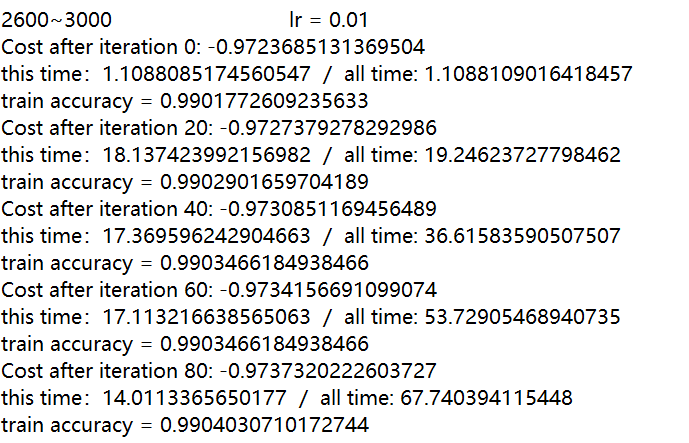


图9.3

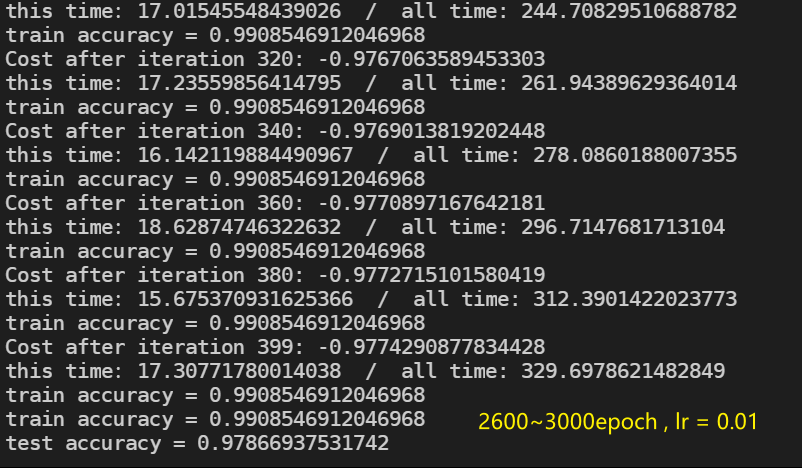


图9.4

Cost变化如下：

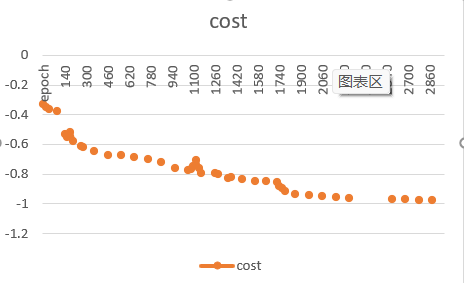


图9.5

Accuracy变化如下：

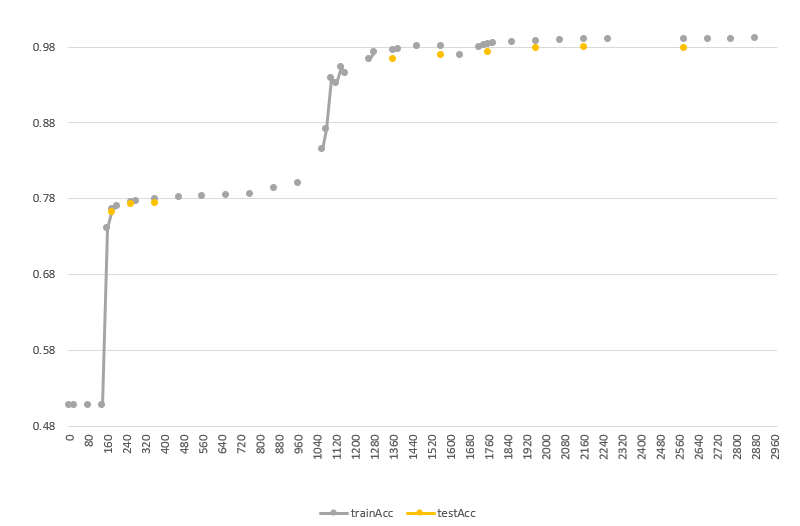


图9.6

不难看出，在训练末期trainAcc达到了0.99以上，testAcc也超过了0.978。

## 9.2 单隐藏层训练结果

为保证实验完整性，另外选用大小为30的单个隐藏层神经网络进行训练，选用0.0075学习率进行3000epoch的训练，最终结果如下：

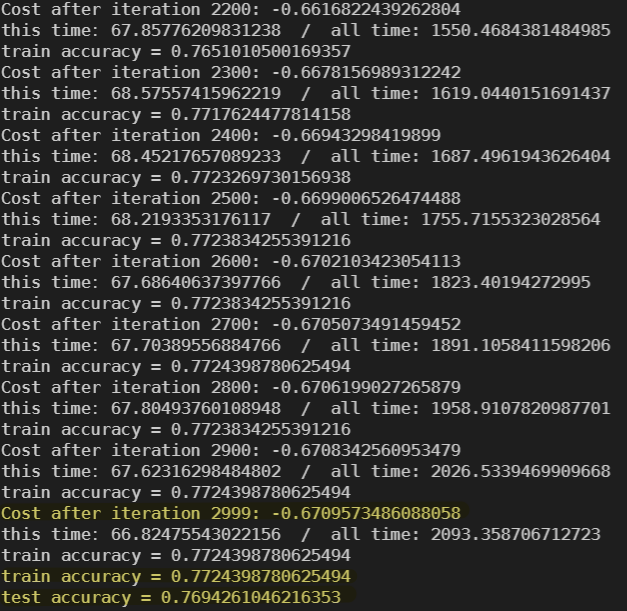


图9.7 单隐藏层训练结果

可以发现单隐藏层得到结果远低于多隐藏层。

# 10. 总结与讨论

## 10.1 实验总结与验证

经上述实验证明，本次课程设计实现的CT图像分类模型，在经过适当训练后得到了较高的精度、较简单的模型结构，较小的模型参数大小与快速的收敛速度。在实验的最后我构建了一个应用该网络模型参数的脚本，利用参数对随机选择的三类图像进行了分类，毫无疑问都取得了正确结果和较高的概率。

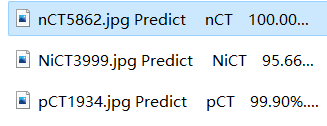


图10.1 从左到右依次是图片名称、预测得到的标签值、预测概率

课程设计实验到此圆满结束，需求所需模型也成功构建。

## 10.2 改进与展望

这次多隐藏层的效果其实远远超出我的预期，我没有想到会有如此高的模型拟合度，毕竟第一次尝试单隐藏层只能训练到0.78不到的acc值。

能改进的地方还有很多：尝试使用cnn卷积网络、尝试更多层数、尝试更优质的梯度下降方法和参数初始化方法、使用正则化防止过拟合等等……

也许在不久的将来，这个模型能被不断完善，焕发光彩。

# 参考文献

[1] Ning, W., Lei, S., Yang, J. et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat Biomed Eng 4, 1197–1207 (2020). https://doi.org/10.1038/s41551-020-00633-5

# 小记

感谢老师一学期如一日的辛勤付出。老师诙谐幽默的课堂氛围和透彻易懂的讲解方式令我陶醉于课堂之中。我知道自己在很多方面还有不足和没有领悟的地方，通过这次课设也渐渐让我学到更多，不只是模型的架构、训练与测试，更多的是对精益求精的思考和完善。通过这段时间的学习，我也更激发了对机器学习和大数据分析的学习热情。

再次感谢！

# 附录

该课程设计github仓库地址为我个人仓库：<https://github.com/ssd777/CT-imageClassification-Covid19>

预处理后的64\*64大小图片数据集也过大，不便呈现。

训练日志文件与excel文件、模型文件已包含于压缩包中。