

**2024年《机器学习》**

**工程报告**

**（个人版）**



**课 程：** 机器学习

**姓 名：** 刘东林

**学 号：** 2022217598

**完成时间：** 2024.5.19

|  |  |  |  |
| --- | --- | --- | --- |
| **“机器学习-大作业”评分细则** | | | |
| **成绩等级** | **具体表现** | **教师评分** | |
| 优秀（100-90] | 报告撰写优秀，题目本身难度大，工作量饱满；Notebook代码清晰准确，实验与可视化优秀 | □ |  |
| 良好（90-80] | 报告撰写良好，题目本身难度较大，工作量较多；Notebook代码较好，实验与可视化良好 | □ |  |
| 中等（80-70] | 报告撰写中等，题目本身难度一般，工作量达标；Notebook代码一般，实验与可视化达标 | □ |  |
| 及格（70-60] | 能完成基本要求，工作量较少，非Notebook代码 | □ |  |
| 不及格（<60） | 未达最低要求或抄袭线上线下资源 | □ |  |

教师签名：

一． 自己所作工作的简介

基于PyTorch的CNN模型进行了改进，帮助同学搭建环境、安装驱动。本项目是对四种血细胞分类，对于给出的血细胞图片预测其为四种中的哪一种。最终测试正确率在88%左右。

https://www.kaggle.com/datasets/paultimothymooney/blood-cells

二、研究背景与意义

在临床检查中，血液疾病的诊断通常需要对患者的血液样本识别。因此，检测和分类血细胞亚型的自动化方法在医学上有重要的应用。

三、模型方法

使用卷积神经网络（CNN）。

模型有6个卷积层、6个池化层、3个Dropout层、5个全连接层。每个卷积层和全连接层后加ReLU激活函数。

输入为240 × 240 × 3，经过第一个卷积层后240 × 240 × 32，经过第一个池化层后120 × 120 × 32；经过第二个卷积层后120 × 120 × 64，经过第二个池化层后60 × 60 × 64；经过第三个卷积层后60 × 60 × 128，经过第三个池化层后30 × 30 × 128；经过第四个卷积层后30 × 30 × 256，经过第四个池化层后15 × 15 × 256；经过第五个卷积层后15 × 15 × 512，经过第五个池化层后7 × 7 × 512；经过第六个卷积层后7 × 7 × 1024，经过第六个池化层后3 × 3 × 1024。

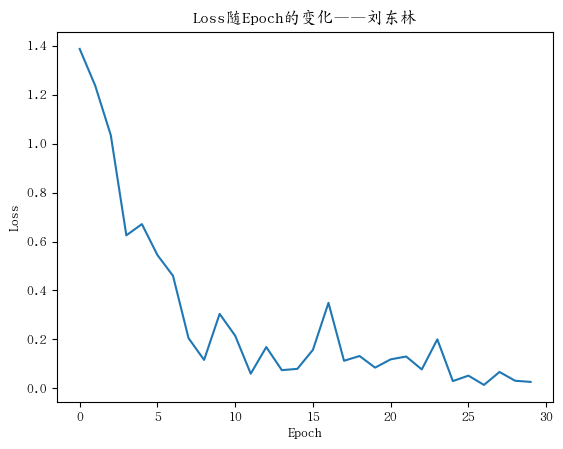
在第二个、第四个、第六个池化层后加一层随机dropout。

四、系统设计

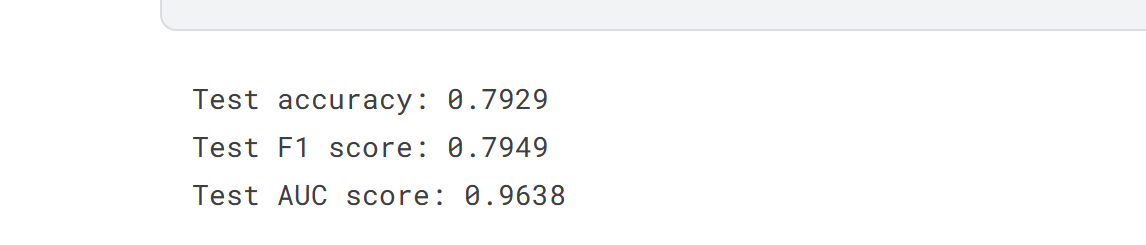
1. 首先使用ImageFolder将数据集导入，transforms使尺寸设为240×240，加入随机反转和随机旋转。
2. batch\_size设为128，然后用DataLoader导入加载训练数据集和测试数据集。
3. 选择一种device。
4. 创建一个CNN的模型，添加相应的卷积层、池化层、全连接层等。
5. 设定loss方法和优化器。
6. 设置epoch数目。
7. for循环训练模型。
8. 以epoch为横坐标，loss为纵坐标画图。
9. 计算模型在测试集上的正确率，F1和AUC。

五．实验结果分析、对比和讨论

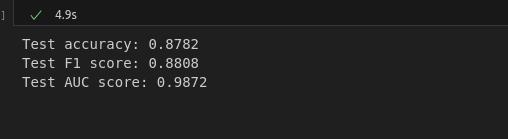
Loss：



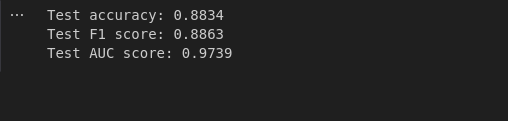
修改前的准确率、F1、AUC：



改进后模型的准确率、F1、AUC：



在截图中准确率最高一次：



六．其他

环境：PyTorch 2.3.0 + rocm 6.0

Ubuntu 22.04

由于设备是AMD GPU，安装PyTorch的过程较为繁琐。需要在Linux下，并安装Rocm后方可使用GPU运行。

七．对本门课的感想、意见和建议

顺带练习了Linux的使用，并学习了conda创建和使用虚拟环境，以及Python包的管理等。

课程难度较高，希望偏向基础一些，并增加实操教程。