

**2024年《机器学习》**

**工程报告**

**（个人版）**



**课 程：** 机器学习

**姓 名：** 魏海严

**学 号：** 2022217594

**完成时间：** 2024/5/18

|  |  |  |  |
| --- | --- | --- | --- |
| **“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

在本项目中，我的主要工作是通过改进模型设计并引入集成学习方法来提高血细胞图像分类的准确性。具体工作包括数据增强与预处理、改进卷积神经网络（CNN）架构、采用集成学习，训练多个模型并结合它们的预测结果，以及进行实验分析以评估改进效果。

具体内容：

**数据增强与预处理：**

使用图像增强技术（包括调整大小、随机翻转、旋转和颜色抖动）来丰富训练数据，提升模型的泛化能力。

标准化处理图像以加快模型收敛速度和提高稳定性。

**模型设计与改进：**

构建了一个改进的卷积神经网络（ImprovedCNN），引入了残差模块（ResidualBlock）以加深网络结构，提高模型的表达能力。

添加了Dropout层以减少过拟合现象。

**训练与优化：**

使用**Adam优化器**和交叉熵损失函数训练模型，设置**学习率调度器**来动态调整学习率。

**采用集成学习**，训练了多个相同架构的模型，并在测试时平均它们的预测结果，以提高整体性能。

实验结果与分析：

通过多次实验记录损失值变化，绘制训练过程中训练与验证的损失曲线。

使用准确率（accuracy）、F1分数（F1 score）和ROC AUC分数（AUC score）评估模型在测试集上的表现。

对单张图片进行预测，并分析集成模型的预测结果。

二、研究背景与意义

说明本工程所涉及的研究背景研究的意义，包括主要涉及领域，主要研究方法，主要存在问题，现有解决方案等等，可以自行补充其他内容。

**研究背景：**

血液疾病的诊断通常涉及对患者血液样本的识别和表征。传统方法依赖于人工显微镜检查，耗时且易受主观因素影响。自动化方法在血细胞亚型的检测和分类方面具有重要的医学应用。

**研究的意义：**

血液性疾病的诊断通常需要检查患者的血液样本，以识别血液细胞形态的异常。由于其效率和准确性，自动化检测和分类血液细胞亚型的方法在现代医学实践中变得越来越重要。

手动分类血液细胞亚型耗时且容易出现人为错误，特别是在数据集较大的情况下。血液细胞形态和染色技术的变化进一步复杂化了分类过程。

传统的血液细胞分类方法涉及由训练有素的血液学家手动进行显微镜检查和目视检查。机器学习和深度学习技术已经成为自动化血液细胞分类任务的有前途的方法。这些技术利用像这里提供的注释数据集来训练能够准确识别和分类血液细胞亚型的模型。

自动化血细胞分类能够显著提高诊断效率，减少人为误差，为医疗领域提供强有力的技术支持，尤其在大规模筛查和实时诊断中具有重要意义。

像这个数据集一样的注释数据集在训练和评估用于血液细胞分类的机器学习模型中发挥着至关重要的作用。这些数据集为训练模型提供了地面真实标签，促进了准确可靠的分类算法的开发。

**主要涉及领域：**

**计算机视觉**：利用深度学习技术对血细胞图像进行处理和分类。

**医学影像分析**：通过自动化方法辅助医学诊断，做到智慧医疗。

**主要研究方法：**

数据增强与预处理：通过图像增强技术扩展数据集，提高模型的泛化能力。

卷积神经网络（CNN）：设计和改进深度学习模型以提升分类性能。

集成学习：通过训练多个模型并结合它们的预测结果提高模型的准确性和稳定性。

**主要存在问题：**

数据集较小：部分血细胞类型的数据较少，可能导致模型难以很好分类。

过拟合现象：复杂的模型可能在训练集上表现良好，但在测试集上表现不佳。

**现有解决方案：**

本项目使用的数据集包含12,500张血细胞图像，分为四种类型：嗜酸性粒细胞、淋巴细胞、单核细胞和中性粒细胞。数据集经过增强处理，增加了数据量，提供了更多训练样本。通过引入集成学习方法，结合多个模型的预测结果，提高了分类模型的性能，为血液疾病的自动化诊断提供了技术支持。传统的血液细胞分类方法涉及由训练有素的血液学家手动进行显微镜检查和目视检查。机器学习和深度学习技术已经成为自动化血液细胞分类任务的有前途的方法。我利用该数据集，采用相关机器学习算法，对数据进行预处理，利用适当模型，对血细胞图像进行分类。

三、模型方法

在本工程中，我采用了深度学习中的卷积神经网络（CNN）及其改进形式来实现血细胞图像的分类任务。主要使用的方法和理论如下：

**数据增强（Data Augmentation）：**

通过对训练数据进行随机翻转、旋转、颜色抖动等操作，增加数据的多样性，提高模型的泛化能力。

**卷积神经网络（CNN）**：

CNN是一种专门处理图像数据的深度学习模型，通过卷积层提取图像的空间特征，通过池化层减少特征图的维度，同时保留关键信息。

残差网络（Residual Networks，ResNet）：

引入**残差模块**，通过跳跃连接（skip connections）来缓解深层网络的梯度消失问题，使得更深层的网络可以更有效地训练。

**系统介绍模型**：

设计的模型是一个**改进的卷积神经网络（CNN）**，用于图像分类任务。首先，模型接受输入图像（RGB格式，即3个通道），然后通过一系列的卷积、池化和全连接层对图像进行处理，最终输出图像的分类结果。

模型的主体结构包括卷积层、批归一化层、残差块和全连接层。卷积层用于提取图像的特征，批归一化层有助于加速模型训练并提高模型的泛化能力。残差块是模型的核心组件，通过残差连接（residual connection）可以解决深度神经网络中的梯度消失和梯度爆炸问题，从而使网络更易于训练和优化。

在残差块中，首先通过一个卷积层和批归一化层对输入进行处理，然后再经过另一个卷积层和批归一化层处理。同时，利用残差连接将输入直接加到最后的输出上，以确保信息的流动。这样的设计使得模型能够更好地捕捉图像中的细节信息，并且具有更好的训练效果。

模型的结构采用了多层的残差块，并在每个残差块之间插入了最大池化层，以逐渐减小特征图的尺寸。最后，通过全连接层将特征图展开成一维向量，并经过几层全连接层进行分类，最终输出图像的类别。

此外，模型还包括了dropout层，用于在训练过程中随机丢弃部分神经元，以防止过拟合。整个模型的设计旨在提高图像分类的准确性和泛化能力，使其适用于各种不同的图像分类任务。

**集成学习（Ensemble Learning）：**

训练多个独立的CNN模型，并将它们的预测结果进行综合，以提高分类的准确性和稳定性。

优化算法：

使用**Adam优化算法**，结合**学习率调度器（learning rate scheduler）**动态调整学习率，加速模型收敛并避免过拟合。

损失函数：使用交叉熵损失函数（Cross Entropy Loss）来衡量模型预测与真实标签之间的差异。

四、系统设计

**数据处理流程：**

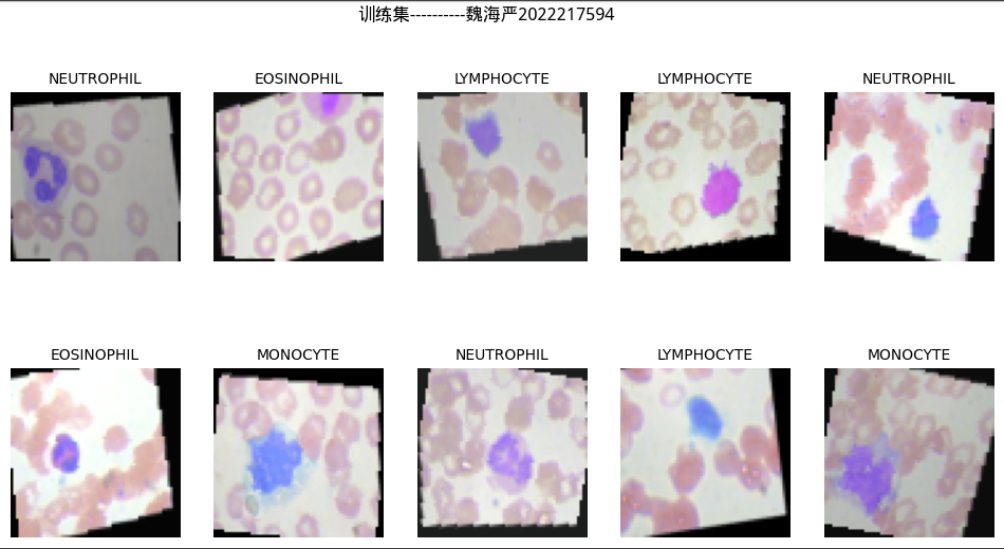
数据加载： 使用torchvision.datasets.ImageFolder加载训练和测试数据集。

数据增强： 使用torchvision.transforms对训练数据进行增强处理，如调整大小、随机翻转、旋转和颜色抖动等。

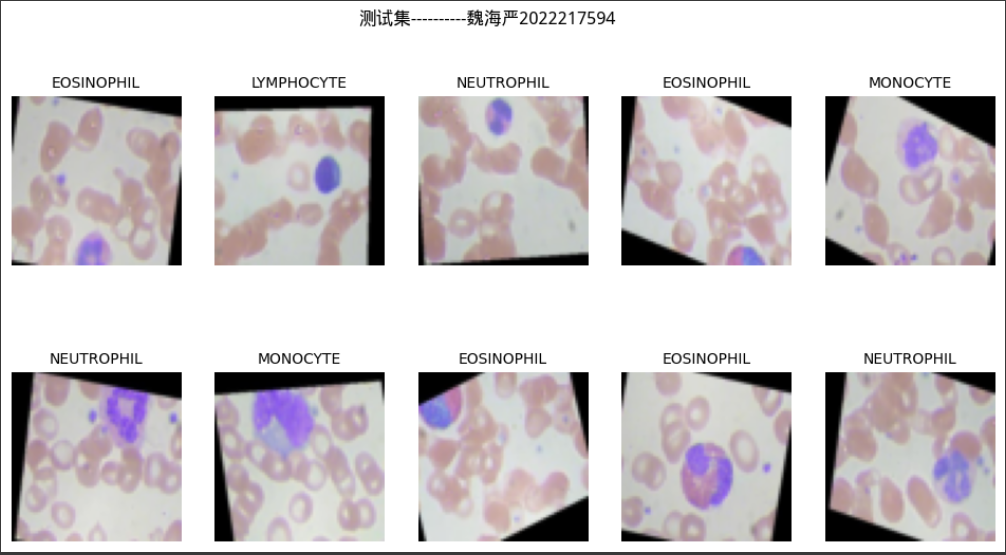
标准化： 对图像数据进行标准化处理，使得图像数据在同一尺度上进行训练。

**查看数据：**

浏览训练数据集和测试数据集中各个类相关数据，随机选取部分图像，将图像输出，分别显示训练集和测试集的图像。在浏览数据之前，需要对数据进行逆标准化处理



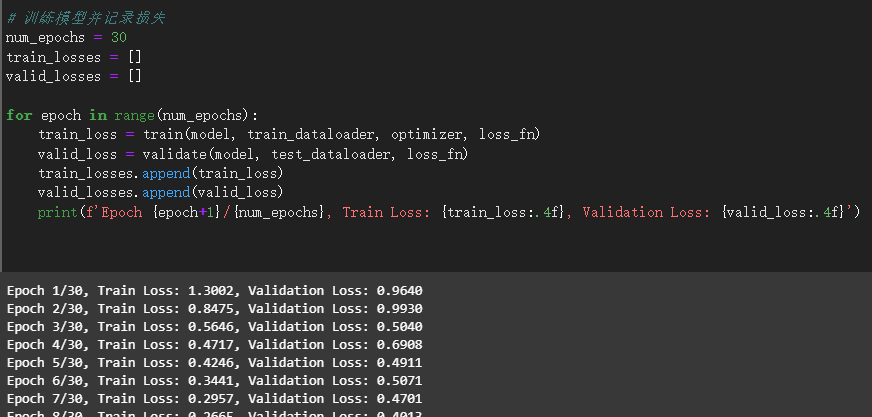
从给定的数据集中随机采样图像，并以网格形式显示这些图像



模型训练流程：

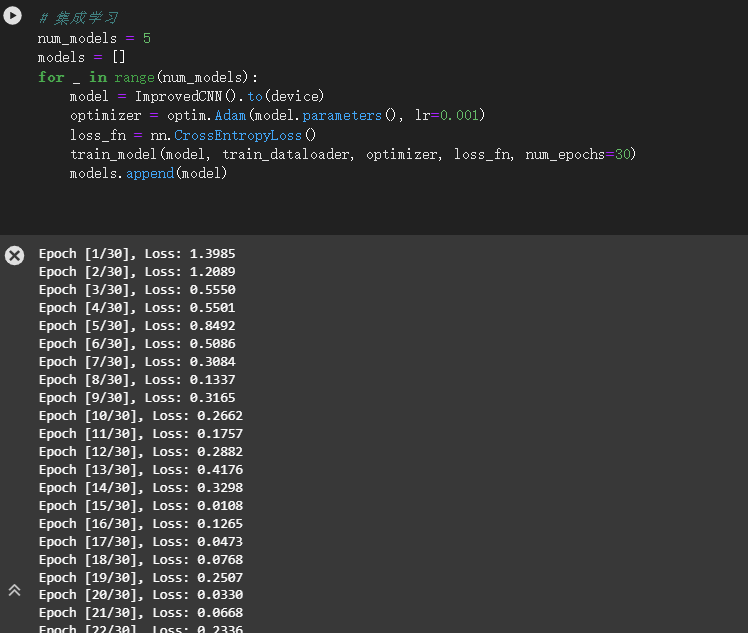
**模型定义：** 定义改进的卷积神经网络（ImprovedCNN），包括卷积层、池化层、残差模块、全连接层和Dropout层。

**单独模型训练**： 使用定义好的模型、优化器（Adam）和损失函数（交叉熵损失）对数据进行训练，设置学习率调度器来动态调整学习率。



**集成模型训练**： 训练多个独立的CNN模型，以构建集成模型。

模型测试流程：



上图为5个学习器的训练损失

**评估单独（非集成）模型**

**评估集成模型：** 对测试数据集进行预测，通过平均多个模型的预测结果来得到最终的分类结果。

评估指标： 使用准确率（accuracy）、F1分数（F1 score）和ROC AUC分数（AUC score）来评估模型性能。采用**混淆矩阵**评估模型性能。将单独（非集成）模型的性能与集成模型的性能做比较。

**单张图片预测流程：**

图像预处理： 对单张图片进行预处理，包括调整大小和标准化。

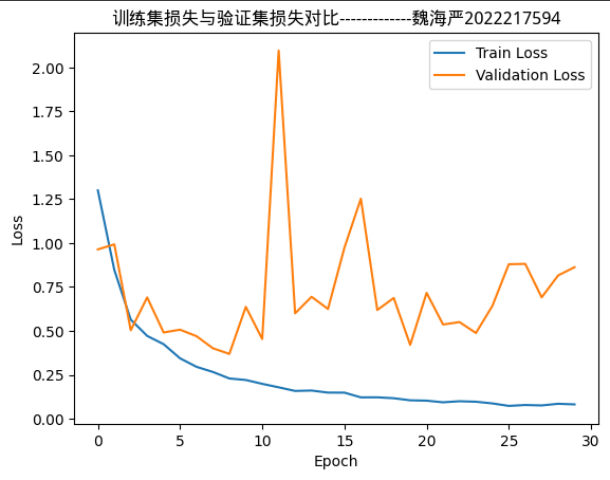
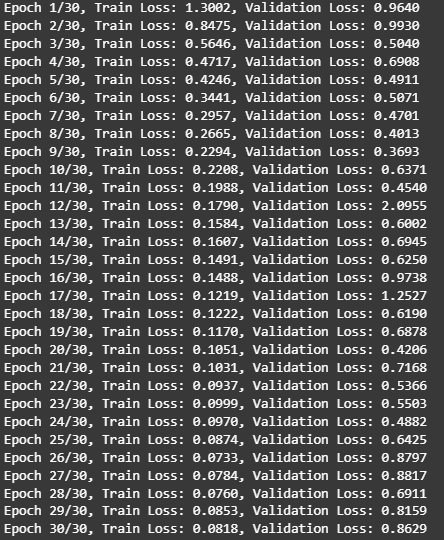
预测： 使用集成模型对预处理后的单张图片进行预测，得到概率分布和最终分类结果。

五．实验结果分析、对比和讨论

**实验结果分析：**

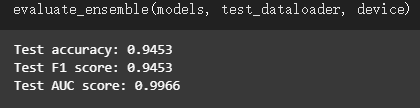
模型训练过程中的损失变化：

在训练过程中，通过记录每个epoch的损失值，可以观察到模型的收敛情况。随着训练进行，损失值逐渐减小，表明模型在不断学习和优化。



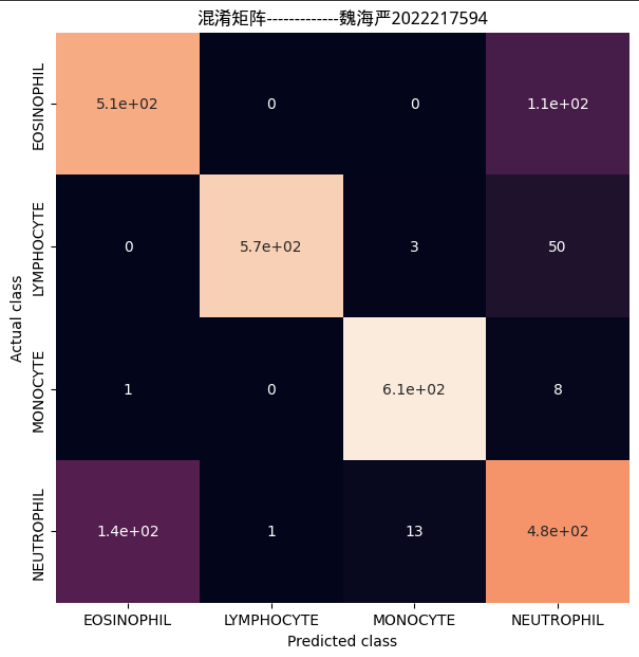
评估结果：

在测试数据集上，通过准确率、F1分数和AUC分数来评估模型性能。实验结果显示，集成模型在各项指标上均表现出较单独模型更高的性能，证明了集成学习方法的有效性。



可以看出，集成模型有很高的准确率，F1值和AUC分值，模型效果好。

通过混淆矩阵可以看出，对角线元素占比很高，非对角线方格数据多数趋近于0，说明模型预测精确。



在单张图片预测任务上，可以看到模型预测正确

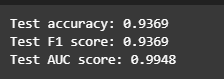


**对比和讨论：**

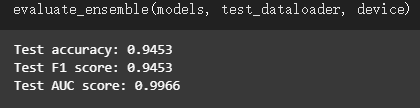
单模型与集成模型的对比：

集成模型的准确率、F1分数和AUC分数均优于单个模型的结果，表明集成模型通过结合多个模型的优势，能够提高整体性能和鲁棒性。

**单个模型分类指标值**：



**集成模型分类指标值：**



可以看到 集成模型性能略有提升。

不同超参数设置的对比：

对学习率、批次大小、迭代次数等超参数进行调整，通过实验发现适当的超参数设置能够显著影响模型的训练效果和最终性能。其中通过多次实验发现epoch为30时，学习率取0.001时，batch size 32时最合适

**改进方案：**

数据集扩展：

引入更多相关样本数据作为训练集的一部分，以进一步提高模型的分类准确率以及模型的泛化能力。

模型架构改进：

尝试融入更复杂的且在其他领域效果较好的网络架构，如GCN、SENet等，以进一步提升模型的分类能力。

超参数优化：

使用超参数优化技术，如网格搜索（Grid Search）或贝叶斯优化（Bayesian Optimization），以找到最佳的超参数组合

。

模型融合技术：

除了平均融合，可以尝试其他融合方法，如加权平均、投票法等，以进一步提升集成模型的性能。

通过以上改进方案，可以进一步提高血细胞图像分类模型的准确性和鲁棒性，推动自动化医学诊断技术的发展。

六．其他

在本项目的研究过程中，通过深度学习和集成学习方法对血细胞图像分类进行了深入的探索和实验。这不仅提升了我们的技术能力，也使我们深刻理解了计算机视觉技术在医学诊断中的重要应用。特别是通过实际数据的处理、模型的设计和优化、以及实验结果的分析，我们进一步掌握了如何将理论知识应用于实际问题解决中的技能。

在实际应用方面，自动化血细胞分类技术可以显著提高医疗诊断的效率和准确性，为医生提供有效的辅助工具，减轻工作负担。此外，这项技术还可以应用于其他图像分类任务，如癌细胞检测、病理图像分析等，为医学影像领域的发展提供更多可能性。

七．对本门课的感想、意见和建议

**感想：**

机器学习课程是一门结合理论和实践的课程，一个非常有趣且有用的领域，拥有广泛的应用场景。它为学生提供了探索和理解人工智能领域的基础。通过学习课程，我深入理解机器学习算法的原理和工作原理。了解不同类型的算法如何在不同情况下运作，并学会选择合适的算法来解决特定的问题。通过本门课程的学习，我系统地掌握了机器学习，特别是深度学习计算机视觉相关领域的核心理论和实践方法。课程内容涵盖了从基础理论到实际应用的各个方面，使我能够全面了解并掌握这一领域的知识。机器学习涵盖了众多的领域，如监督学习、无监督学习、深度学习等，每个领域都有其独特的理论和应用。在课程中，学生将有机会探索这些不同领域的广度和深度，了解它们的优缺点和适用范围。此外，课程中的小组作业和项目大作业，让我有机会将学到的理论知识应用到实际问题中，这种项目驱动的学习方式有助于加深对机器学习原理和算法的理解，同时培养我们的问题解决和创新思维能力。

**意见和建议：**

1. 提供丰富的学习资源帮助学生多渠道学习和掌握知识。提供交互式学习资源，如在线编程环境、实验室和模拟工具。这样的资源可以帮助学生直观地理解算法和技术，并在实践中获得反馈。通过实际项目驱动学习，设立从简单到复杂的项目任务，帮助学生逐步掌握机器学习的应用方法。

2.鼓励学生根据自己的兴趣和目标选择适合的学习路径。提供不同难度级别和专业领域的选修课程，以满足不同学生的需求。

3.引入前沿技术：

机器学习领域发展迅速，课程内容应定期更新和扩展，以反映最新的研究进展和技术趋势。可以在课程中引入一些前沿技术和最新研究成果，帮助学生了解该领域的发展趋势和未来方向，激发学生的学习兴趣和创新思维。

加强互动与讨论：

4.定期对学生的学习进度和理解情况进行评估，并提供及时反馈和建议。这有助于学生发现和解决问题，并及时调整学习策略。

总的来说，这门课程内容丰富，教学方法科学合理，激发了我对人工智能的兴趣。对我的学习和成长起到了很大的帮助作用。希望课程在未来能够不断改进和完善，为更多学生带来更好的学习体验。