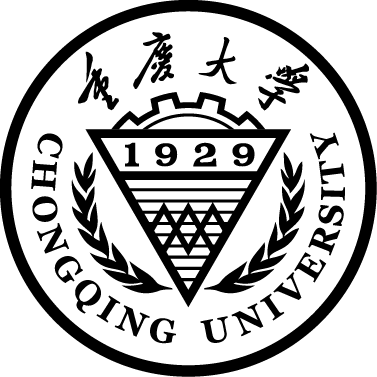
**   **

**大学生创新训练项目**

**申报书**

****

项目名称：基于主动学习的医学图像噪声检测和可视化研究

学　　院：大数据与软件学院

指导教师：葛永新

项目组成员：

教务处制

2023年 5 月 5 日

**填写说明**

1. 凡申报国家级大学生创新训练项目必须填写申报书。
2. 向学校报送本申报书时，一式3份，并报送申报书电子文档。
3. 本表填写内容必须与事实相符，表达准确，数字一律填写阿拉伯数字。
4. 打印格式：

（1）纸张为A4大小，双面打印；

（2）文中小标题为四号、仿宋、加黑；

（3）栏内正文为小四号、仿宋。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | | | | 基于主动学习的医学图像噪声检测和可视化研究 | | | | | | | | | | | | |
| 项目所属学科 | | | | 一级学科代码及名称 | | | | 二级学科代码及名称 | | | | | 三级学科代码及名称 | | | |
| 520 计算机科学技术 | | | | 52040计算机软件 | | | | | 5204070软件工程 | | | |
| 项目开展支撑平台 | | | | 智能服务与软件工程中心 | | | | | | | | | | | | |
| 项目组人数 | | | | 3 | | | 项目实施时间 | | | | |  | | | | |
| 项目所需经费 | | | | 5000元 | | | | | | | | | | | | |
| 项目组成员（含项目申请学生） | | | | | | | | | | | | | | | | |
| 姓 名 | | | 学 号 | | | 学院 | | | | | 年级专业班 | | | | | 联系电话 |
|  | | |  | | | 大数据与软件学院 | | | | |  | | | | |  |
|  | | |  | | | 计算机学院 | | | | |  | | | | |  |
|  | | |  | | | 计算机学院 | | | | |  | | | | |  |
| 主要指导教师 | | | | | | | | | | | | | | | | |
| 姓 名 | 工号 | | | | 职 称 | | | | | 学 院 | | | | | 联系电话 | |
| 葛永新 | 31681 | | | | 副教授 | | | | | 大数据与软件学院 | | | | | 189-8394-5891 | |
|  |  | | | |  | | | | |  | | | | |  | |
| 项目来源 | | | | | | | | | | | | | | | | |
| 项目来源 | | 请打“√”  1. 学生自拟（ ） 2.教师科研项目转化（ ）  3. 学生承担社会、企业委托项目等（ ）  4. 教师发布课题（ √ ） | | | | | | | | | | | | | | |
| 项目名称（限项目来源为2.3填写） | |  | | | | | | | 来源项目类别（限项目来源为2.3填写） | | | | |  | | |

**注**：1.“项目开展支撑平台”指支撑本项目开展的校、院级教学实验中心、科研实验室等，表中填写有关实验室名称，可以多个。

2.“来源项目类别”栏填写“863项目”、“973项目”、“国家自然科学基金项目”、“省级自然科学基金项目”、“教师横向科研项目”、“企业、社会委托项目”以及其他项目标识。

|  |  |
| --- | --- |
| 主要研  究内容  (限200字内) | 基于主动学习的医学图像噪声检测和可视化研究，研究内容由三部分组成：一是医学图像中噪声图像的检测和甄别，这一阶段的目标是在获取的带有噪声样本的医学图像数据集中提取出噪声样本，方法是结合  O2U-Net 和 Curriiculum-Classification 算法检测噪声样本，同时使用 Fine-sample 算法检测噪声样本，同时使用 KNN 对噪声样本集合进行扩充；二是深度学习中的主动学习，这一阶段将训练得到效果好的典型样本加入已标注样本进行迭代，降低标注成本，同时保证模型的性能；三是可视化及反馈模块，这一阶段的目标是将病灶等信息给医生和病人可视化展示，并且医生可提供反馈结果，方法是数据统计和类激活热图。 |
| 项目研究难点及创新点 | **一、本项目的难点**  本项目主要解决基于主动学习的医学图像噪声检测和可视化研究问题。这个问题可以分为以下3个困难点：   1. 如何确定样本是一个带错误标签的噪声样本，还是一个信息丰富的难样本。Curriiculum-Classification 算法能为每个训练样本打上难易程度的等级标签，但是不能够区分出训练样本是标记正确的难样本还是错误标记的噪声样本。O2U-Net 认为周期性训练后平均损失较大的样本是难样本，但无法区分出这个难样本究竟是错误标记的噪声样本还是正确标记的“难”样本。面对复杂的医学图像，仅利用一种噪声检测方法难以保证检测结果的完整性。所以难点在于如何结合上述算法以及其他算法精确的分离出噪声集。 2. 医学领域中，专家标注是昂贵且耗时的。同时，不同的专家面对同一份医学图像的标注也会存在差异。因此，如何利用现在十分流行的深度学习算法来帮助人们减少标注成本。   3）如何做到对信息的高效展示。当前医学噪声图像检测仅仅只是停留在提供一个分类结果的阶段，在信息展示方面以及与病人和专家信息交互方面存在着空缺，但是如果能够使得用户能够自主的上传数据,那么有益于后续模型的迭代训练并进一步改善模型的性能。且大量数据集更有益于深度学习模型学习肺癌图像特征并分类。同时如果使用可视化技术追踪模型训练的每个步骤和性能指标，追踪每个指标随着时间的变化趋势，还有训练过程中使用到的数据信息，就能辅助研究人员简单高效的进行模型训练，同时能够是患者对自身的疾病有着更为清晰的认识。因此难点在于如何设计一个高效地对信息进行可视化以及数据交互系统的方法。  **二、本项目的创新之处**  本项目的最大特点是提出了基于深度学习，主动学习的医学图像噪声检测和可视化解决方案，不仅能有效提升各个模块的准确率和速率，还有着极大的应用价值和意义。我们的主要创新点包括：  **1 噪声检测甄别模块**  使用O2U-Net 和 Curriiculum-Classification 算法检测噪声样本，在O2U-Net算法中创新性的通过循环调整学习率，同时监测损失值，损失值较大的样本数据便有可能是噪声样本或者难学习样本。配合使用 Curriiculum-Classification算法进一步确定数据集是否为噪声集，两者结果相结合确定噪声样本N1。我们同时使用 Fine-sample 算法检测，首先为每个类在噪声训练数据集中创建表示的 gram 矩 阵，并对这些 gram 矩阵进行特征分解。然后，Fine-sample 使用具有最大特征值的第一个特征向量之间的内积值的平方来查找干净且有噪声的实例。通过这种方式，如果数据与第一个特征向量对齐，那么可以说明数据是干净的，而大多数噪声样本则不是对齐的。取前 k%的样本为噪声样本N2。最后使用 KNN 对噪声样本集合进行扩充；图1.1展示了难样本与简单样本的区别：  图 1.1 难易样本展示  **2 主动学习模块**  设计了一个基于深度学习的主动学习的模块。具体来说，首先将得到的模型训练好的典型样本加入到训练样本集中用于模型训练。然后将新的数据集进行迭代，使得标注正确的数据集在原来的数据集上的占比逐渐增大，达到类似于专家标注的效果。  **3 可视化模块和用户交互模块**  对噪声数据进行可视化并且用户可以上传新的数据。数据可视化是进行数据分析的高效工具，也同样可以运用到展示结果当中。通过读取医学影像中噪声鉴别结果数据，将噪声样本，噪声预测概率矩阵，精度，聚类特征图等结果直观的展示出来，以辅助相关专家在没有提前了解病例的情况下对癌症图像的决策和分析。同时能够使使用Grad-CAM 热力图可以可视化网络模型的中间层激活结果和网络中提取特征的区域，对研究者理解、分析模型原理有着实际意义。用户交互模块分为数据导入、噪声查看与修改、以及癌症分类结果展示三个模块如图1.2所示：  图 1.2 可视化交互平台 |

**一、项目组成员分工**

|  |  |
| --- | --- |
| **姓名** | **主要项目研究内容** |
| 黄钰荐 |  |
| 安惠 |  |
| 李金阳 |  |
|  |  |
|  |  |
|  |  |

1. **立项背景和依据（包括研究目的、意义、国内外研究现状分析及评价）**

|  |
| --- |
| **1立项背景及研究意义：**  **1.1 研究背景及意义**  每年全球因癌症而死亡的人数数以百万计，其中前列腺癌给人们的生活带来了较大的危害。面对各不相同的癌症，其对应的治疗方式也大相径庭。因此，准确的诊断对于每位癌症患者选择和规划合理的治疗方案是十分重要的。近几年，人工智能(Artificial Intelligence，AI) 和深度学习技术已成功应用于医学领域，尤其在医学图像的识别和病理诊断方面起着显著作用。然而，在现实场景下，噪声医学图像广泛存在于数据集种且严重影响着模型的识别性能。因此，研究如何有效进行医学图像的噪声检测以提升模型的识别性能具有重大的现实意义。在过去几十年里，病理学家通过眼睛观察计算机断层扫描(CT)、核磁共振(MR)、X光、X射线、超声，正电子发射型(PET)等医疗诊断结果影像对疾病进行诊断。 然而，病理学家的稀缺和诊断流程的繁琐耗时加剧了临床诊断的需求和实际诊断工具低效的矛盾。为了解决这个矛盾，医学领域的研究者们开始把目光聚集在计算机辅助诊断中。随着科学技术的进步，计算机处理器(CPU)和图形处理器(GPU) 性能的提升，让计算机能够处理大量的复杂数据；深度神经网络的不断优化，计 算机能够提到取更深层次的特性；伴随着医学图形成像技术的提升，数字化全幻 灯片图像(WSI)技术问世。通过这些技术相辅相成，计算机视觉和深度学习技术在各种医学领域图形分析中卓有成效。此外，各种先进的算法也在不断完善，为深度学习算法能够在医学领域进行准确分析和完成各种实际场景难题时提供技术支持。  数据可视化是一门用图像或图形表达数据形式的科学技术。数据可视化主要以图表、图像的形式简洁且有效的表达数据信息，是数据分析领域中常用的一种分析手段[1]。在医学领域，可视化技术常用于将大量的数据集处理成可用于医疗实践的图像。用各具优势可视化图形，将复杂的医疗数据映射到图像当中，通过图像的表达让复杂的医疗数据变得简单易懂，能够做到让研究者能更好的观察数据 的分布以及变化，帮助他们提高分析和决策效率。  **1.2研究意义**  本项目设计了一个基于主动学习的医学图像噪声检测算法，并结合可视化技术辅助展示不同方式噪声监测对噪声的预测概率，同时专家可以通过系统进行噪声样本的查看与矫正。这样既能给使用者自主修改诊断结果的可能性，又能有效减少相关专家对病理的批阅和误诊的现象出现。简单来说，本项目的研究意义有如下几点：  **（1）噪声检测准确化**  前列腺癌对于人们的正常生活造成了巨大的影响，同时由于癌症之间的差别与人体器官之间的界限较小，所以计算机能在繁多的医学图像中找到前列腺癌是一项严峻的挑战。实验数据来自PESO公共数据集中的前列腺数据集和Hubmap的人体器官功能组织单位集（FTU）。在图像预处理时用无覆盖的方式去切割WSI，根据每个切割的中癌症的占比来分标签，然后分已标签和未标签并添加噪声。具体流程如图1.3所示：  图 1.3 WSI\_process  Curriculum-Classification 算法（简写 CC）是一种有效的阶段性训练策略，它让深度学习模型从较容易的数据开始学习，后加入稍难的样本对深度学习模型进行训练。具体来说，训练集被划分成多个子集，这些子集包含信息丰富的难样本图像，简单可靠的样本图像和带错误标签的噪声样本[2]。针对这些子集进行简单到复杂的排序，可以对每个样本赋予等级标签，具体操作如图1.4所示。O2U-Net（overfitting to underfitting）[3]的原理来源于常见的深度神经网络的训练过程。简单的样本在早期对梯度计算的贡献更大，因此它们的损失急剧下降。相反，通常难样本是在训练后期学习到的。如图1.5。最后将两者相互结合得到准确的甄别结果，如表1.1所示。另外再结合Fine-sample算法得到最终的结果。    图 1.4 curriiculum-Classification 算法流程    图 1.5 O2U算法周期性训练    表 1.1 O2U 和 CC 算法结果  **（2）主动学习策略设计**  深度学习算法在图像分类任务中表现出了卓越的性能，但模型的效果严重依赖于专家标注样本的质量和数量。医学领域中，专家标注是昂贵且耗时的。因此，如何能在不降低模型性能的前提下，显著地减少样本的标注成本是一个需要被解决的问题。主动学习是机器学习中的一个领域[4]，目的是尽可能的减少标注样本数量且保证模型的性能。常见将训练得到效果好的典型样本加入已标注样本进行模型训练，这样可以很大程度地减少专家标注的样本数量，降低标注成本，同时保证模型的性能。在本项目中，我们分别将两个数据集迭代了20次。其结果如图1.6和图1.7所示：    图 1.6 Hubmap数据迭代20次  图 1.7 PESO前列腺癌数据迭代20次  **（3）分析结果可视化**  该项目利用可视化技术，将噪声特征提取，将其可视化，可以直观地展示原图，在该项目中，我们一方面方便医生的直接诊断，另一方面方便患者自行标注，减轻标注图片的难度，有利于扩充数据集，为以后的医学噪声图像鉴别提供训练材料。如图1.8所示：  图 1.8 分析结果可视化  我们整个项目研究流程图如图1.9所示：    图 1.9 项目流程图  **2国内外研究现状**  **2.1 医学图像噪声集甄别研究现状**  在计算机视觉领域，基于深度学习的方法在众多任务中取得了令人惊叹的性 能。但面对存在大量噪声样本的医学图像数据集，现有的大多数模型易受到噪声 样本的影响从而呈现较差的识别性能。为进一步促进深度学习在医学图像领域的 应用，对于医学图像数据集的噪声检测是有必要的。目前，研究者在一般图像和 医学图像上进行了许多噪声检测的尝试性的研究工作。大体上，针对训练集的噪 声样本检测可分为两类：1）设计噪声检测算法进行噪声样本检测，然后对判定为 噪声的样本去除或矫正标签，以减少噪声样本对模型训练的影响 2) 直接在不干净 的数据集上训练噪声鲁棒的模型。 第一种噪声清理的解决方案：首先检测到潜在的噪声标签样本，然后将这些 标签样本从训练集中删除。Lee 等人[5]提出 CleanNet 的联合神经嵌入网络，从带 有噪声标签的图像中提取特征向量。使用 CleanNet 后的每个类代表的特征向量是从一个小的干净数据集中计算出来的。用这些特征向量之间的相似度用来决定标 签是否正确。或者，用这种相似性为训练样本分配权重。这种方法总结了一部分 需要人工验证的噪声标签知识，并使用迁移学习处理噪声标签。Han 等人[6]提出 了一种噪声标签检测方法，名为 Co-teaching，其中同时训练两个深度神经网络。 每个网络选择另一个网络用于训练的样本。每一个网络都相互选择识别噪声标签。 最重要的是，它们通过在两个网络迭代在框架中估计了正确的标签，消除了模型 对干净数据集的需求。此外，它们使用多个原型（而不是在 CleanNet 中只有一个） 来表示每个类。Guo 等人[2]提出了 Curriculum-Net，在 Curriculum-Net 中通过将训 练的数据的分布密度作为衡量复杂性的标准，然后将复杂度按从小到大的顺序进 行排序，将训练数据划分为简单到复杂的子集。这些子集是作为一种课程学习模 型训练，用于逐步让模型从简单到困难理解标签噪声学习。Huang 等人[3]提出了O2U-Net 网络，让网络不断自动从过度拟合状态到欠拟合状态之间变化。在这个网络状态中通过监测网络训练的损失变化，启发性的发现噪声样本的损失变化较大，因此他们检测出噪声标签并删除。Northcutt 等人[7]使用置信标签识别数据点，并仅使用这些数据点更新分类器。但是该方法基于这样一个假设，即预测概率接近1 的数据样本更有可能具有正确的标签。还提出了从 one-hot 标签中获得这种 标签分布的方法，用于多种应用，包括多分类和语义分割。Kim 等人[8]提出一种 噪声检测器 Fine-sample来过滤噪声标签实例，使用无导数的简单方法提供一个鲁 棒检测器。 第二种抗噪解决方案通常侧重于引入正则化或以减少过度拟合对噪声标签的影响。Xiao等人[9]提出了一个概率模型用来描述图像、真实标签、噪声标签和噪声类型之间的关系，用概率模型验证一部分的干净的标签。Reed 和 Lee 等人[10] 提出了与噪声标签建模的概念，在该方法中样本重建误差被用作估计噪声分布的一致性。Ghosh 等人[11]表明，平均绝对误差（MAE）可能对噪声标签具有鲁棒性。Zhang和Sabuncu等人[12]认为，MAE在使用DNN时表现不佳，并提出了一个GCE损失函数，可以将其视为 MAE 和交叉熵（CE）的推广。Wang 等人[13]引入了交叉 熵项（RCE）的反向版本，并提出SCE损失函数是CE和RCE的加权和。一些研 究表明，早期停止模型可以防止噪声标签的记忆现象[14-15]。基于这一假设，Liu 等人[16]提出了一种早期学习正则化（ELR）损失函数，通过利用半监督学习技术禁止记忆噪声数据。Xia等人[17]阐明了哪些神经网络参数会记忆噪声标签，并针对这些噪声标签参数提出了一种稳健的训练策略。该方法通过一个平滑热向量[18]，在 数据实例之间使用线性插值[19]，以及提取其他模型噪声标签的重新缩放预测[20]，在预测层面上将结果正则化。但是，随着数据集噪声率的增加，这个方法在性能方面存在局限性。  **2.2 深度学习在肺癌医学图像的研究现状**  肺癌是世界上许多国家癌症相关死亡的主要原因之一。目前，主要基于组织 病理学显微镜呈像结果决定最佳的治疗方案。以往病理学家用眼睛观察光学显微 镜下的结果进行诊断和确定癌症的亚型，然而，病理学家的稀缺和诊断流程的耗 时加剧了临床诊断需求高和实际医学诊断工具低效的冲突。此外，不同观察者的 理论知识和评判标准存在差异，这会对诊断造成偏差并引入额外风险。随着科学 技术的发展，出现了数字化全幻灯片图形(WSI）技术[21]。随着数字化全幻灯片图形 技术的问世，见证了医学图像从传统人工诊断到如今计算机诊断的转变。随着数字化全幻灯片图形的逐步使用，以及计算机视觉和深度学习技术在各种医学领域 图像分析方面取得的显著成功，计算机在解决复杂的现实问题方面表现出强大的 计算能力和图像分析能力。先进的深度学习算法可以增强组织病理学图像分析能 力，能广泛的用于医学领域挑战性的诊断任务。  Yang 等人[22]提出一种基于深度学习的六分类器，以监督或弱监督方式对肺腺 癌、肺鳞状细胞癌、小细胞肺癌、肺结核、机化性肺炎和正常肺的 WSI 图像分类。 该方法使用在来自不同医疗中心的四个队伍的 1067 张幻灯片数据进行测试，AUC 指标分别达到了 0.970、0.918、0.963 和 0.978，分类器性能高于组织病理学人工 诊断。Kanavati 等人[23]训练了一个基于 EfficientNet-B3 架构的卷积神经网络(CNN)， 使用迁移学习和弱监督学习两个方式，对 3,554 个幻灯片图像 WSI 的训练数据集 预测所有幻灯片图像(WSI)中的癌症，结果在四个独立的测试集上的曲线下面积 （AUC）较高，在区分肺癌和非肿瘤性肺癌方面获得了非常理想的结果。Wang 等人[24]提出了一种弱监督方法，能够对整个肺癌幻灯片图像进行快速有效的分类。利 用基于图像块的全卷积网络（FCN）进行检索，高效地提供具有判别性的图像特征。然后，探索不同的上下文感知块选择特征聚合的策略，生成全局的 WSI 描述符，最终将其结果输入随机森林(RF)分类器进行图像分类预测。在大规模的肺癌 WSI 数据集进行评估，验证了方法的有效性和可行性。 以上研究显示，深度学习在分析组织病理图像方面具有多重优势，包括特征 定义的简化、对复杂对象的识别能力强、计算节省时间。因此，将深度学习方法 应用于肺癌病理图像分析的研究迅速增加。针对相对简单的任务，如肿瘤检测和组织学亚型分类，结果普遍令人满意。  **2.3 医学图像可视化研究现状**  虽然深度学习方法在组织病理学的应用中有着优越的性能，但是这些方法缺 乏解释性。以肺癌图像分类为例，深度学习模型只需以肺癌图像为输入就能直接 输出图像的分类结果。因此，基于组织病理学在深度学习推理和决策的可视化方 面研究开始引起人们的关注。一些专家在研究医学图像方面将深度学习和可视化 技术结合起来，提供可解释的展示，帮助使用者了解和信任深度学习技术，并且 将深度学习辅助技术用于病理诊断、预测和决策。 Faust 等人[25]在结肠癌数据方面利用随机邻域嵌入(t-SNE)来降低维度，然后在 softmax 转换之前来描述 CNN 的高维组织学数据结构的平面信息。使用 t-SNE 图，显示了 CNN 最终隐藏层中的 13 个经过训练的组织类内部的高维组织平面。 绘制了来自每个类的 350-600 个训练图块，t-SNE 内的每个点代表一个 1024 × 1024 像素的训练图像。用于在组织病理学中定量可视化分析和减少常规错误发生，通过可视化为相关病理学家提供预测结果和决策模型。Wang 等人[26]提出了一个基于 深度学习的自动化框架 HEAL，实现在多类别分类（例如癌症分型）和多标签分类 （例如基因突变识别）模式之间自动切换。此外，Wang 等人还对模型进行可视化， 并对 CONC-Heatmap 使用 Grad-CAM 和 Feature-map 两种方式对预测结果进行可视 化图形分析。HEAL 在两个肺癌案例研究和一个结肠癌案例研究展示出实用性和 功能性，可以进行复杂的组织病理学分析，并将深度学习应用到癌症图像分析中。 Xu等人[27]提出一种简单、高效组织病理学图像自动分析系统。该方法利用深度卷 积神经网络 (CNN) 激活特征使其在大规模组织病理学图像中执行分类、分割和可视化，并在MICCAI2014 脑肿瘤数字病理学挑战赛的脑肿瘤数据集和结肠癌组织病理学图像数据集上显示出最先进的性能。Jamieson 等人[28]使用三种不同的成像形 式的数据：超声图像 1126 例，磁共振图像 356 例，乳房 X 射线全视野数字影相 245 例，探索无监督非线性降维(DR)的可视化表示形式，即使用拉普拉斯特征图 表展示降维数据和特征数据。将初始的高维特征空间映射到人类可解释的低维特 征空间，同时保留数据局部信息和全局信息，可视化的展示有着复杂数据结构的特征数据。以上研究表明，将人工智能和可视化技术结合，可以帮助研究者对深度学习模型学习进行推理和决策。结合可视化的方式为组织病理学常规组织形态分析提供了视觉直观、客观和“分级”性能指标的展示。  **参考文献：**   1. 刘春鹤,张晗,惠文,李艳君,赵玉虹.国内外医疗数据可视化研究的现状分析与展望[J].世界科技研究与发展,2021,43(03):312-330.DOI:10.16507/j.issn.1006-6055.2020.12.026.   [2] Guo S, Huang W, Zhang H, et al. Curriculumnet: Weakly supervised learning from large-scale web images[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 135-150.  [3] Huang J, Qu L, Jia R, et al. O2u-net: A simple noisy label detection approach for deep neural networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3326-3334.  [4] Felder R M, Brent R. Active learning: An introduction[J]. ASQ higher education brief, 2009, 2(4): 1-5.  [5] Lee K H, He X, Zhang L, et al. Cleannet: Transfer learning for scalable image classifier training with label noise[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5447-5456.  [6] Han B, Yao Q, Yu X, et al. Robust training of deep neural networks with extremely noisy labels[C]//Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS). 2020, 2: 4.  [7] Northcutt C G, Wu T, Chuang I L. Learning with confident examples: Rank pruning for robust classification with noisy labels[J]. arXiv preprint arXiv:1705.01936, 2017.  [8] Kim T, Ko J, Cho S, et al. FINE Samples for Learning with Noisy Labels[J]. arXiv e-prints, 2021: arXiv: 2102.11628.  [9] Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2691-2699.  [10] Reed S, Lee H, Anguelov D, et al. Training deep neural networks on noisy labels with bootstrapping[J]. arXiv preprint arXiv:1412.6596, 2014.  [11] Ghosh A, Kumar H, Sastry P S. Robust loss functions under label noise for deep neural networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1). [12] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels[J]. Advances in neural information processing systems, 2018, 31.  [13] Wang Y, Ma X, Chen Z, et al. Symmetric cross entropy for robust learning with noisy labels[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 322-330.  [14] Arpit D, Jastrzębski S, Ballas N, et al. A closer look at memorization in deep networks[C]//International conference on machine learning. PMLR, 2017: 233-242.  [15] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107-115.  [16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.  [16] Liu S , Niles-Weed J , Razavian N , et al. Early-Learning Regularization Prevents Memorization of Noisy Labels[J]. 2020.  [17] Xia X, Liu T, Han B, et al. Robust early-learning: Hindering the memorization of noisy labels[C]//International conference on learning representations. 2020.  [18] Lukasik M, Bhojanapalli S, Menon A, et al. Does label smoothing mitigate label noise?[C]//International Conference on Machine Learning. PMLR, 2020: 6448-6458.  [19] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.  [20] Kim T, Oh J, Kim N Y, et al. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation[J]. arXiv preprint arXiv:2105.08919, 2021.  [21]Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides[J]. Medical Image Analysis, 2021, 70: 102032.  [22] Yang H, Chen L, Cheng Z, et al. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study[J]. BMC medicine, 2021, 19(1): 1-14.  [23] Kanavati F, Toyokawa G, Momosaki S, et al. Weakly-supervised learning for lung carcinoma classification using deep learning[J]. Scientific reports, 2020, 10(1): 1-11. Khan R U, Zhang X, Kumar R. Analysis of ResNet and GoogleNet models for malware detection[J]. Journal of Computer Virology and Hacking Techniques, 2019, 15(1): 29-37.  [24] Wang X, Chen H, Gan C, et al. Weakly supervised deep learning for whole slide lung cancer image analysis[J]. IEEE transactions on cybernetics, 2019, 50(9): 3950-3962.  [25] Faust K, Xie Q, Han D, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction[J]. BMC bioinformatics, 2018, 19(1): 1-15.  [26] Wang Y, Coudray N, Zhao Y, et al. HEAL: an automated deep learning framework for cancer histopathology image analysis[J]. Bioinformatics, 2021, 37(22): 4291-4295.  [27] Xu Y, Jia Z, Wang L B, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features[J]. BMC bioinformatics, 2017, 18(1): 1-17.  [28] Jamieson A R, Giger M L, Drukker K, et al. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and ‐ SNE[J]. Medical physics, 2010, 37(1): 339-351. |

**三、主要研究内容和目标（包括研究方案和技术路线）**

|  |
| --- |
| **1 研究目标和研究内容**  **1.1 研究目标**  基于主动学习的医学图像噪声检测和可视化研究，研究目标由三部分组成：一是实现基于O2U-Net和Curriiculum-Classification的噪声检测对训练集样本进行划分，对每个样本赋予等级标签，能初步区分出简单、难样本和检测噪声；二是实现基于Fine-Sample的噪声检测和KNN噪声集扩充，利用无监督学习进行特征分解从而分离干净样本和噪声。KNN实现噪声样本的邻域扩充，减少噪声样本在模型训练时出现的遗漏；三是实现肺癌图像噪声检测可视化平台设计以及实现，癌症图像噪声监测可视化平台主要分为用户交互模块和可视化展示模块。用户交互模块功能包括导入数据、修改噪声标签和查看分类结果。可视化平台可视化了噪声样本预测概率矩阵，热力图，散点图，特征聚类图和精度和准确率折线图，便于研究者对数据进行分析。  **1.2 研究内容**  **1.2.1 基于O2U-Net和Curriiculum-Classification的噪声检测**  在癌症图像噪声检测问题中，需要对样本图像进行划分。我们希望通过对每个样本赋予等级标签初步区分出简单、难样本和检测噪声，提高训练过程的效率。我们的模型通过结合O2U-Net和Curriiculum-Classification算法的思想来实现：Curriiculum-Classification从较容易的数据开始学习，将ResNet50 作为骨架网络来实现图像特征的提取，并利用PCA降维处理，k-means 算法用于特征上的聚类分组将样本分类；O2U-Net算法根据循环学习率对网络进行周期性训练并监控每个样本损失变化，从而区分出噪声样本。  **1.2.2 基于Fine-Sample的噪声检测和KNN噪声集扩充**  对于复杂的医学图像，仅利用一种噪声检测方法难以保证结果的完整性与准确性，  故Fine-sample被引入用于更好提升模型最终的检测性能。Fine-sample利用无监督学习算法实现特征分解，以产生潜在的特征表示成分。通过分析这些成分，可以将干净数据和噪声数据进行分离，能够高效地检测出噪声样本。为进一步扩充噪声样本集合减少噪声样本在模型训练时出现的遗漏，采用KNN算法对噪声样本集合进行扩充，将Fine-sample方法与KNN噪声集扩充相结合能够很大程度提高样本预测的准确率。  **1.2.3 癌症图像噪声检测可视化平台设计以及实现**  癌症噪声检测可视化平台能够辅助研究者进行决策并提高决策效率，同时通过对模型训练效果的展示来提高使用者对于深度学习模型诊断结果的信任。  癌症图像噪声监测可视化平台的页面使用React框架和Ant-Design UI库实现布局。可视化部分使用D3.js自定义实现可视化图形，后端使用flask进行数据传递。数据处理过程主要为：将模型预测结果保存为pkl文件，并通过flask，用接口的形式将数据传送到服务器。可视化部分则是通过相应的接口对服务器上数据进行获取而后进行可视化显示。同时，用户对数据的上传与修改可通过可视化平台的网页接口提交数据到服务器，然后通过flask的更改算法部分的pkl文件，完成数据的交互过程。  **2 拟采取的研究方法**   * 1. **基于主动学习的癌症图像噪声检测**   主要采用了O2U-Net算法,Curriiculum-Classification算法，Fine-Sample算法和KNN算法对噪声图像进行筛选，具体介绍如下。   * + 1. **O2U-Net算法**   O2U-Net算法，即overfitting to underfitting，在训练过程中通过不断循环调整学习率，监测损失值来区分较噪声样本和正常样本。具体来说，对于一个等待训练的包含噪声集的样本，损失值首先会下降较大，因为机器会优先学习较为简单，也就是较为纯净的样本集，等到损失值下降不明显的时候，学习率会调到较大的值，使样本跳出过拟合转态，开始下一轮学习。在噪声样本或者难样本较多的训练集中，损失值相对于其他的数据集较大。对于同一个训练 epoch，从每个样本的 loss 中减去这个 epoch 中所有样本的平均损失来归一化不同 epoch 的损失。在循环训练过程中，假设最大 循环学习率为 r1，最小学习率是 r2，其中 r1>r2。最后，采用线性递减函数来循环 调整学习率。循环训练时的学习率调整公式如下：    其中 t 是指循环训练中的第t个epoch，c是每个循环轮中 epoch 的总数，r(t) 是 在 t 处的学习率。整个循环训练之后，计算每个样本的归一化损失的平均值。然后 按照所有样本平均损失值进行由大到小的排列，取前 k % 的样本作为噪声标签。最后，重新初始化网络的参数，在去噪后的数据集上对模型进行重新训练，直至损失下降不明显。    图 3.5 周期性训练     * + 1. **Curriiculum-Classification算法**   Curriculum-Classification 算法（简写 CC）是一种有效的阶段性训练策略，它 仿照人类的学习过程，让深度学习模型从较容易的数据开始学习，后加入稍难的 样本对深度学习模型进行训练。简而言之，课程学习算法根据训练集样本的复杂 性将它们分为“简单”、“中等”和“困难”。具体来说，训练集被划分成多个 子集，这些子集包含信息丰富的难样本图像，简单可靠的样本图像和带错误标签 的噪声样本[19]。针对这些子集进行简单到复杂的排序，可以对每个样本赋予等级标签。  首先使用ResNet50提取图像特征,ResNet50是一种经典的卷积神经网络模型，通常用于图像分类和特征提取任务。要使用ResNet50提取图像特征，可以按照以下步骤进行：  a.导入ResNet50模型  首先，需要导入ResNet50模型，这可以通过使用PyTorch中提供的预训练好的ResNet50来导入。  b.加载图像并进行预处理  接下来，需要加载图像并进行预处理，以使其与ResNet50模型的输入格式相匹配。这可以通过使用Keras的图像处理工具ImageDataGenerator和相关函数完成。  最后，可以使用ResNet50模型对加载的图像进行特征提取。  在提取完图像的特征后，使用PCA进行降维处理：它是一种常用的数据降维技术，它可以将高维数据转换为低维数据，从而减少数据的冗余性和复杂性。基本思想是将原始数据通过线性变换，转化为新的一组正交的坐标轴，称为主成分。主成分的数量可以选择，一般选取其中包含大部分数据方差的前几个主成分作为新的坐标轴，其余的主成分被丢弃。然后对数据进行标准化，将所有特征都缩放到相同的尺度上。接着计算数据的协方差矩阵。再对协方差矩阵进行特征值分解，得到特征值和特征向量。最后选取前k个特征值对应的特征向量作为新的坐标轴，将原始数据投影到新的坐标系中得到降维后的数据。  在PCA降维完成后，用k-means 算法用于特征上的聚类分组。根据特征间的距离关系得到不同的分组结果，每个组内的样本具有相似的特征表示。为计算每组内图像特征间的距离，求得欧几里得距离矩阵 D ∈ ||f(pi)-f(pj)||^2,n 指当前组内的图像数量，pi,pj为第 i 张和第 j 张图像样本，f(pi),f(pj)分别表示图像pi和pj的特征向量，Di,j表示pi和pj间的距离。通过设置特定阈值dc，可以得到每张图像的局部密度ρi，其具体的数学公式表示如下：    X(Di,j-dc) 表示在同一类别下样本 i 到其他样本之间的距离。如果距离小于dc， 那么说明这个样本与另一个样本是相似的，记为 1。反之记为 0。用数学公式表示 如下：    对每个样本 X(d)的数量进行累加。对于具有正确标签的图像来说，如果他们的外 观是相似的，那么他们对应的投影的距离应该比较相近，这就会产生一个较大的 局部密度；反之对于噪声图像，他们对应的投影距离比较远，他们的局部密度就 比较小。最后，针对每个样本局部密度ρi，用 k-means 算法进行聚类。根据数据点 到聚类中心的距离进行分类，将样本分成“简单”、“中等”和“困难”。  **2.1.3基于Fine-Sample的噪声检测和KNN噪声集扩充**  Fine-sample 首先为每个类在噪声训练数据集中创建表示的 gram 矩阵，并对这些 gram 矩阵进行特征分解。Gram矩阵（Gram Matrix）是一个由向量内积构成的矩阵，通常用于计算特征映射的相似度或相关性。在机器学习和计算机视觉领域中，Gram矩阵被广泛应用于图像风格迁移和神经风格化等任务中。假设有一个由n个向量x\_1, x\_2, ..., x\_n组成的向量集合，其中每个向量的维度为d。则该向量集合的Gram矩阵定义为：  G\_{ij} = x\_i^Tx\_j  其中，i和j是向量的索引。可以看出，Gram矩阵是一个对称矩阵，且对角线上的元素是每个向量的模长的平方。在深度学习中，Gram矩阵常用于计算神经网络中卷积层的特征映射之间的相似度。给定一个卷积层的输出，可以通过将输出的每个特征映射的像素展开成一个向量，然后计算这些向量的Gram矩阵来获得特征映射之间的相似度。  然后，Fine-sample 使用具有最大特征值的 第一个特征向量之间的内积值的平方来查找干净且有噪声的实例。通过这种方式，如果数据与第一个特征向量对齐，那么可以说明数据是干净的，而大多数噪声样 本则不是对齐的。即设 x为标记为 k 类的数据，z为输入的 x 的特征提取器输出，以及数据集 D 中记为 k 类的所有特征的 gram 矩阵∑k,公式如下:    然后，通过 <u1,z>^2 来计算 x 的对齐度，其中u1是uk特征分级的第一行，Uk通过∑k特征分解，结果公式如下：    将得到特征值按从大到小顺序排列。对于数据 D 中标记为 k 类的所有特征，在其 对齐分布上拟合高斯混合模型，高斯混合模型（Gaussian Mixture Model，GMM）是一种常见的概率模型，用于描述由多个高斯分布组成的混合分布。GMM是一种非监督学习方法，通常用于聚类分析和密度估计。  假设有一组观测数据x\_1, x\_2, ..., x\_N，其中每个观测数据x\_i是一个d维向量。GMM假设这些观测数据是由k个高斯分布组成的混合分布生成的，每个高斯分布都有自己的均值向量和协方差矩阵。因此，GMM的目标是从观测数据中估计出这k个高斯分布的参数，即每个高斯分布的均值向量和协方差矩阵，以及每个高斯分布在混合分布中的权重。  GMM的概率密度函数可以表示为：    其中，是第i个高斯分布在混合分布中的权重，是第i个高斯分布的均值向量，是第i个高斯分布的协方差矩阵，是均值为，协方差矩阵为的d维高斯分布的概率密度函数。  将样本分为干净样本集和噪声集。均值较大的集 合被视为干净集，则另一个集合是噪声集。其中干净的数据（蓝色点）主要与主 成分（黑色虚线）对齐，而嘈杂的数据（橙色点）则不对齐。因此，通过将数据 集拟合到高斯混合模型（GMM）中，数据集可以很好地聚类，表示与主成分对齐， 如图 3.7 所示，通过聚类将噪声样本和干净样本分离。这里，噪声数据检测器的质 量通过 F 分数进行衡量.  F分数（F Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/F1%E5%88%86%E6%95%B0/_blank)。F1分数可以看作是模型精确率和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/F1%E5%88%86%E6%95%B0/_blank)的一种调和平均，它的最大值是1，最小值是0。    F 分数越高的说明是更干净样本，分数低的说明是噪声样本。我们将 F 分数 归一化处理为 0 到 1 的值，归一化后按分数值从小到大进行排序，取前 k%的样本为噪声样本。    **2.2肺癌图像噪声检测可视化平台设计以及实现**  数据分析在我们的日常生活中发挥着越来越重要的作用，其中数据可视化是数据分析的一个重要组成部分，利用图表、图形等可视化元素，让人们能够可视化数据并进一步观察数据特征与变化。通过该平台的可视化模块，该平台提供了一种人机交互的方式，对于训练后得到的噪声集交给相关医学专业人员对噪声样本进行查阅和矫正。  **2.2.1用户交互模块及可视化平台开发**  肺癌图像噪声监测可视化平台的页面部分使用React框架和Ant-Design UI库 实现布局。可视化部分使用D3自定义实现可视化图形，后端使用 Flask 进行数据传递。用户交互模块数据导入、噪声查看与修改、以及癌症分类结果展示三个模块。  D3（Data-Driven Documents）是一个非常受欢迎的可视化开源库，D3图库的内容种类丰富，包含树图、分层边缘捆绑图、密度图、高线图等等，此外，D3的动画和交互功能十分强大。D3整合HTML、SVG和CSS等功能。D3还将强大的可视化技术和交互技术与数据驱动DOM操作方法相结合，可以编写简单浏览器页面以实现全部的功能  React是Facebook在2013年开源在GitHub上的JavaScript库。React用于构建用户界面，它把用户界面抽象成一个组件。使用者通过组合这些简单的组件，来实现丰富的功能和可交互的页面。如图 2.11 所示，React通过引入JavaScript的 eXtension(JSX)来操控虚拟文档对象模型(DOM)，用虚拟DOM来对真实DOM进行更新，这样避免直接操控DOM从而提升性能。  Screenshot 2023-03-15 194406  图15 React的DOM更新操作  目前肺癌图像噪声监测可视化平台的核心功能界面已经使用React初步开发，如图16所示。    图 16 可视化平台界面  **2.2.2可视化模块**  可视化模块是本平台最重要的部分，其利用可视化技术展示数据分布特征和重要的数据变化趋势，以辅助相关专家在没有提前了解病例的情况下对肺癌图像的决策和分析，提高了决策的效率和准确度。可视化模块分为四部分：噪声预测概率矩阵、Grad-CAM热力图、散点图和聚类特征图和算法性能展示折线图。  噪声预测概率矩阵展示不同算法处理噪声样本的结果并进行对比，以使研究者比较直观地观测到不同噪声检测算法对于噪声样本的预测概率以及它们之间的区别，如图18所示。    图18 噪声预测概率混淆矩阵  CAM全称为Class Activation Mapping，即类别激活映射图，也被称为类别热力图。热力图是一种通过对色块着色来显示数据的统计图表。可视化网络模型的中间层激活结果可以作为医疗专业人员判断病情的重要依据。得到的类别激活热力图如图19所示：    图19 热力图结果展示  散点图常用于多维数据进行可视化展示。在散点图上能查看数据的大概排列方式，但若是数 据量过大，用户也不能直观地从数据中获取信息。而通过将散点图和聚类结合的方式，用户就可以非常直观地了解多维数据中可能存在的特征关系及聚类情况。  折线图的使用可以直观展示不同噪声对模型训练的影响和不同比例的噪声对模型训练的影响，例如不同比例噪声的模型准确率可以通过如图20的折线图展示：    图20 不同比例噪声的模型准确率折线图 |

**注：如为重庆大学大学生科研训练计划（SRTP）立项项目需写明在原项目的基础上进一步深入研究的内容。**

**四、研究计划和进度**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | 研究计划和进度 | | | | 工作阶段 | 时间 | 具体研究计划 | | 第一阶段  （准备工作阶段） | 2022.6-2022.7 | 1. 准备好项目所需硬件设备：GPU服务器。 2. 深度调研项目需求，完善项目功能。 3. 制定一套详细地实施方案，包括：模块划分、功能实现、功能检测等。 | | 第二阶段  （功能实现阶段） | 2022.7-2023.4 | 1.实现基于孪生卷积神经网络模型的冗余图像检测系统。  2.实现基于小样本的胶囊内镜图像分类系统，研究迁移学习和元学习相关算法。  3.将整套系统组合并上传至服务器，编写调用API接口，完成系统封装。  4.前端界面设计，编写前端代码，建立人机交互友好的界面，展示项目内容。 | | 第三阶段  （测试总结阶段） | 2022.4-2023.6 | 1. 测试：对项目整体功能做测试，检测基本功能是否实现；对各模块做相应测试，检查模块内部、模块耦合漏洞。 2. 总结：编写项目结题报告，完善项目相关文档，对项目作总结分析。 | |
| **五、预期提供的成果及形式** |
| 1.通过本项目的研究，能够构建基于小样本学习的无线胶囊内镜分类模型，实现在标注数据较少的情况下对胶囊内镜图像的分类；同时提供冗余检测和数据可视化，做到对冗余图像的筛查、删除和对分类结果的可视化分析；最终建立一个基于小样本的无线胶囊内镜的分类系统。  2.申请国家发明专利1项。  3.上线无线胶囊内镜软件1款，并申请软件著作权。  4.发表EI收录论文1篇。 |

**六、项目研究支撑条件**

|  |
| --- |
| **工作环境：**  依托智能服务与软件工程中心，拥有自己的GPU计算服务器及软件开发实验室。  **主要实验设备：**  计算机、GPU服务器。  **项目组成员技术经验：**  C、C++、JAVA、python、数据库及及SRTP微信小程序项目开发经验，有良好的编程技术基础和大量前期总结的算法数据、研究成果。  **现有数据集：**  目前已有mini-imagenet数据集以及一定数量无线胶囊内镜图像用于训练模型。 |

**七、项目经费概算**

|  |  |  |
| --- | --- | --- |
| 序号 | 开支内容 | 金额（单位：元） |
| 1 | 资料费（包括书籍购买、资料复印等） | 1500 |
| 2 | 差旅费 | 1500 |
| 3 | 耗材（硬盘等硬件设施） | 2000 |
| 合计 | | 5000 |

**八、其他**

|  |  |  |
| --- | --- | --- |
| 是否重庆大学大学生科研训练计划（SRTP）立项项目 | | 是□， 否 ☑ |
| 经费配套说明 |  | |

**九、评审、审批意见**

|  |
| --- |
| 学院推荐意见：  主管院长签字： （公 章）  年 月 日 |
| 校评审专家组评审意见：  专家组组长签字：  年 月 日 |
| 学校意见：  (公 章)  年 月 日 |