**独创性声明**

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研 究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他 人己经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构 的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均 己在论文中作了明确的说明并表示了谢意。

签 名：

日 期： 年 月 日

**关于论文使用授权的说明**

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权 保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部 分内容，可以采用影印、缩印或其他复制手段保存论文。

(保密的论文在解密后应遵守此规定）

签 名： 日 期： 年 月 日

导师签名： 日 期： 年 月 日

# 摘 要

健康是人类终生的话题，一个人从孩童到暮年的各个阶段都伴随着医疗资源的消耗。截至2021年5月11日，[第七次全国人口普查](https://baike.baidu.com/item/%E7%AC%AC%E4%B8%83%E6%AC%A1%E5%85%A8%E5%9B%BD%E4%BA%BA%E5%8F%A3%E6%99%AE%E6%9F%A5/23542107?fromModule=lemma_inlink)结果公布显示全国人口总人口为144350万人，同年数据显示，我国全国医疗卫生机构数约102.3万个，床位约944.8万张，医疗人员约1066万人。在人口数量与医疗资源的矛盾日益凸显的背景下，人工智能技术以其高速有效的分析决策能力全面赋能医疗行业的各个环节，提升了医疗机构的诊断水平和效率，促进了医疗资源的合理配置。

在儿童先天性肾盂输尿管连接部梗阻性积水（Ureteropelvic Junction Obstruction Connection, UPJO）领域，B超是最常用的筛查手段，且主要治疗手段仍以手术为主，诊断阶段通过观测超声图像中肾盂经线、肾盏扩张程度、肾实质厚度、皮质回声等指标判断病理等级（SFU 0-4），这一阶段需要经验丰富的影像科医师通过细致的检测做出判断从而确定手术时机。而目前该疾病的首选高效诊断技术在国内发展仍很落后，国内分布的相关医学影像人才也不平衡、不充分，因此在诊断阶段需投入大量的医疗时间成本和人力资源成本，从而需要计算机图像处理技术来降本增效。传统的计算机图像处理方对超声图像的处理准确度不高，而超声图像本身为灰度图像，具备边界模糊性、视角多样性，且各病理等级对应的数据集不平衡，对本文所研究的智能诊断分类方法提出了一定的挑战。为解决上述问题，本文研究基于深度学习的计算机视觉技术，提出一种基于肾脏超声图像的智能诊断方案，具体所做工作如下：

1. 构建基于肾脏超声图像的语义分割模型，针对超声图像的背景区域和器官区域边界模糊的问题，分割网络的注意力需要集中在器官和病变区域的问题，研究提出与注意力模块（Triplet Attention Module，NAM）结合的基础特提取网络并在高级特征提取阶段采用金字塔池化模块（Pyramid Pooling Moudle , PPM）的图像语义分割网络作为解决方法。在包含1850张带注释的超声图像的肾积水数据集上进行实验，包括注意力模块的排列，参数量的计算以及性能。
2. 在语义分割模型提取的分割特征图并结合原图的数据集下，构建疾病严重程度的图像分类模型，为尽可能提高分类准确率，并针对特征信息丢失最严重的池化层，研究提出一种具有小波池化增益层的WaveConvNeXt网络，在利用上一步分割模型预测的语义特征图和原图结合下，补充预测并构造的共3289张图像带标注的数据集上（SFU 0-4）进行实验，最终结果表面我们的诊断模型在分割任务和分类任务上均具备优越性与有效性。

关键词：肾脏超声图像；深度学习；注意力机制；金字塔场景解析；小波变换

# **Abstract**

Health is a life-long topic of human beings, and every stage of a person from childhood to old age is accompanied by the consumption of medical resources. As of May 11, 2021, the results of the seventh national census showed that the total population of the country was 1,443.5 million. The data of the same year showed that there were about 1.023 million medical and health institutions in my country, with about 9.448 million beds and about 9.448 million medical personnel. 10.66 million people. In the context of the increasingly prominent contradiction between population and medical resources, artificial intelligence technology fully empowers all aspects of the medical industry with its high-speed and effective analysis and decision-making capabilities, improves the diagnostic level and efficiency of medical institutions, and promotes reasonable allocation of medical resources.

In the field of congenital ureteropelvic junction obstruction (Ureteropelvic Junction Obstruction Connection, UPJO) in children, B-ultrasound is the most commonly used screening method, and the main treatment method is still surgery. The pathological grade (SFU 0-4) is judged by indexes such as degree, renal parenchyma thickness, and cortical echo. At this stage, experienced radiologists need to make judgments through careful testing to determine the timing of surgery. At present, the preferred and efficient diagnosis technology for the disease is still lagging behind in domestic development, and the distribution of relevant medical imaging talents in China is unbalanced and insufficient. Therefore, a large amount of medical time cost and human resource cost need to be invested in the diagnosis stage, which requires computer imaging processing technology to reduce costs and increase efficiency. The traditional computer image processing methods do not have high accuracy in processing ultrasound images, and the black and white characteristics, boundary ambiguity, and viewing angle diversity of ultrasound images themselves also pose certain challenges to the research on auxiliary diagnosis technology. In order to solve the above problems, this paper studies the computer vision technology based on deep learning, and proposes an intelligent diagnosis scheme based on renal ultrasound images. The specific work is as follows:

(1) We built a semantic segmentation model based on kidney ultrasound images. Aiming at the problem of blurred boundaries between the background area and the organ area of the ultrasound image, the attention of the segmentation network needs to focus on the problem of organs and lesion areas. The research proposal and attention The basic feature extraction network combined with the Triplet Attention Module (NAM) and the image semantic segmentation network of the Pyramid Pooling Moudle (PPM) are used as a solution in the advanced feature extraction stage. Experiments are performed on the hydronephrosis dataset containing 1850 annotated ultrasound images, including the arrangement of attention modules, the calculation of parameter quantities, and the performance.

(2) Based on the segmentation feature map extracted by the semantic segmentation model and the data set of the original image, an image classification model of disease severity is constructed. In order to improve the classification accuracy as much as possible, and for the pooling layer with the most serious loss of feature information. This research proposes a WaveConvNeXt network with wavelet pooling gain layers. Combined with the semantic feature map predicted by the previous segmentation model and the original image, the predicted and constructed data of a total of 3289 images with annotations are supplemented. Experiments are carried out on the set (SFU 0-4), and the final results show that our diagnostic model is superior and effective in both segmentation and classification tasks.

**Key words:** ultrasound image, deep learning, attention mechanism, feature pyramid networks, discrete wavelet transform

目录

[摘 要 I](#_Toc127898294)

[**Abstract** III](#_Toc127898295)

[第1章 绪论 1](#_Toc127898296)

[1.1 研究背景和意义 1](#_Toc127898297)

[1.2 研究现状 3](#_Toc127898298)

[1.3 本文的主要工作 4](#_Toc127898299)

[1.4 本文的主要组织结构 5](#_Toc127898300)

[第2章 相关知识 6](#_Toc127898301)

[2.1 计算机视觉 6](#_Toc127898302)

[2.1.1 图像分类 6](#_Toc127898303)

[2.1.2 语义分割 8](#_Toc127898304)

[2.1.3 评估标准 11](#_Toc127898305)

[2.2 背景知识 13](#_Toc127898306)

[2.2.1 深度神经网络 13](#_Toc127898307)

[2.2.2 注意力机制 14](#_Toc127898308)

[2.2.3 频域图像处理 16](#_Toc127898309)

[2.3 神经网络优化算法 16](#_Toc127898310)

[2.3.1 梯度下降算法 16](#_Toc127898311)

[2.3.2 归一化算法 16](#_Toc127898312)

[2.3.3 正则化算法 16](#_Toc127898313)

[2.3.4 标准化算法 16](#_Toc127898314)

[2.4 本章小结 17](#_Toc127898315)

[第3章 融合注意力机制和金字塔池化的图像语义分割网络 17](#_Toc127898316)

[3.1 模型提出的动机 17](#_Toc127898317)

[3.2 基础特征提取模块 18](#_Toc127898318)

[3.3 高级特征提取模块 22](#_Toc127898319)

[3.3 实验与结论 24](#_Toc127898320)

[3.3.1 数据集与实验设置 24](#_Toc127898321)

[3.3.2 消融实验与分析 26](#_Toc127898322)

[3.3.3 对比实验结果与分析 26](#_Toc127898323)

[第4章 基于离散小波池化增益的图像分类神经网络 28](#_Toc127898324)

[4.1 模型提出的动机 28](#_Toc127898325)

[4.2 神经网络中的离散小波变换(DWT) 29](#_Toc127898326)

[4.3 图像分类网络-WaveConvNeXt 30](#_Toc127898327)

[4.4 实验与结论 32](#_Toc127898328)

[4.4.1 数据集与实验设置 32](#_Toc127898329)

[4.4.2 对比实验与分析 33](#_Toc127898330)

[结论 36](#_Toc127898331)

[参考文献 36](#_Toc127898332)

[攻读硕士期间的主要研究成果 36](#_Toc127898333)

[致谢 36](#_Toc127898334)

# 绪论

随着人工智能（artificial intelligence，AI）时代的到来，在国内医疗资源匮乏、不平衡的现状下，基于大数据医疗的AI方法变得越来越重要。但是医疗领域中的应用场景十分丰富，各类医疗背景下的AI处理方案并不能适用于所有类别，各类背景下与之对应的处理方法研究成果尚不成熟，因此最近几年，随着医疗人工智能的应用越来越多的需求产生，大量的研究人员对其产生了研究兴趣，许多成果以及技术方法也都被应用在各种基于大数据AI的医疗任务上，并且从成果数量上看已经是百花齐放，在各个医疗处理场景下都有相当的优秀产出，大大节省了各地的医疗资源的利用。随着医疗AI获得越来越多的关注度，其研究意义和应用场景也会进一步增多。

本章首先对医疗AI的相关知识以及在当前的研究价值进行了细致的阐述。然后对国内外医疗AI 研究现状进行了简述，并通过调研国内外现有方法的相关内容，得到了现有方法在各自医疗场景下的缺点与不足，最后再详细介绍本文的研究内容和组织架构。

## 1.1 研究背景和意义

人工智能已经发展了大半个世纪，经历几次大起大落。从上世纪80年代到本世纪初，人们对于深度学习探索较多，但受制于计算机的计算能力，以及算法本身的限制，效果不太好，直到2006年，Hinton等人解决了神经网络的大量参数训练的问题。从2009年开始人工智能飞速发展。2015年、2017年，两场世界瞩目的人机围棋大赛之后，人们对人工智能的认识将进一步的加深。而AlphaGo用于下围棋的高效算法是一种通用型的算法，这种算法可以推广到其他算法，把人工智能运用到各个领域，其中就包括医疗领域。

AI医疗可以基于数据采集分析应用于睡眠监测、临床护理、慢性病监测等各种医疗领域。AI赋能医疗，人工智能医疗简单说即以互联网为依托，通过基础设施的搭建及数据的收集，将人工智能技术及大数据服务应用于医疗行业中，提升医疗行业的诊断效率及服务质量。然而我国医疗资源分配严重不均，优质的医疗设备和医护资源大多集中在发达城市与地区，而使得大量外地病患由于在本地得不到良好的医疗，转而向大城市、大医院集中。如图1-1所示，根据国家卫计委数据，截止至2020年底，我国共有医院约32476个，其中三级医院仅有约2498家，约占8%; 然而，三级医院就诊人数却达到16.46亿人次，约占全国总人次的51%。医疗资源供需明显不匹配。



图1 - 1 医疗资源比例图和对应的就诊人次比例图

根据预测，未来我国医疗体系劳动力不足的现象将会进一步加剧。人工智能医疗是解决医疗生产力的根本之道。在我国，人口老龄化、慢病高速增长、医疗资源供需严重失衡以及地域分配不均等问题，造就了对医疗人工智能的巨大需求；同时，我国人口基数大、产业组合丰富、人才储备充分等特点，又给人工智能的发展提供了很好的基础。人工智能有能力减轻临床医生的负担，并为医生提供更好的医疗工具，弥补我国医疗服务中劳动力短缺的情况，提升医疗诊断效率。目前，人工智能技术已广泛地应用于各个领域，与医学领域的联系愈加密切，人工智能技术主要应用在医学影像的诊断环节，解决以下三种需求：

1. 病灶识别与标注：针对医学影像进行图像分割、特征提取、定量分析、对比分析等工作。
2. 靶区自动勾画与自适应放疗：针对肿瘤放疗环节的影像进行处理。
3. 影像三维重建：在人工智能进行识别的基础上进行三维重建，针对手术环节的应用。

对于第一个需求，其主要工作是与医学影像（CT、MRI、PET）相结合的科学应用。因为超声具有无电离辐射、无痛无创、轻便简洁、即时成像、高复用性等等优势优点，现已广泛应用于肾、肝、心血管、乳房等内脏器官以及浅表结构的初步检查与医学诊断[<akkus2019survey>]中。但是，超声扫描检查具有一定的主观性，从而衍生出主观差异性，不同超声医生的判断也不尽一致，并且一名合格的超声医师需长期大量的培训与学习。相比之下，人工智能与超声影像结合可大大简化操作步骤，降低甚至避免主观差异性、节约医师珍稀资源、缩短报告的时间、提高诊断的效率。

本文研究背景下的主要人群为儿童以及先天待产儿童，而医疗背景为肾盂输尿管交界处梗阻即UPJO，该疾病是可能引起肾脏功能损害的一种泌尿系畸形，是引起新生儿肾积水的最常见病因, UPJO会使得尿液流出受阻，导致肾盂、肾盏内压力增高，肾小管压力增大，集合系统内压力长期持续增加，肾脏血流量减少，导致肾脏缺血性损伤和不可逆的病理改变，称为梗阻性肾病。如果不予治疗将导致进行性肾功能损害或限制儿童发育肾脏最终的功能潜能。目前主要的检查手段是产前超声检测和产后超声检查，UPJO在超声检查的图像上主要表现为器官内部出现一定面积的阴影区域，人工的分辨方式会根据超声图像中的几何、纹理、阴影分布等特征来判断疾病的病理等级，然而，由于超声波图像是高噪声的，即使是训练有素的医学影像医师也要获得足够的信息[<akkus2019survey>]，这种人工的特征提取处理方式使得医学影像医师的处理速度不高。因此为解决该问题，可以通过人工智能分支中的计算机视觉技术对超声图像做大数据AI分析，快速有效地提出一个初步的诊断结果，从而加快医疗进程。在基于超声图像的病灶识别与标注领域，人工智能的发展也在不断地促进这一领域蓬勃向上的趋势，也不断有创新性的研究成果在各医学影像的应用场景下孵出，解放医学影像人才的劳动力，节省大量的超声医疗资源，帮助医疗机构与患者节省财力，并且能够辅助医生对患者的手术时机做出及时有效的判断，因此具备极大的社会价值。

## 1.2 国内外研究现状

随着计算能力的发展，人工智能技术已应用于众多地区，尤其是医学图像分析。它可以极大地帮助医生的诊断能力，同时缓解欠发达地区的医疗资源失衡和短缺[<smail2020using>, <turco2020contrast>, <shokoohi2019enhanced>, <dhindsa2018grading>, <blum2018early>, <he2019practical>]。近年来，医学影像与人工智能的结合正在盛行。其中深度学习作为人工智能的一个子集，可以从数据集中学习，找出自动得出结论的重要特征。如果我们能够结合人工智能医学成像技术，在超声检查阶段自动区分病变区域，我们不仅可以支持泌尿外科医生的进一步诊断，还可以节省大量的医疗资源、人力、金钱，帮助负担过重的患者，不仅人工智能医学影像技术在很大程度上缓解了医疗短缺资源和不同地区之间的不平衡，并且促进医生的诊断能力[<gulshan2016development>, <jussupow2021augmenting>, <kumar2022artificial>, <lee2021application>]。

对于大多数结合人工智能的医疗相关的研究，其重心都倾向于超声图像的病灶识别与标注[<rotemberg2021patient>]，人工智能长期以来一直应用于医学图像处理等各个领域，极大地促进了医生的诊断能力。Zhu[<zhu2021generic>]等人开发了一种使用深度卷积神经网络对超声图像中的甲状腺和乳腺病变进行分类的自动方案，其结论表明其模型在对乳腺癌和甲状腺癌进行分类时具有更深潜力，比放射科医生的准确性更高。人工智能在肝脏超声方面也有相关的应用，主要为肝脏脂肪检测及评估肝纤维化等[<lassau2021three>]。 Biswas[<biswas2018symtosis>] 等人运用深度学习方法评估脂肪肝，并与其他 2 种方法即 SVM、极限学习机（ELM）相比较，诊断准确率分别为 100%、82%、92%，其结论表明相比机器学习、使用深度学习能更好地判别脂肪肝。Burgos等人开发了一种人工智能方法，通过分析胎儿大脑信息自动估计胎龄，将其性能与4种目前使用的胎儿生物测定公式进行了比较，与胎儿生物特征参数相比，使用标准超声胎儿平面的自动化人工智能方法在孕龄估计方面产生了相似或更低的误差。可见在医疗领域，人工智能正在逐渐探索自己的位置，Shah等人[<shah2020artificial>]总回顾了47篇文章，报告了AI在泌尿系统癌症中的应用，在所有良性条件下，AI都被用来预测手术的结果，影像组学在肾脏肿块的分类和核分级、膀胱癌的膀胱镜诊断、格里森评分预测以及磁共振成像与前列腺癌的计算机辅助诊断中都有AI的相关应用，结论表面临床范式即将转变，AI应用将在该领域中找到自己的位置，并彻底改变决策过程。

关于肾脏器官领域，与泌尿科息息相关，其医学影像病灶识别也有不少研究成果，其中肾脏病理是肾脏疾病诊断的金标准，迄今为止已有多项研究显示AI技术能够在这项金标准中得到高效且成熟应用,比如肾小管、肾小球以及间质等肾脏基本结构和病变的病灶识别、预测肾脏病理分型等,从而实现自动化的特征量化识别、病理分型和判断，Gallego等人[<gallego2018glomerulus>]使用深度学习框架从数字化肾脏切片中自动进行肾小球分类和检测，结论表明，该技术适用于在全幻灯片图像 (WSI) 中正确检测肾小球，显示出鲁棒性并降低了假阳性和假阴性的检测率。对于UPJO诊断，早期已有试图从机器学习领域寻求效率突破的研究，Lorenzo等人[<lorenzo2019predictive>]结合机器学习与云计算技术，提出了一个通过数据分析解决当代挑战的机会，报告了一个超越当前标准的创造性解决方案，探索利用云机器学习平台预测产前婴儿UPJO手术时机的潜在价值。Blum等人[<blum2018early>]引入了一种动态解决方案，使用自动信号分析和机器学习来分析利尿肾图的引流曲线，研究结果表明，机器学习方法具有潜在的临床实用性，可以更早地发现梗阻病例，但以上研究可靠性不足，仅仅证明机器学习方法的潜在实用性价值。

另外，在深度学习与肾脏UPJO的跨领域内，已有学者探索并发表研究，然而已有的可靠研究成果并不丰富，大部分还停留在理论阶段，截至2021年，Hameed等人[<hameed2021artificial>]在该领域收集大量资料，其综述结论表明AI在该领域的未来是光明的，但在提供可靠的结果方面，仍然有很大的改进和增长空间以积极影响更多的生命。Adree等人[<khondker2022machine>]研究结论表明使用定量指标对膀胱输尿管反流（VUR）进行分级是可能的，未来可以应用机器学习方法对 VUR 进行客观评分。Drysdale等人[<drysdale2021personalized>]提出一种模型预测肾盂成形术后再次干预的风险和时间，结论表面模型表现良好，其方法的实施在该科领域是新颖的，可能有助于对接受肾盂成形术的患者进行个性化的风险分层，但需要进一步的真实世界验证。Smail等人[<smail2020using>]探讨了针对SFU分级分类的问题，提出一种平均准确率为71%深度学习模型并首次证明了CNN方法对肾积水超声图像的适用性，Tabrizi等人[<tabrizi2021pediatric>]提出了一种基于深度学习的方法来预测肾盂积水的严重程度，该方法由肾脏和输尿管交界处是否存在阻塞或阻塞来定义。首先对肾脏进行半自动分割，以分析其表征梗阻的外观。然后开发了一个基于深度学习的模型，使用超声图像中的每个切片来预测阻塞，将平均准确率提高到78%。 Lin等人[<lin2021evaluation>]提出一种结合注意力的模型将肾脏和扩张的盆腔系统与液体分开，结果表明模型检测扩张盆腔系统的敏感性和特异性分别为99%和83%，检测肾积水的敏感性和特异性分别为90%和80%，将检测肾积水的准确度进一步提高。Lien等人研究结论表面Res-UNet算法对中度/重度肾积水分类具有显著的准确度，但数据集来源患者年龄跨度太大，并不适用于儿童先天性UPJO的临床诊断，也缺乏对正常和轻症患者的科学评估。综上现有方法对先天性儿童UPJO超声图像的准确度以及可靠性，均未达到该领域临床诊断的要求，以及本文研究所涉及的应用场景，因此在该领域的研究仍在尝试着新的突破，本文研究也在前人学者的基础上做出了一定的贡献。

## 1.3 本文的主要工作

随着人工智能应用在医疗领域的横向拓展，为缓解当前医疗资源与医疗需求的不平衡的矛盾，节省大量的人力、医疗 资源、金钱和帮助受苦的病人，基于医学影像的结合计算机视觉的技术也应运而生。伴随着医疗器械的发展，如何利用超声图像来辅助医生完成最优的医疗诊断、治理过程也变得尤为关键。本文研究项目来源于国家重点研究项目，与首都医科大学附属儿童医院协同合作，并将先天性儿童UPJO在医学诊断阶段的智能分级作为主要的研究方向，突破性地提出了一种从语义分割模型再到图像分类模型的分段诊断方法，模型会利用超声图像所展示的器官特征、病变特征、纹理特征等信息，分别对不同类别的患者的肾脏病理等级做出高效率、高准确度的初步诊断。本研究的模型将在基于现有研究的基础上，进一步提高诊断的准确率，为该领域医疗AI处理落地临床实用的最终目标贡献科研成果。以下是本文的主要研究内容：

1. 分析数据并对现有医疗图像方法的调研与分析

首先对医院提供的超声检查的视频进行了图像帧的周期性抽取，然后对图像数据集中的脏数据进行做清理与过滤，分析思考其中运用计算机视觉技术的难点与痛点。接着针对性地调研相关该领域内前人的可靠的研究成果，探索创新性的解决方法，分析现有研究成果的模型与其优劣性，然后研究合适的优化算法，并提出完整的方法论。

1. 提出基于肾脏超声的融合注意力机制的语义分割模型

在图像语义分割阶段，针对器官区域、病变区域和非相关背景区域的位置关系和边界模糊的问题，我们提出了一种结合注意力机制和金字塔场景解析结构的语义分割模型，帮助神经网络关注标注的器官与病变区域、更好地学习全局信息以区分各区域，最终提高图像语义分割的精度，然后将分割结果按一定比例与原图结合生成新的数据集。

1. 提出融合小波池化增益层地图像分类模型

在图像分类阶段，创新性地引入离散小波替换池化层，从频域解析出三个低频分量与一个高频分量，并行学习，在神经网络中构造分支完成参数传播并完成异化特征组合的学习，其中在归一化处理模块、激活函数模块、神经网络剪枝优化模块方面也均调研了相关技术并应用其中，综合提高了最终分类模型对于UPJO超声影像的诊断准确率。

尽管在UPJO病理研究领域内，结合深度学习方法的研究比较少，对于UPJO引发的肾积水的诊断仍具有挑战性，因为其表现各不相同且非特异性，超声具有易于接近、无辐射暴露和重复评估的特点，成为肾积水的补充诊断工具，但经过耗时的培训后，影像科医师之间的主观差异仍然存在，也不断有学者提出最新研究成果并评估深度学习算法通过超声图像检测肾积水的可行性，表明深度学习算法对肾积水的检测具有显著的召回率、特异性、精密度和准确度，减少了超声医师之间的变异性并提高临床条件下的效率，但是这些研究忽略了轻症甚至正常的超声图像的诊断识别，直接对图像分类的诊断过程也欠缺鲁棒性。相对而言，本研究的主要贡献有三方面:。

1)我们提出了一种分割-分类的诊断框架，并将其应用于肾积水超声图像的智能诊断，取得了优异的综合性能。

2)在分割阶段不破坏骨干网络结构的前提下满足迁移学习条件并轻量级地嵌入注意力机制提高图像分割精度，结合金字塔场景解析模块针对性地解决了肾脏超声图像的上下文特征难题。

3)在分类阶段思考图像频率域的运用可能性，并创新型地加入小波池化增益层辅助神经网络模型学习到更多维度的、更加鲁棒的特征信息，并提高最终的分类性能。

## 1.4 本文的主要组织结构

本文的组织结构如下：

第一章：绪论。

本章主要从研究的相关背景和各研究意义的角度出发分别介绍了AI与医疗领域的结合，细分到AI在医学影像领域内的研究现状，再然后到UPJO应用场景下的现有研究，分析各研究的贡献与不足。之后引出本文研究所解决的问题和以研究内容的应用价值，以及本研究所提出的解决方案，并阐述主要工作，最后阐述了本文的组织结构。

第二章：相关技术

本章首先对深度学习在肾脏超声领域内的相关研究分析，提出不足并引出本文研究，并基于本文对肾脏超声图像的诊断模型所用的技术进行了分解，大体分为语义分割和图像分类两个阶段，并分别介绍了两个阶段的相关技术发展、相关的评价指标和计算方式，然后背景知识内分别介绍了各个阶段内的更细粒度的技术模块，包括注意力模块、金字塔场景解析模块和各种优化算法，并简单介绍了这些模块下对应领域内的相关研究，以及对本次研究的结合的最优方案。

第三章：融合注意力机制和金字塔场景解析的图像语义分割网络

本章节首先会大致勾勒并介绍分割网络模型的大致结构，并按输入到输出的顺序一一进行阐述与理论分析，对注意力机制的算法和结合网络的方式进行公式化的介绍，具象化地介绍了注意力机制所解决的问题，然后详细介绍了在高级特征处理阶段的金字塔场景解析过程，阐述其算法原理与解决的问题，最后再详细地介绍这一阶段对整个诊断过程所做出的贡献、其作用与意义。

第四章：基于离散小波池化增益的图像分类神经网络

本章节首先大致介绍了分类网络模型的大致结构，同样地以输入到输出的顺序进行解析，在对离散小波变换的运用介绍中，由连续小波变换循序渐进到图像中的离散小变换，再由一维的计算公式推导出二维图像下的计算公式，由浅入深、由简到繁的分析其理论原理与神经网络的结合。然后分别介绍了其中所用的优化方法，包括剪枝算法、归一化算法等，并从理论原理角度分析其算法过程，以及与本文研究方向为何、如何相结合的细致介绍。

# 第2章 相关知识

## 2.1医疗图像处理与深度学习

### 2.1.1 医学图像分割

用于图像处理的典型卷积神经网络（CNN）架构由一系列卷积网络组成，其中包含有一系列数据缩减即池化层。与人脑中的低级视觉处理一样，卷积网络检测提取图像特征，这些特征可能表示为直边的线或圆，例如器官特征，或者是圆圈，例如结肠息肉的特征，然后是更高阶的特征，例如局部和全局形状和纹理特征提取，特征上下文信息，例如器官与器官内组织的空间关系。图像分割作为经典的计算机视觉问题，其涉及将一些原始数据作为输入并将它们转换为具有突出显示的感兴趣区域的掩膜，其中图像中的每个像素根据其所属的对象被分配类别标签[<liu2019recent>]。

对于医学图影像数据，准确分割医学图像是放射治疗计划期间轮廓的关键步骤，并且早在2010年就有权威研究论述了AI技术在医疗图像处理中的应用[<sharma2010automated>]，其中医学图像分割就是一个根据区域间的区分度，把图像分割成若干区域并掩模化的过程[<castleman1996digital>, <pratt2007digital>]。目前，主要以各种细胞、组织与器官的图像作为处理的对象,传统的图像分割技术有基于区域的分割方法和基于边界的分割方法，前者依赖于图像的空间局部特征，如灰度、纹理及其它象素统计特性的均匀性等，后者主要是利用梯度信息确定目标的边界。通常，医学图像分割可以用集合论模型来描述：给定一个医学图像和一组相似性约束 ,的分割即是得到它的所有子集区域的划分，即：

其中满足信息相似约束 下所有像素的集合。对于同理， 用来区分不同区域， 为不小于2的正整数，表示区域划分后的序号。

早起的分割算法主要是灰度分割，条件随机场等一些较为传统的算法：

1. 最简单的语义分段形式涉及分配区域必须满足的硬编码规则或属性，以便为其分配特定标签。规则可以根据像素的属性（例如灰度级强度）来构建。使用此技术的一种方法是拆分（Split）和合并（Merge）算法。该算法递归地将图像分割成子区域，直到可以分配标签，然后通过合并它们将相邻的子区域与相同的标签组合。这种方法的问题是规则必须硬编码。此外，仅用灰色级别的信息来表示复杂的类(如人)是极其困难的，在面对超声图像这类复杂的图像场景运用下，这种方式显然已经因为其滞后性与低效性，使得无法应用于医疗图像处理领域。因此，需要特征提取和优化技术来正确地学习这些复杂类所需的表示。
2. 再后来机器学习渐渐兴起，考虑通过训练模型为每个像素分配类来分割图像。如果模型不完美，可能会得到自然界不可能得到的噪声分割结果。可以通过考虑像素之间的先验关系来避免这些问题，例如，对象是连续的，因此附近的像素往往具有相同的标签。为了模拟这些关系，可以使用条件随机场（CRF）。CRF是一种用于结构化预测的统计建模方法。与离散分类器不同，CRF可以在进行预测之前考虑“相邻上下文”，比如像素之间的关系。这使得它成为语义分割的理想候选。图像中的每个像素都与一组有限的可能状态相关联。在我们的示例中，目标标签是可能的状态集。将一个状态(或标签，u)分配给单个像素（x)的成本称为它的一元成本（unary cost）。为了对像素之间的关系建模，将一对标签(u,v)分配给一对像素(x,y)的成本，即成对成本（pairwise cost），还可以考虑它的近邻像素对(Grid CRF)或者我们可以考虑图像中的所有像素对(Dense CRF)
3. 与此同时随着深度学习技术的不断发展，深度学习在语义分割任务中大放异彩，取得了骄人的成绩。从第一篇真正意义上的深度学习方法语义分割模型FCN（[Fully Convolutional Networks for Semantic Segmentation](https://arxiv.org/abs/1411.4038)）介绍起:
4. FCN作为语义分割始祖之作，其主要贡献有三点：a、将端到端的卷积网络推广到语义分割中;b、重新将预训练的ImageNet网络用于分割问题中;c、使用反卷积层进行上采样;提出了跳跃连接来改善上采样的粗糙程度。该模型的关键在于网络中的全连接层可以看作是使用卷积核遍历整个输入区域的卷积操作，这相当于在重叠的输入图像块上评估原始的分类网络，但是与先前相比计算效率更高，因为在图像块重叠区域，共享计算结果。该模型用卷积运算实现的全连接层结构，在将预训练网络模型的全连接层卷积化之后，由于CNN网络中的池化操作，得到的特征图谱仍需进行上采样。反卷积层在进行上采样时，不是使用简单的双线性插值，而是通过学习实现插值操作。此网络层也被称为上卷积、完全卷积、转置卷积或是分形卷积。然而，由于在池化操作中丢失部分信息，使得即使加上反卷积层的上采样操作也会产生粗糙的分割图，因此该算法还从高分辨率特性图谱中引入了跳跃连接方式，但仍然这类分割效果图在噪声高、细节信息与边缘信息模糊的超声图像运用场景下略显乏力。
5. 针对FCN在语义分割时感受野固定和分割物体细节容易丢失或被平滑的问题,SegNet被提出。SegNet和FCN思路十分相似，编码部分主要由VGG网络的前13个卷积层和5个池化层组成，解码部分同样也由13个卷积层和5个上采样层组成，最后一个解码器输出的高维特征被送到可训练的分类器中，用于分类每个独立的像素。特别地，SegNet网络采用池化指数来保存图像的轮廓信息，降低了网络参数数量。
6. 在医学图像领域，但由于医学影像处理本身的特殊性,能够使用用于训练的数据数量还是比较少的, 因此U-Net模型算法应运而生，它有效地提升了使用少量数据集进行训练检测的效果，提出了处理大尺寸图像的有效方法。其主要贡献也有三点：a、以重叠-拼接的策略能够实现对任意大的图像进行无缝分割，同时每个图像块也获得了相应的上下文信息。b、随机弹性变形进行数据增强搭配图像分块在数据量较少的情况下起到了扩充数据量的作用。更重要的是，这种策略不需要对原图进行缩放，每个位置的像素大小与原图保持一致，不会因为缩放而带来误差。c、引入加权损失函数以解决数据不平衡的问题。该模型继承了FCN的思想，继续进行改进。但是相对FCN，U-Net是完全对称的，且对解码器进行了加卷积加深处理，FCN只是单纯的进行了上采样，两者都用了跳跃连接的结构，这样的结构所带来的明显好处是可以联合高层语义和低层的细粒度表层信息，就很好的符合了分割对这两方面信息的需求。在FCN中，跳跃连接的联合是通过对应像素的求和，而U-Net则是对其的通道的连接过程。
7. 对于DeepLab系列，DeepLabv1 是由深度卷积网络和概率图模型级联而成的语义分割模型，由于深度卷积网络在重复最大池化和下采样的过程中会丢失很多的细节信息，所以采用扩张卷积算法增加感受野以获得更多上下文信息。考虑到深度卷积网络在图像标记任务中的空间不敏感性限制了它的定位精度，采用了完全连接条件随机场（Conditional Random Field，CRF）来提高模型捕获细节的能力。DeepLabv2 语义分割模型增加了 ASPP（Atrous spatial pyramid pooling）结构，利用多个不同采样率的扩张卷积提取特征，再将特征融合以捕获不同大小的上下文信息。DeepLabv3 语义分割模型，在 ASPP 中加入了全局平均池化，同时在平行扩张卷积后添加批量归一化，有效地捕获了全局语境信息。DeepLabv3+语义分割模型在 DeepLabv3 的基础上增加了编-解码模块和 Xception主干网络，增加编解码模块主要是为了恢复原始的像素信息，使得分割的细节信息能够更好的保留，同时编码丰富的上下文信息。增加 Xception 主干网络是为了采用深度卷积进一步提高算法的精度和速度。在inception结构中，先对输入进行1×1的卷积，之后将通道分组，分别使用不同的3×3卷积提取特征，最后将各组结果串联在一起作为输出。
8. 聚焦全局上下文信息的网络同样还有本次研究参考使用的PSPNet (Pyramid Scene Parsing Network，PSP) 金字塔场景稀疏网络语义分割模型（Pyramid Scene Parsing Network，PSP）首先结合预训练网络和扩张网络来提取图像的特征，得到原图像八分之一大小的特征图，然后，采用金字塔池化模块将特征图同时通过四个并行的池化层得到四个不同大小的输出，将四个不同大小的输出分别进行上采样，还原到原特征图大小，最后与之前的特征图进行连接后经过卷积层得到最后的预测分割图像。其主要特点包括：金字塔场景解析网络是建立在FCN之上的基于像素级分类网络，将大小不同的内核集中在一起激活特征图的不同区域创建空间池金字塔。特性映射来自网络被转换成不同分辨率的激活，并经过多尺度处理池层，稍后向上采样并与原始层连接进行分割的特征图。学习的过程利用辅助分类器进一步优化了像残差网络这样的深度网络模型，不同类型的池模块侧重于激活的不同区域特征图

以上内容主要对于图像语义分割技术的研究发展历程进行了详细评述，对于传统的语义分割方法到当前主流的基于深度学习的图像语义分割理论及其方法做出了综合性的介绍，对基于深度学习语义分割技术需要用到的网络模型、网络框架、分割流程进行了详细的评估。该领域仍然存在着非常多的未知问题值得深入探究，且该领域的未来研究方向有：a、实时语义分割技术：现阶段评价应用于语义分割的网络模型主要着重点在精确率上，但是随着应用于现实场景的要求越来越高，需要更短的响应时间，因此在维持高精确率的基础上，尽量缩短响应时间应是今后工作的方向。b、弱监督或无监督语义分割技术:针对需要大量的标注数据才能提高网络模型的精度这个问题，弱监督或无监督的语义分割技术将会是未来发展的趋势。c、三维场景的语义分割技术:目前的诸多基于深度学习的语义分割技术所用以训练的数据主要是二维的图片数据，同时测试的对象往往也是二维的图片，但是在实际应用时所面对的环境是一个三维环境，将语义分割技术应用至实际中，未来需要针对三维数据的语义分割技术进行研究。

### 2.1.2 图像分类

图像分类是计算机视觉中很基础的一个任务，也是几乎所有的基准模型进行比较的任务。从最开始比较简单的10分类的灰度图像手写数字识别任务MINIST，到后来更大一点的10分类的 CIFAR10和100分类的CIFAR100 任务，到后来的ImageNet任务，图像分类模型伴随着数据集的增长，一步一步提升到了今天的水平。现在，在ImageNet这样的超过1000万图像，超过2万类的数据集中，计算机的图像分类水准已经超过了人类。图像分类顾名思义就是一个模式分类的问题，它的目标是将不同的图像，划分到不同的类别，实现最小的分类误差。总体来说，对于单标签的图像分类问题，它可以分为跨物种语义级别的图像分类，子类细粒度图像分类，以及实例级图像分类三大类别。其中所谓跨物种语义级别的图像分类，是指在不同物种的层次上识别不同类别的对象，比较常见的包括如猫狗分类等，这样的图像分类，各个类别之间因为属于不同的物种或大类，往往具有较大的类间方差，而类内则具有较小的类内误差；而子类细粒度图像分类相对于跨物种的图像分类，级别更低一些，它往往是同一个大类中的子类的分类，如不同鸟类的分类，不同狗类的分类，不同车型的分类等；如果我们要区分不同的个体，而不仅仅是物种类或者子类，那就是一个识别问题，或者说是实例级别的图像分类，最典型的任务就是人脸识别。

图像分类任务从传统的方法到基于深度学习的方法，经历了几十年的发展。本文只关注于深度学习的进展，下面重点讲述几个重要的节点。在计算机视觉分类算法的发展中，MNIST 是首个具有通用学术意义的基准，它是一个手写数字的分类标准，包含 60000 张训练数据， 10000张测试数据，图像均为灰度图，通用的像素大小为 28×28。在上个世纪90年代末本世纪初，SVM 和 K-最邻近方法被使用的比较多，以SVM为代表的方法，可以将MNIST分类错误率降低到了0.56%，彼时仍然超过以神经网络为代表的方法，即LeNet系列网络。LeNet网络诞生于1994年，后经过多次的迭代才有了1998年的LeNet5，是为我们所广泛知晓的版本。虽然LeNet5当时的错误率仍然停留在0.7%的水平，不如同时期最好的SVM方法，但随着网络结构的发展，神经网络方法很快就超过了其他所有方法，错误率也降低到了0.23%，甚至有的方法已经达到了错误率接近0的水平。在本世纪的早期，虽然神经网络开始有复苏的迹象，但是受限于数据集的规模和硬件的发展，神经网络的训练和优化仍然是非常困难的。MNIST和CIFAR数据集都只有60000张图，这对于10分类这样的简单的任务来说，或许足够，但是如果想在工业界落地更加复杂的图像分类任务，仍然是远远不够的。在ImageNet发布的早年里，仍然是以SVM和Boost为代表的分类方法占据优势，直到2012年AlexNet的出现。AlexNet是第一个真正意义上的深度网络，与LeNet5的5层相比，它的层数增加了3层，网络的参数量也大大增加，输入也从28变成了224，同时图像处理器(GPU)的面世，也使得深度学习进入新的时代。相比LeNet5, AlexNet包括5个卷积层和3个全连接层;使用更好的激活函数，收敛很快，解决了Sigmoid在网络较深时出现的梯度弥散问题;加入了随机剪枝层，防止过拟合；加入归一化层，对局部神经元的活动创建竞争机制，抑制反馈较小的神经元放大反应大的神经元，增强了模型的泛化能力；使用裁剪翻转等数据增强的手段，增强了模型的泛化能力；其总体的数据参数大概为240M，远大于LeNet5。

之后分类任务的网络相继出现GoogleNet和VGGNet, 其中VGGNet包括16层和19层两个版本，共包含参数约为550M。全部使用3×3的卷积核和2×2的最大池化核，简化了卷积神经网络的结构。VGGNet很好的展示了如何在先前网络架构的基础上通过简单地增加网络层数和深度就可以提高网络的性能。虽然简单，但是却异常的有效，时至今日，VGGNet仍然被很多的任务选为基准模型。GoogLeNet是来自于Google的Christian Szegedy等人提出的22层的网络，其核心是Inception模块，它采用并行的方式，包括有四个成分。1×1的卷积，3×3的卷积，5×5的卷积，3×3的最大池化层，最后对四个成分运算结果进行通道上的组合，即Inception模块的核心思想。通过多个卷积核提取图像不同尺度的信息然后进行融合，可以得到图像更好的表征。自此，深度学习模型的分类准确率已经达到了人类的水平(5%~10%)。

直至2020年，受到NLP领域中Transformer成功应用的启发，Vison Transformer算法尝试将标准的Transformer结构直接应用于图像，并对整个图像分类流程进行最少的修改，将整幅图像拆分成小图像块，然后把这些小图像块的线性嵌入序列作为Transformer的输入送入网络，然后使用监督学习的方式进行图像分类的训练，一度让研究者们认为CNN结构分类网络失去了竞争优势，直到ConvNeXt的出现，重新依靠CNN结构达到了ImageNet第一名的准确率。

### 2.1.3 评估标准

图像分类的概念：是一个模式分类的问题，将不同的图像划分到不同的类别、当想知道类别之间分类错误的情况，就是查看是否有特定的类别相互混淆，就可以绘制混淆矩阵画出分类的详细预测结果。对于包含多个类别的任务，混淆矩阵很清晰的反应了各个类别之间分类错误的概率；其实就是把所有类别的预测结果与真实结果按类别放置到了同一个表里，在以下表格中我们可以清楚看到每个类别正确识别的数量和错误识别的数量。

表 2-1 预测结果表

Tab. 2-1 Table of Prediction Result

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 正样本且分类为正  （TP） | 负样本但分类为正  （TN） |
| F | 正样本但分类为负  （FP） | 负样本且分类为负  （FN） |

需要现根据表2-1-2，计算四个基本的指标：分别为TP:正样本且分类为正样本的数目为（True Positive）；FN:标签为正样本分类为负样本的数目（Flase Negative）；FP:标签是负样本分类为正样本的数目（Flase Positive）；TN：标签是负样本且分类为负样本的数目（True Negative），基于以上四个指标：单标签分类的评价指标有：准确率(Accuracy),精确率（Precision）,召回率（Recall），F1-score，混淆矩阵，ROC曲线和AUC，

在所有样本中，预测正确的概率，即准确率（Accuracy）计算方式为：；在正样本中正确的概率,即精确率（Precision）的计算方式为: ；在正样本被找出来的概率，即召回率（Recall）的计算方式为：

召回率越高，精确率越低。例如概率阈值T的影响（概率阈值T是一个设定的在[0,1]范围的值，使用该值判断网络预测出的结果是正类还是负类），当T小，导致召回率的升高，分类精确度下降，当T大，精确度升高，召回率降低。如果样本中包含很多的类，那么预测类别就是预测概率最大的那一类。基于精确率（Precision）和召回率（Recall）共同组成的曲线是PR曲线。PR曲线中的P代表的是精准率），R代表的是召回率），其代表的是精准率与召回率的关系，一般情况下，将召回率设为横坐标，精确率设为纵坐标。一条PR曲线要对应一个阈值。通过选择合适的阈值对样本进行合理的划分，概率大于阈值的样本为正例，小于阈值的样本为负例，曲线越靠近右上角说明性能越好，可以根据曲线和坐标轴包围的面积来定量评估.

基于精确率和召回率的指标，同样还有F1-score，其计算方式为：，在众多模型对数据进行学习后，如果其中一个模型的PR曲线完全包含另一个模型的PR曲线，则可断言其性能的优越性。但假如二者发生交叉，则可以根据曲线下方的面积大小来进行比较，但更常用的是平衡点F1-score,F1值越大，我们可以认为该模型的性能较好。

准确率,精确率,召回率,F1-score都是单一数值指标，如果我们想观察分类算法在不同参数的表现，可以用一条曲线表示[ROC曲线](https://so.csdn.net/so/search?q=ROC%E6%9B%B2%E7%BA%BF&spm=1001.2101.3001.7020)；ROC曲线可以评价一个模型在不同阈值下的表现。对于ROC曲线，其横坐标是FPR（False Positive Rate）计算方式为：、纵坐标为TPR（True Positive Rate），计算方式为：。

ROC曲线上的每一点对应于一个阈值，当阈值最大时，TP=FP=0，对应于原点，当阈值最小时，TN=FN=0，对应于坐标（1，1）；随着阈值增加，TP和FP都减小，ROC点向左下移动。ROC曲线来评判分类器的性能，则一个分类器中的ROC完全包含另一个ROC曲线，说明该模型性能优于被包含的模型。

因为语义分割实质上是像素级别的分类，因此对于图像语义分割的评估标准，图像分类的指标同样适用。除此之外，还有平均像素准确性（Mean Pixel Accuracy）,计算预测类别正确数占总像素数的比例，然后计算在其类别的预测值中，真实属于其类别的像素准确率，即类别像素准确率（CPA），对于二维图像的所有像素点，其计算方式为：，最后分别计算每个类别的CPA后累加求平均值，计算方式为：，另外MIoU（Mean Intersection over Union）平均交并比，也可作为语义分割系统性能的评价指标，其计算的是所有类别的交并比的均值结果。其中交并比是预测值和真实值的交集和并集之比，其计算方式为。这种评价标准相比MPA而言，直观上如果某一类物体体积较小，在图像上所占的像素比例就小，如果这类物体预测全对或者全错，即交并比为0或者1，对最终的平均交并比的结果影响就较大，但是MPA是统计的每一个像素的预测准确率，这种极端情况的影响就较小。因此一般结合多种评价指标综合分析分割模型的有效性。

## 2.2 背景知识

### 2.2.1 深度神经网络

在早期，神经网络模型的研究起源与认知心理学、计算神经科学领域。1943年Warren McCulloch和Walter Pitts发明了基于数学和阈值逻辑的算法神经网络计算模型，即MCP神经元数学模型，为神经网络的发展拉开了序幕。1958年，Rosenblatt创造了感知机并将其成功运用于简单的模式识别任务中，将神经网络的研究推向了第一次的高潮。一些学者对神经网络的发展过度乐观，甚至认为“20年内机器可以做所有人做的事”。与此同时各国政府，如美国国防部（DoD）也开始大量资助神经网络的研究。然而好景不长，1969年Minsky和Papert直接指出目前的神经网络存在两个关键的缺陷：

（1）无法解决异或这类简单的线性不可分问题；

（2）计算机没有足够的计算力来求解大型的网络。

这些问题在当时直接决定了神经网络的不可行性，神经网络的研究进入了第一次寒冬。在神经网络的第一次寒冬期间也有部分学者取得了一定的成就。例如，1959年Hubel、Wiesel通过对猫的视觉皮层的研究发现，哺乳动物的视网膜上同时存在着两种细胞即视锥细胞和视杆细胞，其分别对颜色和明暗敏感。视网膜接收光信号后将其转化为电信号并分两路向视觉皮层传播。视神经传来的信号主要经过初级视觉皮层（V1）、二级视觉皮层（V2）、三级视觉皮层（V3）等层层处理抽象，最终在我们大脑中形成影像。该研究于1981年获得诺贝尔生理学或医学奖，并直接启发了日本学者Fukushima。1980年Fukushima根据猫的视觉皮层中视觉感受野以及视神经信号层层传播处理的思想发明了感知机。该网络可以视为CNN最早期的雏形，但当时该网络的训练为自组织的方式，其并未采用误差反向传播，当数字的位置或形态稍有变化时其识别效果并不理想，然而这一工作却为日后LeCun研究的重要基石。

1983年加州理工学院的物理学家John Hopfield利用神经网络，通过电路模拟仿真的方法求解了旅行商NP难题，在学术界引起的较大的轰动，这也推动了神经网络第二次的快速发展。同时1986年BP误差反向传播算法被Hinton等人再次发明，这也为后来神经网络的发展奠定了基础。1989年George Cybenko首次提出了sigmoid激活函数的万能逼近定理，2年后Kurt Hornik指出万能逼近定理并不依赖于特别的激活函数，而是由多层前馈网络结构所决定的，例如对于任何的Boolean function 我们只需使用一层MLPs，同时对于任意的分类问题我们也只需使用一层的MLPs，当然若增加MLPs的层数，其每一层神经元的数目将会减少，即宽度减少（对于网络设计时我们一般倾向于增加深度而非宽度这主要是因为，深度的增加对精度的提高更加有效。同时在1989年LeCun大神正式登场，其通过利用BP算法完成了对手写数字的识别。1998年LeCun正式提出LeNet-5，该网络即为标准的CNN，其采用了卷积、池化的结构，并选择双曲正切激活函数和极大似然估计损失函数，同时使用BP算法进行训练。该网络在手写数字识别上获得成功，并应用于美国的邮政系统中。然而受限于当时计算机的发展、数据的匮乏以及网络优化的困难（梯度消失），神经网络的发展再次受挫，进入第二次寒冬。同时，1992年Bernhard E. Boser, Isabelle M. Guyon等人利用kernel的技巧发明了非线性SVM这种基于表层学习的分类器。1993年Corinna Cortes和Vapnik提出了软间隔的思想，再次强大了SVM的能力。这也使得多数学者转而研究SVM，神经网络方法受到了极大地打压。

2006年Hinton在 Science上发文指出利用RBM编码预训练深度神经网络与PCA相比在高维特征抽取方面有更佳的性能，即深度网络拥有强大的特征提取能力。同时还指出深层的网络训练可以通过逐层训练的方式实现，这也使得更深网络的训练成为了可能。虽然这篇文章现在看来并没有在理论上做出较大的创新，尤其是逐层训练的方式早已被弃用，但是该文章却使得深度学习重新回到人们的视野。与此同时随着计算机技术的进步以及互联网的普及，其为深度学习技术提拱了爆发向前的燃剂。真正使深度学习得到广泛关注的是在2012年ImageNet比赛中，当时Hinton组凭借AlexNet深度神经网络以领先第二名10.8个百分点的优势一举多得比赛冠军，此后便进入了深度神经网络的时代。

### 2.2.2 注意力机制

注意力机制(Attention Mechanism)是人们在机器学习模型中嵌入的一种特殊结构，用来自动学习和计算输入数据对输出数据的贡献大小。本文以一个基于注意力机制的机器翻译模型为例，从人的直觉、中英文翻译的常识、特征工程等角度，对注意力机制的思想和机理进行了阐述；并介绍了一种常见的注意力机制实现形式，即基于感知机的注意力机制；还介绍了一种比较经典的注意力机制，即自注意力机制。注意力机制是上世纪九十年代，一些科学家在研究人类视觉时，发现的一种信号处理机制。人工智能领域的从业者把这种机制引入到一些模型里，并取得了成功。目前，注意力机制已经成为深度学习领域，尤其是自然语言处理领域，应用最广泛的“组件”之一。这两年为人广为人知的BERT、GPT、Transformer等模型或结构，都采用了注意力机制。当研究者们用机器学习模型时在情感分析任务上，通常会做一点特征工程的工作，即将原始文本转换为数值向量，其中特征工程，就是注意力机制在数据科学领域里的一种体现，它帮助模型选择有效、适当规模的特征，进而让模型可以有效、高效地完成任务。例如用逐步回归分析方法对原始特征集进行筛选，得到一个高质量的特征子集，就可以让下游模型聚焦于和任务关系最密切的信号。

而对于深度学习领域的注意力机制，一些学者尝试让模型自己学习如何分配自己的注意力，即为输入信号加权。他们用注意力机制的直接目的，就是为输入的各个维度打分，然后按照得分对特征加权，以突出重要特征对下游模型或模块的影响，这也是注意力机制的基本思想。在深度学习领域，模型往往需要接收和处理大量的数据，然而在特定的某个时刻，往往只有少部分的某些数据是重要的，这种情况就非常适合注意力机制发光发热。在深度学习的机器翻译任务中，以中译英为例，机器翻译是将一串中文语句翻译为对应的英文语句，其中主要的功能模块即编码器（Encoder）、注意力模块和解码器（Decoder）, 其中Encoder用于将中文语句进行编码，这些编码后续将提供给Decoder进行使用；Decoder将根据Encoder的数据进行解码。在中间过程，将某时刻的前一个时刻的状态和编码器输出进行注意力计算，得到一个当前时刻的翻译内容，其中注意力算法会有个打分函数，其大小描述了当前的时刻在编码器的结果上的关注程度，在此基础上的下一个时刻会更加关注源中文语句中的相应内容。

除此之外注意力机制也存在一定的变体，其中包括硬性注意力机制、键值对注意力机制、多头注意力机制以及自注意力机制。其中，在经典注意力机制章节我们使用了一种软性注意力的方式进行注意力机制，它通过注意力分布来加权求和融合各个输入向量。而硬性注意力（Hard Attention）机制则不是采用这种方式，它是根据注意力分布选择输入向量中的一个作为输出。这里有两种选择方式：选择注意力分布中，分数最大的那一项对应的输入向量作为注意力机制的输出。根据注意力分布进行随机采样，采样结果作为注意力机制的输出。这两种方式会使得最终的损失函数与注意力分布之间的函数关系不可导，导致无法使用反向传播算法训练模型，硬性注意力通常需要使用强化学习来进行训练。因此，一般深度学习算法会使用软性注意力的方式进行计算；另外是键值对注意力机制，该种机制对输出的要求是更为一般的键值对（K,V），并需要相关的查询向量Q，在这种模式下，使用查询向量Q和相应的键值K，进行计算注意力权值A，当计算出在输入数据上的注意力分布后，利用注意力分别和键值对中的值V进行加权融合计算出结果，当键与值相等时，该注意力方式就退化成了普通的经典注意力机制；还有就是多头注意力机制，其利用了多个查询向量Q，并行地输入键值对信息（K,V），在查询过程中，每个查询向量q将会关注输入信息的不同部分，即从不同角度上去分析当前的输入信息，最终向所有查询向量的结果进行拼接作为最终的结果；最后在自注意力机制中，其查询向量可以通过输入信息生成，而不用选择上一个任务相关的查询向量，即当模型读到输入信息后，根据输入信息本身决定当前最重要的信息。

### 2.2.3 频域图像处理

传统的图像处理方式是通过计算机对图像进行去除噪声、增强、复原、分割、提取特征等处理，随着人工智能的发展，大部分技术已被人工智能的子集——计算机视觉所取代。一般来讲，对图像进行处理（或加工、分析）的主要目的有三个方面：

（1）提高图像的视感质量，如进行图像的亮度、彩色变换，增强、抑制某些成分，对图像进行几何变换等，以改善图像的质量。

（2）提取图像中所包含的某些特征或特殊信息，这些被提取的特征或信息往往为计算机分析图像提供便利。提取特征或信息的过程是模式识别或计算机视觉的预处理。提取的特征可以包括很多方面，如频域特征、灰度或颜色特征、边界特征、区域特征、纹理特征、形状特征、拓扑特征和关系结构等。

（3）图像数据的变换、编码和压缩，以便于图像的存储和传输。

不管是何种目的的图像处理，都需要由计算机和图像专用设备组成的图像处理系统对图像数据进行输入、加工和输出。

图像是人类获取和交换信息的主要来源，因此，图像处理的应用领域必然涉及到人类生活和工作的方方面面。随着人类活动范围的不断扩大，图像处理的应用领域也将随之不断扩大。在航天和航空、生物医学工程、通信工程、工业与工程、军事公安、文化艺术、机器人视觉、视频与多媒体、科学可视化和电子商务等方面都有大量的应用。而对于肾脏超声图像，主要的目的便是提取超声图中所包含的特征与特殊信息，以帮助神经网络学习到更多维度的特征知识。深度学习模型中，主要还是通过卷积学习低级特征，比如边缘、纹理与幅值，但对于模糊性强、噪声高的超声图像，仅仅是低级特征不足以支撑模型最后的拟合任务与分类效果。需要让神经网络学习到频域内的高级特征……

### 2.3.2 归一化算法

### 2.3.3 正则化算法

相比常见用的正则化方法DropOut，即将每个隐藏的激活层乘以一个独立的伯努利随机变量，而当与批处理归一化结合使用时，DropOut会失去有效性{2ndPaper\_ioffe2015batch}。正则化算法为Drop Path{huang2016deep}

### 2.3.4 标准化算法

{touvron2021going}

## 2.4 本章小结

# 第3章 融合注意力机制和金字塔池化的图像语义分割网络

## 3.1 模型提出的动机

之前章节有提到过对于儿童先天性UPJO的病理检测，在诊断阶段以超声为主要手段，需要医生根据超声图像做出科学权威的分析，给出其对应的病理情况，决定手术与否与手术时机，然后国内面临医疗资源紧缺的情况，医学影像科的权威人才更是少之又少，在这一阶段引入大数据与人工智能技术下的计算机视觉以帮助，甚至在大部分医师短缺的情况下替代其工作职能，能够大大节省医疗资源与时间成本，帮助实现医疗工作的降本增效。其中面临的挑战就有超声图像本身的一些难点痛点，如图3-1-1所示，每一帧关于肾脏的超声图像会包括四类区域：病变区域（a）、器官区域（b）、其他组织区域（c）、非相关超声背景区域（d）。不难发现，对于疾病诊断最重要的a、b、c区域之间的边界存在模糊性，且图像噪声点多且密集，传统的图像处理方法或深度学习分类方法缺乏足够的相关信息、以及边缘信息，从而降低最终分类结果的准确度。基于以上问题，本研究提出一种先将肾脏超声图像通过语义分割算法，然后融合原图以增强边缘等信息，然后再通过分类算法输出诊断结果的智能诊断方法。本章首先介绍诊断方法中图像语义分割阶段的深度学习网络（Attention & Pyramid Pooling Network，APPNet)模型，经过实验与分析，本研究决定先通过专业的且经验丰富的医学影像科医生对超声图像数据集做人为的边界标注，然后基于这些标注的数据集训练一个语义分割神经网络，得到分割结果后，通过数据清洗与图像处理算法创建新的分类网络数据集，再完成之后的分类工作。



图3 - 1 超声图像视频中一帧图像，并用标注工具完成注释后的样例，成分包括a、b、c、d四类区域

针对如同3-1的四个类别的区域，在专业医师的帮助下，制作语义分割数据集，下一步就是选择什么样的语义分割算法，上一章节提到过各类经典的语义分割算法即从图像处理到机器学习到深度学习的一个发展过程，也列举了各算法在超声图像应用领域内的缺点。其实仔细观察图3-1还可以发现，随着病理严重程度增加，器官区域b内的病变区域a会越来越饱满，整张图像的其他无关区域比如c、d区域也会造成干扰，边缘信息更加难于识别，边界上下文更难以区分，需要分割算法具备一种注意力机制使得模型关注局部区域，同时还需要提供一种额外的算法分支能够让模型学习到多尺度的上下文信息，即学习不同尺度下、各感受野范围内不同类别之间关系的信息。因此，在基础特征提取阶段，融合注意力模块到特征提取网络中去，生成注意力权重图并使得在基础特征提取阶段的神经网络聚焦标注的病变、器官区域以及边缘信息；然后在高级特征处理阶段以金字塔场景解析方式做空间金字塔池化处理实现多尺度上下文的信息解析。因此对于肾脏超声图像分割，本研究提出了一个融合注意力机制和金字塔池化模块的新型金字塔场景解析网络APPNet，它能满足既定应用场景下、专注于关键部分，聚合上下文信息的能力，并很好地完成语义分割任务。其核心是一个融合了注意力模块的基本的特征提取网络和高级特征处理阶段的金字塔池化模块。整体结构如图3-2所示，详细解释如下跟随整体网络下面本文将详细地从数据输入到输出的方式端到端地介绍该网络的各个模块。

C:\Users\Administrator\Desktop\毕业论文\picture\分割图.tif分割图

图3 - 2 APPNet的整体架构，输入为原始超声图像，输出为红绿黑颜色区分的分割结果。a. 该模块为ResNet32网络并在前后端嵌入两个NAM注意力模块,即f，即；b. 基础特征阶段所提取的特征结果；c. 金字塔池化处理的高级特征；d. 对c进行统一维度运算后的结果；e. d和b集成的最终特征图，经上采样并预测后输出结果。

## 3.2 基础特征提取模块

对于语义分割任务，由于其分类级别是基于像素的，低级别的图像信息往往不够，例如传统图像处理中，为了得到某些信息（例如边缘信息），需要对图像进行滤波处理，滤除不是边缘的内容，从而得到边缘特征。例如还进一步地需要纹理信息，通过卷积神经网络将这些低级特征提取出来。而像素级分类是需要高级特征的，即假设有两张图片分别为猫和狗，只用上述低级的纹理、边缘特征是无法区分其中的耳朵、眼睛、鼻子等特征区域是否属于猫类别还是狗类别，因此需要通过环境信息等高级特征去结合低级特征并给出一个综合分析的结果。

目前人工智能在欠缺可解释性的大背景下，由实验验证网络的深度对模型的性能至关重要，当增加网络层数后，网络可以进行更加复杂的特征模式的提取，所以当模型更深时理论上可以取得更好的结果，网络深度增加时，网络准确度出现饱和。在本研究的基础特征提取阶段，医学超声图像受本身成像原理的限制,致使其图像的分辨率不高而且噪声污染严重,在深度网络的传播过程中随着维度的降低，饱和率增大、易出现过拟合、网络退化等问题，而ResNet可以在保留深度网络均值的同时避免网络退化的问题。它还具有低复杂性，相比传统的基础特征提取网络具有更少的参数量，这意味着处理速度更快。关于 ResNet 的结构，其基本构成单元为瓶颈模块，具体结构如图3-3所示。

C:\Users\Administrator\Desktop\毕业论文\picture\基本块.tif基本块

图3 - 3 ResNet的基本构成单位，主要由卷积、归一化、激活层和残差连接组成

深度网络的退化问题说明深度网络不容易训练。但是假设我们现在有一个浅层网络，想通过向上堆积新层来建立深层网络，一个极端情况是这些增加的层什么也不学习，仅仅复制浅层网络的特征，即这样新层只完成了的恒等映射。在这种情况下，深层网络应该至少和浅层网络性能一样，也不应该出现退化现象。事实上，基于对生成模型不断地实验后，发现了在训练的过程中发生了梯度消失的问题。ResNet提出了用残差学习来解决网络退化的问题，对于一个由图3-3所示基本单元堆积层的结构块，当输入为时，则网络学习到的特征记为, 而ResNet希望其可以学习到残差，从而原始的学习特征为。这样做的优点是残差学习相比原始特征学习更加容易，因为当残差为0时，会使得堆积层在输入特征基础上学习到新的特征从而获取更优性能，从数学的角度分析为何残差学习，首先将残差单元定义为:

(3-1)

其中和分别表示的是第n个残差单元和输入和输出，为残差函数即学习到的残差，表示恒等映射，表示激活函数，基于上式可得到从浅层神经网络 到深层神经网络的学习特征为：

(3-2)

通过链式求导法则，定义损失函数到达深层神经网络时的梯度为，结合公式3-2计算偏导，那么反向传播过程中的梯度为：

(3-2)

由公式3-2可发现残差连接的短路机制可以无损地传播梯度，即括号内的1，即使主干的残差梯度即比较小，有1的存在也不会导致梯度消失。对于超声图像，由于分辨率有限，我们在任何时候都会接收到分辨率单元内的大量分布的散射体的散射信号，这些散射信号相干相加，或者说它们根据每个散射波形的相对相位进行相加叠加和相消叠加。图像中就出现亮点和暗点不规则相间分布的信号。这样的散斑噪声带来的负面影响就是在局部区域随着网络深度的不断加大，梯度消失的概率会逐渐增加，因此采用ResNet网络作为基础特征提取模块，利用其残差连接的优点弥补超声图像自带的噪声所带来的缺陷。本研究在其基础上融合了注意力机制，另外考虑到在本研究中肾脏超声图像的分割任务中，需要更多地识别一些不显著的特征，通过使用基于规范化的注意力模块(NAM),它抑制了较少显著性的权值，对注意力模块应用一个权重稀疏惩罚，相比其它注意力机制缺乏对权重的影响因素的考虑并抑制了不显著的像素，它通过利用预训练模型权重的方差度量来突出显著特征，其整体结构如图3-3所示。

C:\Users\Administrator\Desktop\毕业论文\picture\NAM.tifNAM

图3 - 4 NAM的结构图，分别有通道注意力与空间注意力两种模式

对于通道注意力模块，以作为输入特征、为输出特征，权重值等于使用批归一化（BN）中的缩放因子，如公式3-3，

(3-3)

其中为均值，为标准差，其平方根表方差，和表示可训练的仿射变换参数，对应尺度和位移的变换，利用BN的缩放因子反映出各个通道的变化的大小，也表示了该通道的重要性,缩放因子即BN中的方差，方差越大表示该通道变化的越厉害，那么该通道中包含的信息会越丰富，重要性也越大，而那些变化不大的通道信息单一且重要性小，通过这种方式对各个通道的重要性给出分值作为注意力层的输出。而对于空间注意力模块，以作为输入特征、为输出特征，权重值等于 ,将BN的比例因子应用于空间维度来衡量像素的重要性，并称之为像素归一化，即给各个像素的重要性给出分值并作为注意力层的输出，具体计算方式分别如下：

(3-4)

其中ReLU表示激活函数。对于以上描述的注意力模块，其与ResNet的具体融合方式不同于传统的方式，即在每个卷积层之后加入注意力层，因为这种方式有两个缺点，一是会过高地提高网络的参数量，二是会破坏原有网络的结构而不利于预训练参数的初始化，于是研究并思考了一种新的在不破坏ResNet原有结构的情况下添加NAM的巧妙方法，即在主干网络的整体输入输出两端分别前后加入通道和空间注意力单元，由于NAM是一个轻量级的通用模块，这种嵌入方式几乎无需任何额外的计算开销，经实验验证两种嵌入方式的最终的准确率、召回率等指标几乎一致，为了更好地介绍ResNet与注意力层的结合原理，对于给定的特征图作为输入，对于通道注意单元和空间注意力单元, 其整体的注意力特征处理过程可以概括为：

(3-5)

其中表示逐像素乘法，分号表示串联。同样的，对于输入特征，将ResNet主干网络的特征处理过程定义为，结合公式3-5，其最终的基础特征提取模块处理过程可以概括为：

(3-6)

其中和分别表示NAM在ResNet中的嵌入顺序，并且都由通道注意力单元和空间注意力单元串联组成。以上即基础特征提取模块，其作用是初步提取低级的超声图像的特征如纹理、边缘等特征信息，其主干网络为ResNet32，以残差连接的方式消除了网络退化的问题，并其基础上嵌入了NAM注意力层，赋予主干网络在特征提取阶段更多地关注非显著特征的能力，并将提取的初步特征为下一步高级特征提取模块做输入准备。

## 3.3 高级特征提取模块

上一节提到过对于语义分割这样的像素分类任务，低级的纹理等特征信息是不够的，需要知道像素间的关系等高级特征信息，以此来提高像素分类的准确性。而完成这一目标，需要通过高级特征信息解决语义分割的传统的三个难点，一是环境下的不匹配导致的错误识别；二是难以识别的不明显的类别（例如过大儿超越感受野的物体或者体积微小难以区分识别的物体），三是相近的类别难以统一而混淆。如图3-5所示，图中的a所表示的对比图即在水上行驶的真正的船和在岸上未使用的船，同样是船却需要上下文的环境信息去区分；图中b所表示的对比图即不同尺度下的人脸难以识别的问题，若对两张图片做人脸分割，将会因为极端尺度的问题而导致特征难以识别；图中c所展示的即相同的类别易混淆、难统一的问题，二者同样从人工观察上看都是房子，其中左图具备墙面、右边只有半面墙，若需要对其从计算机视觉的视角去区分的话，同样也需要结合不同感受野去学习其易混淆的相关特征。

C:\Users\Administrator\Desktop\毕业论文\picture\语义难点.tif语义难点

图3 - 5 语义分割中的难点样例图，a. 岸上的船和海上的船；b. 小尺度人脸和大尺度人脸；c. 有墙的房子和不全带有墙的房子

许多错误都与不同感受野获取的全局信息和语境关系有着部分甚至是完全的关联，因此，一个拥有适当全局信息的深度网络可以大大提高场景解析的能力。学界将感受野定义为使用上下文信息的大小，通过不同大小的感受野引入更多的上下文信息进行解决，当分割层有更多全局信息时，出现误分割的概率就会低一些。因此本研究在基础特征提取结束过后，对特征图进一步处理，并使用金字塔池化结构作为高级特征处理模块的主干结构，具体过程如图3-6所示。

高级特征提取

图3 - 6 高级特征提取的过程，先经过不同尺寸的池化，然后通过卷积调整通道数量，最后上采样并拼接结果输出高级特征图

从图3-2中可以看到，在基础特征提取模块处理过完原图过后生成的低级特征图b过后，会产生分支，金字塔池化模块融合了四种不同尺度下的特征。。其中对于图3-2中的c模块，即金字塔池化模块，就是一种充分利用全局信息的方式，结合图3-6的细节过程展示，首先对特征图分别池化到不同大小的特征块儿，且池化大小以此为1×1、2×2、3×3、6×6、如图3-6所示例如由2048个通道的、30×30像素大小构成的特征图经过4层金字塔模块池化后分别生成30×30、15×15、10×10、5×5的结果，因为这里相当于产生了4个分支，相比于原通道数2048多了4倍的特征块，为了压缩通道数，对结果分别进行1×1卷积将通道数减少到2048的，这里的N大小为4即金字塔模块池化层的数量。之后为了统一像素尺寸大小，将四个分支的高级特征通过双线性插值的图像处理算法将四个高级特征采样到同样大小的像素尺寸，最后如图3-2所示，将高级特征c与低级特征d拼接生成融合细节特征与全局特征的输出，最后的通过一层卷积生成最终的预测结果。

以上小结较为详细地介绍了基于肾脏超声的智能诊断方案的图像语义分割阶段过程，我们从原始的超声图像输入到最终语义分割结果的输出进行详细地过程描述，首先利用融合了注意力机制的基础特征提取模块提取超声图像的低级特征，然后利用金字塔池化算法在低级特征的基础上提取高级特征，通过一些图像处理的算法统一尺寸并与低级特征组合成最终预测特征图，然后输出预测结果。

## 3.4 实验与结论

本小节对融合注意力机制和金字塔池化的图像语义分割网络进行实验评估，实验建立在首都医科大学附属儿童医院提供的肾脏超声图像数据集上，首先简单的描述我们的数据集，以及实验的设置与参数配置，之后根据所设计的实验来评价本研究所提出图像语义分割算法，最后分析实验结果并评估该模块的作用。

### 3.4.1 数据集与实验设置

我们所有的实验都是在同一个 UPJO 数据集上进行的，该数据集包含来自 17 名患者的 1850 张超声图像。 源数据由北京儿童医院于2019年9月10日至2020年3月10日收集，由专业超声医师在肾脏冠状面和横切面（标准位置）截取的174张截图组成，另外还有120 个扫描仪存储的视频，记录超声波扫描期间超声波技师的操作。 将所有源视频以大小为5的帧速率切成图像，并且通过数据清洗，去掉脏数据、包括模糊而无法区分的图像。 然后经验丰富的医生使用图像标注软件—Labelme在每张图像上标注病变区域和肾脏区域。 最后，经过这样的预处理得到了 1850 张带注释的 UPJO 超声图像，目前胎儿肾积水的诊断标准使用最多的是胎儿泌尿外科协会(Society of Fetal Urology)制定的先天性肾积水分级系统，其中病理等级分为了 的共5级(SFU，0~4级)共5级。本文对模型和算法进行了对比试验，并且在注意力机制的融合设置消融实验，均使用了相同的实验环境，表3-1列出了本次实验环境的具体配置，包括实验室用的计算机配置。实验的模型训练的开发与调试工作主要在Windows操作系统下进行，代码主要基于Python的深度学习框架Pytorch。

表 3 - 1 实验环境

|  |  |  |
| --- | --- | --- |
| 环境 | 参数 | 具体配置 |
| 硬件环境 | CPU | Intel(R) Core(TM) i7-10700F 八核2.90GHz |
| GPU | NVIDIA GeForce RTX 3070 |
| 内存 | 32GB |
| 显存 | 8GB |
| 软件环境 | 操作系统 | Windows 10专业版 |
| 开发环境 | JetBrains PyCharm |

我们将标注好的1850张图像数据分为比率为9：1的训练集与测试集，采用批量训练并将批量规模设置为8，训练世代为200，训练过程加入Adam优化器，初始学习率设为0.005，并设置学习下降的倍率为0.9，频率为1世代的方案更新学习率，以平均交并比作为模型结果的评价指标，输入分辨率为473×473，因此，在数据加载的阶段，我们将重置图像的大小。由于图像的长度和宽度不相等，为了避免失真，我们通过添加灰度边框将图像填充成一个正方形，然后重置大小。然后，基于迁移学习[IEEE Transactions on Knowledge and Data Engineering]的思想，利用预训练的模型的权重进行优化，从ImageNet上的公共数据集生成的学习过程。此外，还对其他基于ResNet的特征提取网络模型进行了比较训练，并保持了相同的学习速率、优化器和训练代数。我们所有的实验都是使用Intel CPU和NVIDIA GPU进行的。由于我们的GPU内存大小是8GB，因此我们使用批处理训练，批处理大小设置为8。然后，我们使用自适应矩估计（Adam）优化器（一种常见的优化算法）来训练我们的模型直到收敛。损失函数上结合了Dice Loss和Cross Entropy Loss作为组合损失函数，补充定义真实值为，预测值为，损失值为，其计算公式为：

(3-7)

为了完整地描述一个特定的深度学习模型的性能，除了准确性外，还应该考虑模型的复杂性，如参数的数量和计算量。为了提高其在精度和速度上的性能，我们在分割模型的中分别在顶部和底部位置分别集成了独立的注意力模块。为了找出基础特征提取模块种注意力的效能，以及其子结构、空间域与通道域注意力单元的位置对整体结果的影响，我们分别对两个单元和两个串联组合进行了比较实验。结果如表3-2所示，以经过50个世代训练的最终模型（所有实验中其余参数均一致）的MIoU和MPA作为评价指标。此外，为了确认我们提出的基于注意力的ResNet的网络结构比原始的ResNet轻，我们比较了参数的数量，并计算Giga级的定点每秒乘的累加运算量（GMACs）来测量表3-3中每种方法的计算复杂度。

### 3.3.2 消融实验与分析

我们进行了消融实验来比较NAM中四种注意力单元的排列，通过删除部分结构和改变顺序来研究网络的性能从而得到最优的网络结构，并得到结果如下：

表 3 - 2 注意力单元排列组合的对比结果

|  |  |  |  |
| --- | --- | --- | --- |
| 实验序号 | 描述 | MIoU | MPA |
| 1 | ResNet50 + 空间注意力 | 88.17 | 94.03 |
| 2 | ResNet50 + 通道注意力 | 88.16 | 94.05 |
| 3 | ResNet50 + 空间 - 通道注意力 | 88.39 | 93.78 |
| 4 | ResNet50 + 通道 - 空间注意力 | 88.53 | 94.63 |

如表3-2所示，不同的模型采用了不同组合的混合域的注意力单元，我们对各模型有效性做了MIoU和MPA两个维度上的评估。通过比较发现每种方法的性能差距并不明显，其中MIoU的最大差值仅为0.27。相对而言，实验4的最终模型效果最好。单独集成空间域或通道域中的某一个注意力单元的情况下，表现均劣于混合域组合的情况，且各自对模型的影响基本一致。混合域集成注意力模块的情况下，实验4在平均交并比上略优于实验3，仅相差0.14个百分点。在像素准确性上，实验4明显优于实验3且相差0.85个百分点。综上可得出实验4的方案使得最终模型的有效性最高，之后小节的对比试验也是在实验4方案基础上，与各个图像语义分割模型完成多个指标维度上的比较分析。

### 3.3.3 对比实验结果与分析

这一节我们将从准确度、复杂度两个维度去对评估我们的最终模型APPNet，对比一些经典语义分割模型的计算性能。对比最终分割模型与基础特征提取模块，并且评估了各经典基础特征提取模块的可训练的参数量和计算量，包括评估了基于表3-2最优结果的基础特征提取模块的可训练的参数量和计算量。首先，我们选择了几个在基础特征提取网络模型上同样采用ResNet [Dual attention network for scene segmentation, Rethinking atrous convolution for semantic image segmentation]的图像语义分割网络，以及几个现有的流行的图像语义分割网络[FCNs,RefineNet]。并对所有深度学习网络进行了参数一致、数据集一致、训练过程一致的对比实验。选择最优的分割模型与我们的NAM模块相结合，作为最终的图像语义分割网络并再次进行训练。

表 3 - 3 有效性的比较

|  |  |  |  |
| --- | --- | --- | --- |
| 实验序号 | 方法 | MIoU | MPA |
| 1 | PSPNet | 87.61 | 92.01 |
| 2 | Deeplab\_v3 | 85.47 | 91.58 |
| 3 | DANet | 80.54 | 83.14 |
| 4 | **APPNet** | **88.93** | **93.52** |
| 5 | FCNs | 81.66 | 82.46 |
| 6 | RefineNet | 84.79 | 88.68 |

结果如表3-3所示。对比实验2、3、4，结果显示，在基于ResNet作为基础特征提取网络的前提上，PSPNet的有效性最佳，因此将其与两个轻量级注意单元相结合，对其进行改进然后构建了APPNet，并进行对比实验来彻底评估最终模型的有效性。我们验证了融合NAM的ResNet的性能优于其它没有融合NAM模块的语义分割网络，即对比实验1和实验4，结果显示，与原始的PSPNet相比，其性能提高了1.32个百分点。对比实验4、5、6，可以发现与其它的不同种类的语义分割算法相比，在肾脏超声数据的场景下，我们的模型仍然具有优越性。

除此之外，我们还分别评估了空间注意单元和通道注意单元的计算参数，以论证它们的轻量性。将我们构造的基础特征提取网络作为基线，对比各经典基础特征提取网络模型和基线的复杂度。

表 3 - 4 复杂度的比较

|  |  |  |  |
| --- | --- | --- | --- |
| 实验序号 | 描述 | Params(M) | MACs(G) |
| 1 | ResNet50 | 25.55703 | 18.84624 |
| 2 | 通道注意力 | 0.00020 | 0.00004 |
| 3 | 空间注意力 | 0.52480 | 0.00378 |
| 4 | **Ours** | **26.08203** | **18.85144** |
| 5 | ResNet101 | 45.07416 | 35.96785 |
| 6 | ResNet152 | 60.7178 | 53.10086 |
| 7 | AlexNet | 61.10084 | 13.09258 |
| 8 | VGGNet | 138.35754 | 68.2407 |

我们计算了空间注意单元和通道注意单元的复杂度，即实验2和实验3，实验结果如表3-3所示，与基线相比，空间注意单元小了5个数量级的参数量，通道注意单元相差了2个数量级的参数量，分别相差了6和4个数量级的计算量，由此分别证明并展示了他们的轻量性。总体上，对比实验1和实验4，基线模型和原始结构相比，参数量增加了约2%、计算量增加约了0.03%。对比实验4、5、6，即ResNet同族模型见比较，无论是可训练参数量，还是计算复杂度都比融合了注意力模块的ResNet50高。对比AlexNet，实验7表明其参数量约为的基线的2.4倍，但计算量略小于基线。对比VGGNet, 实验8表明其参数量约为基线的5.3倍，计算量约为基线的3.5倍。因此从计算复杂度的角度去分析，融合了注意力机制的基线模型同样具备轻量级特性，此外，结合表3-3实验1和表3-4实验1，以及表3-3实验4和表3-4实验4，对比结果分析，模型增加了2%参数量，和0.03%的计算量，最终训练模型结果的MIoU提高了1.32个百分点，MPA提高了1.51个百分点，可以发现前后模型在增加了极少的复杂度的情况下，有效性明显提高了。

综上所示，我们评估了模型的复杂度和有效性，最终验证了在极小的开销下我们的模型的性能有了一定的提升，也印证了注意力机制在肾脏超声图像的应用场景下，能够有效地提高预测准确率，接下来将从抽象层面到可视化层面进一步分析注意力机制在超声图像应用场景下的实际效果。

### 3.3.4 注意力可视化分析

从可解释性的角度阐述，为了更加抽象地理解注意力机制的作用，分别从超声扫描过程中的横截面与冠状面的数据中，挑选了左、右肾的图像数据通结果，并可视化了注意力层之后在数据集中的权重热力图，具体如图3-所示，颜色从小（蓝色）到大（红色），波长越长即表示权重越大。



图3 - 7 左肾右肾的横截面和冠状面，共四张样本图，左侧为分割结果，右侧为权重热力图

理论上注意力机制得到的权重乘以输入特征图可以帮助网络进行自适应特征细化，权重数值表征了特征图该点的重要程度，通过感受野反推至原图像，即表示了该区域的重要程度，观察图3-7可以发现，网络的权重主要集中在病变区域，其次是肾脏区域，再其次就是其它的无关区域，包括超声图像自带的黑色背景的权重都很低，由此可证明，注意力机制能够让网络自适应关注需要关注的地方，从而使网络按照我们的期望，学习到更加重要的特征信息，并尽可能忽略掉无关区域的特征分量。

3.5 本章小结

# 第4章 基于离散小波池化增益的图像分类神经网络

## 4.1 模型提出的动机

之前章节有提到在诊断阶段分为语义分割阶段和图像分类阶段，前一阶段是为了增强肾脏超声图像的边缘信息，为最后的分类任务提供特征信息更丰富的数据集，经过上一章节所介绍的语义分割算法过后，我们获得了一个可以将原始超声图像分割成背景区域、器官区域与病变区域的分割模型，并且在图像分类阶段，我们通过语义分割模型补充分割了新的数据集，将分割结果与原图按比例叠加后作为图像分类神经网络的新的数据集。这一阶段的数据集仍然具有超声图像自带的白斑噪声，且器官的纹理特征等信息也与白板噪声高度耦合。从通信工程中的数字信息处理中可以思考这种白斑噪声从频域上处理，因此在图像本身是一种二维的数字信号载体的前提下，我们试图从数字图像频域处理的角度去挖掘不同维度中的特征信息。

图像中的频域处理算法主要有傅里叶变换和小波变换，傅里叶变换是全局的变换，函数的各个点的值都对变换后的结果有影响。而小波变换使用的是局部基，对于指定基的系数，函数对该系数是否有影响取决于其是否在当前基所支撑的点上。这就导致小波不仅包含频率信息（由基的频率反应出来），也包含时间信息（由基所在的局部时间轴位置反应出来），在图像中级即表现为位置信息。小波的尺度收缩特性使得小波具有了分形的特性。这种特性能够小波捕获了信号和数据的局部相似性。这种能力可以用来压缩和特征提取，这是傅立叶所没有的，小波也是有缺点的，在对信号频域分析的时候，由于小波使用了下采样，违反了柰圭斯特准则，造成了频谱的混叠，因此我们仅仅在网络中对不同的小波变换下采样结果中的在通道维度上叠加并进行神经网络深度学习的过程而不做小波变换的逆变换过程。也正是这里小波变换带来的下采样效果，使之可以完全替代一般模型前前端的池化操作，这里传统的池化操作有最大值池化、均值池化等算法，会丢失部分图像特征信息，对于一般图像，这种操作对分类结果几乎没什么影响甚至能节省网络计算复杂度，但对于超声图像这种非显著特征的密度较大的图像来说，池化操作会大大降低最终的分类精度，因此我们尝试在目前计算机视觉上已有的表现优异的网络模型中尝试这种替换方式来达到更好的分类效果，并构建了结合小波池化增益层的网络模型，命名为WaveConvNeXt。

## 4.2 神经网络中的离散小波变换

正如之前所说，为了对超声图像做病理分级判断，我们在智能诊断系统中设计添加了分类器，在分类网络中池化层以2倍下采样的方式分解图像，其目的是为了减少总体计算量，但是同样带来的还有信息量的损失，既需要降低复杂度，又不想损失信息量，所以我们在分类器中加入了一种数字信号处理的方法来取代传统的池化层以折中之前所述之难题。从数学角度来讲，图像特征是具有局部变化统计特性的亮度值的二维阵列，如边缘和反差鲜明的同质区域，以多个分辨率来表示图像的结构（也称图像金字塔）非常有效且概念简单，与其相关的重要技术就是子带编码，将图像分解为一组频带受限的分量，也称子带，子带可以重组用以复原图像，在特征维度上也可以表示图像中人眼难以捕捉的信息。在深度学习模型中，我们不需要使用子带以复原图像，但对于计算机而言，这种信息的捕捉与学习十分容易。如前所述，我们添加了一个分类器来获取 UPJO 病理分级。 在传统的分类网络中，池化层采用下采样来分解图像，减少总的计算量，但会导致信息丢失。 由于DWT[ sci-32]及其逆运算是可逆的，从而保证了在降低复杂度的同时不丢失信息。 DWT被认为是分解图像并获得所需频域信息的有效算子，因此被应用于分类器以取代传统的池化层。为了方便理解替换过程的原理，我将从一维的角度去讲解，首先定义两个基函数，一个和尺度函数，对于一维的离散序列，其正向的离散小波级数展开系数为

(4-1)

(4-2)

在以上等式中，和 是在 尺度下的基函数的支撑区上使用 个等间隔的取样，通常，变换本身由个系数组成，最小尺度为0，最大尺度为，这里再补充定义

, (4-3)

决定既定方向上的位置， 决定取样的宽度， 为限定的任意开始尺度，则得到的 是的一个子集。

与一维DWT类似，二维DWT使用二维尺度函数和小波函数。 我们先对二维数组的行进行一维变换，然后对上一步结果的列进行一维变换。结果，然后再取结果的列的一维变换，但在二维情况下会得到水平、垂直和对角细节系数，对于大小为 图像，其DWT是

(4-4)

(4-5)

最终得到4个1/4大小的输出子图：，分别表示原图近似分量和水平、垂直和对角上的细节分量。我们将这四个分量构成的网络层命名为小波增益池化层，相比传统池化层的效率换时间的特性，这种小波增益池化层缓和了直接池化带来的信息损失，从特征工程的角度理解，通过这些人工设计的细节分量把较低级别的特征表示成更加抽象的特征，提高了表示学习下，输入数据的转换层次，利用深度神经网络模型的自动提取特征的能力，使模型从频率域中学习到更加抽象的特征。

## 4.3 图像分类网络-WaveConvNeXt

该分类模型在ConvNeXt基础网络上，添加了小波增益池化层，其中ConvNeXt参考了ResNet的多阶段设计思想，对于其核心宏观结构，以为ConvNeXtBlock为单元，在每个各个阶段具有不同的特征分辨率，各阶段且计算比率 调整为（3：3：9：3），。考虑到图像的信息冗余性，需要对输入图像下采样并聚合到适当的特征尺寸，其中Sampleblock起到至关重要的降采样操作，通过归一化用来保持模型的稳定性，并以步长与尺寸相同的卷积核实现无覆盖的卷积运算从而实现降采样过程。对于ConvNeXtBlock, 相比于ResNet使用具有更大感受野的卷积核、更少的激活函数 ，并且使用LayerNorm以替换表现欠佳的BatchNorm。



图 4 - 1 WaveConvNeXt模型的框架构成，其卷积运算中的K表卷积核大小，S表卷积核滑动步长，P为填充大小，ConvNeXt Block中的C为输出通道数，激活函数为GELU，归一化算法为LayerNorm，标准化算法为Layer Scale, 正则化算法为Drop Path，最终输出正常、轻症、重症三类结果

如图4-1所示，我们在首次下采样阶段用替换为小波池化增益层，并对高频分量做一个通道剪枝，以达到后续的维度一致性。模块内使用Layer Scale辅助优化，其是一种为了保持各层值和梯度量级的一致性的标准化方法，通过加入一个可学习的缩放参数矩阵以解决残差连接带来的方差增大的问题并提高训练效率。以表示可学习的缩放参数矩阵，表输出， 表输入，其计算公式如下：

(4-6)

其中表均值，表样本方差，为一个无穷小参数以防分母为零，为偏移量参数。模块内还引入DropPath正则化方法。DropPath通过加入可学习的超参数，即存活率，以自动化寻找Dropout模式。以表存活率，通过一个简单的线性衰减规则，从表示对于最初输入的存活率开始，对于最后一个ConvNeXtBlock的存活率，其过程中的存活率计算如下，

(4-7)

其中 表计算所至的次序，为计算总次数。通过该方法从而可以提高网络鲁棒性和泛化能力，并提高训练速度。

## 4.4 实验与结论

### 4.4.1 数据集与实验设置

在实验环境上，因为我们所有的实验主要分为分割阶段与分类阶段，因此实验环境与第三章的3-1表格所示的实验环境一致，且都是在UPJO数据集上按处理顺序进行的。上一章节实验完成了整个智能诊断系统的分割阶段，最终产出了一个肾脏器官、病变区域、无关区域三类分割的图像语义分割模型，所有所用数据集包含了来自17名门诊病人的3289张超声波图像，并由具备专业医学影响技术的医生采用标注工具对病变区域和肾脏通过打点环绕的方式标记并注释了1850张图像用于分割阶段，由经验丰富的医生对所有共3289超声波图像进行病理分级，其中分级标准采用了胎儿泌尿外科协会制定的先天性肾积水分级系统。其中在训练出具备良好准确度的分割网络之后，加入之前未标注的所有剩余图像（1439张），经过训练好的图像语义分割模型处理并输出，现定义图像语义分割结果为S、原图为F，然后定义输出的融合图像结果为F.I.，合并算法如下：

(4-8)

并将病理分级完毕的3289图像结合医生的需求，再分为SFU 0-2、SFU 3和SFU 4共三类，即正常、轻症合重症。将分类好的3289张数据经过分割器，以尽量保留器官区域的原则裁剪不必要区域，结果大小为810×608为并分成比率为9：1的训练集与测试集，这一阶段加入了对比实验以证明我们的分类网络的优越性，并以准确度（Accuracy）作为评价标准，对于分类模型与大小为 的测试集，指标值 的计算方法为：

(4-8)

所有对比试验除初始学习率为以外，其它参数一致，使用批量规模为32的批量训练，训练世代50，加入Adam优化器并设置衰减系数为0.9、频率为1世代的方案更新学习率，对于最终的分类结果：正常、轻症与重症，数据集中各个分类的数据量并不平衡,针对数据集的不均衡问题，以各分类的数据量间的比值作为权重系数加入到损失函数的计算中去，并以此算法作为整个训练过程的损失函数。整体训练网络的输入是大小为224×224的图像，而对于我们目标网络中的小波变换，采用哈尔基小波变换，并采用标准分解方法，首先使用一维小波对图像每一行的做变换处理，再同理对每一列进行变换，进而生成产生近似分量与细节分量，其中近似分量很好的替换了池化层的输出，为了印证这一效果，我们做了对比实验，即在其它网络上以小波变换的近似分量替代池化层输出并训练，训练世代30，其它参数一致。

### 4.4.2 对比实验与分析

在上一章节，我们进行了实验并选取了MIoU最高的分割模型作为我们分割阶段的图像语义分割器，将分割结果与原图合并然后裁剪作为分类阶段的输入。参考公式4-8，我们在实际实验过程中取α等于0.1，总共将3289张图像通过语义分割算法并完成图像融合，具体过程如图4-2所示。



图 4 - 2 图示为分割阶段的处理过程，包括语义分割过程和融合图像过程

在分类器中，我们在不改变特征维度的情况下结合了离散小波变换，获取了与2倍下采样效果一致的近似分量，和具备高频信息的细节分量，增强神经网络模型能力，帮助神经网络从频域学习到更抽象的纹理等特征信息，其中维度由最初的64×56×56增加到256×56×56，原池化过程后的特征图如图4-3所示

图 4 - 3 图示为原始网络中在经过池化层后的数据集中随机选取的超声图像的特征图，维度为64×56×56

将池化层更改为由小波池化增益层之后，将原始池化层的输出替换为近似分量，然后并行添加细节分量的网络路径，具体特征分别如图4-4所示。



图 4 - 4 a.小波池化增益层后近似分量的特征图，单张维度为56×56； b. 从水平、垂直和对角线三个细节分量中随机选择的特征图，其总数为192。

图4-3和图4-4.a相对比，离散小波变换的近似分量是池化操作的有效替代分量，从图4-4.b可以看出，新的特征分量包含密集的高频噪声信息和轮廓信息。

表 4 - 1 各分量网络分类的测试的结果和我们的模型

|  |  |  |
| --- | --- | --- |
| 模型 | 准确度 | 迭代阈值 |
| AlexNet | 84.45 | 35 |
| VGG | 86.36 | 27 |
| GoogleNet | 84.75 | 38 |
| ResNet | 89.94 | 38 |
| ConvNeXt | 91.65 | 49 |
| **Ours** | **93.47** | 50 |

除此之外，我们还对6组分类网络的性能，在数据集以及参数与超参数的设置均一致的情况下，进行了对比实验。具体结果如表4-1所示。表中给出了不同网络训练过程中的最优精度模型和训练迭代阈值。由此我们观察到，ConvNeXt在四个基准网络中具有更好的准确性，我们提出的WaveConvNeXt模型更是优于所有模型，这证明了我们选择ConvNeXt的正确性。



图 4 - 5 在经典的基础分类网络上，用小波池化增益层替换池化层后的性能

另外，我们还尝试在基于表4-1的各分量网络上做了同样的小波池化增益层替换传统池化层的操作。由于过多的池化层会大大增加网络的复杂度，因此我们只选择和测试了几个池化层在重要位置上的网络，从而剔除了VGG网络。以30迭代阈值训练其它基础分类网络在采用小波池化增益层前后的模型，然后收集训练过程中准确度曲线变化。如图4-5所示，用小波池化增益层替换池化层，增加特征提取的维度与数量，可以提高所有网络的性能，其中仍然以WavaConvNeXt为最佳。进一步证实了我们所提出的方法的优越性与有效性。



图 4 - 6 ConvNeXt模型与我们提出的模型之间的训练曲线对比图

由于我们提出的分类网络的方法是基于ConvNeXt的，因此我们将以相同实验条件下的训练过程的曲线图为切入点，进一步挖掘小波池化增益层的其它优越性。如图4-6所示，从曲线的振动振幅和频率来看，在约迭代次数从1-30范围内，原网络的震荡幅度明显高于我们的模型，可以证明我们的模型相比原模型，具有更加鲁棒的训练过程和更高的性能。再观察模型训练的后半段，原模型的损失曲线出现了回升的情况，且准确度曲线趋于平滑，意味着训练过程陷入了局部最优解，甚至是过拟合，相比之下，我们的模型损失曲线稳定下降并趋于平滑，证明模型接近饱和状态。以上分析进一步证实了我们的网络结构的优越性。基于以上结果，用小波池化增益层合理地替换池化层，将从频率维度上增加特征的数量与特性，使训练过程更加稳定、模型学习的特征更加鲁棒，模型的性能更加优秀。

# 结论

学位论文的结论作为论文正文的最后一章单独排写，但不加章号。

结论是对整个论文主要成果的总结。在结论中应明确指出本研究内容的创造

性成果或创新点理论（含新见解、新观点），对其应用前景和社会、经济价值等

加以预测和评价，并指出今后进一步在本研究方向进行研究工作的展望与设想。

结论内容一般在 2000 字内。

# 参考文献

# 攻读硕士期间的主要研究成果

# 致谢