• 密 级 公开

桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

题目 改进的病理学图像分类多示例学习方法研究与应用

Research and Application on Improved Multiple-instance Learning Method for Histopathological Image Classification

研 究 生 学 号: 19032201043

研 究 生 姓 名: 叶天鸽

指导教师姓名、职称: 罗笑南 教授

申 请 学 位 门 类: 工学硕士

学 科、专 业: 计算机科学与技术

论 文 答 辩 日 期: 2022 年 6 月 2 日

摘要

在使用传统深度学习进行病理学图像分类时，研究人员经常会因为需要对病理学图像进行大量标记而耗费大量时间和人力。为了降低标记图片的时间和人力成本，本文对现有解决方法进行了研究。在进行组织病理学图片分类时，因为全视野数字切片(Whole Slide Image, WSI)非常巨大，人们会将 WSI 裁剪为很多个小的图像块。在裁剪图像块的同时，医生会对每个图像块进行标记。这样就能将这些图像块放入神经网络中进行训练。标记图像块的工作对医生来说是费时费力的。研究人员想到使用多示例学习(Multiple Instance Learning, MIL)注意力机制来解决这个问题。但是多示例学习注意力机制是一种硬注意力机制，无法对每个图像块内部进行权重评估，从而降低最终分类准确率。

本文提出使用两种方法来解决多示例学习注意力机制缺乏软注意力机制的问题，并且通过在结肠癌数据集上进行分类来证明这两种方法的优越性。本文提出使用的第一种方法是一种软的、连续的、空间的、自顶向下的注意力机制(Soft, Sequential, Spatial, Top-down Attention mechanism, S3TA)。该注意力机制受到灵长类动物视觉系统启发。人类并不是将图像视为一个静态的场景，而是在一系列扫视中探索图像，并且在过程中收集和整合信息。这就是人类进行图片分类总比通常的神经网络强的原因。S3TA 是一种循环神经网络(Recurrent Neural Network, RNN)，它模拟了注意力瓶颈和视觉皮层的顺序的、自上而下的循环控制的功能。在多示例学习注意力机制中加入 S3TA 有利于弥补硬注意力机制的不足之外，还能获得人类进行图片分类的优点。本文提出使用的第二种方法则是从通道域角度对多示例学习注意力机制进行改进。本文使用一种压缩和激励(Squeeze and Excitation, SE)注意力机制模块来改进多示例学习注意力机制。SE 模块对卷积层中各个通道的相互依赖关系进行表示，从而让网络可以对各个通道域的权重进行重新校准，也就是说通过 SE 模块可以在通道域中对重要信息进行强调并且抑制无效特征。因为本文中所有网络模型的深度较浅而且原 SE 模块在浅层中表现很差，所以本文对 SE 模块进行了改进。改进后的 SE 模块不仅可以在任何深度插入网络模型，而且还保留了简单灵活有效的优点。实验结果表明两种新模型比原多示例学习注意力机制模型有更高的分类准确率，而且两者相结合的模型表现更好。本文为降低 WSI 标记成本提供了更加有效的解决方案。

关键词：注意力机制, 多示例学习, 结肠癌, 循环神经网络，通道域

Abstract

When using traditional deep learning for pathology image classification, researchers often consume a lot of time and labor costs due to the need to label a large number of pathology images. In order to reduce the time and labor cost of labeling images, this thesis studies existing solutions. When classifying histopathological images, people will crop the whole slide image (WSI) into many small patches because of the huge WSI. While cropping the patches, the doctor labels each patch. This allows these patches to be fed into the neural network for training. The job of labeling patches is time-consuming and labor-intensive for doctors. The researchers thought of using the Multiple Instance Learning (MIL) attention mechanism to solve this problem. However, the MIL attention mechanism is a hard attention mechanism and cannot evaluate the weight inside each patch, thereby reducing the final classification accuracy.

This thesis proposes the use of two methods to address the lack of soft attention mechanisms for MIL attention mechanisms, and demonstrates the superiority of both methods by classifying them on the colon cancer dataset. The first method proposed in this thesis is a soft, sequential, spatial, top-down attention mechanism (S3TA). This attention mechanism is inspired by the primate visual system. Rather than viewing an image as a static scene, humans explore the image in a series of glances, gathering and integrating information in the process. This is why humans are always better at classifying pictures than usual neural networks. S3TA is a recurrent neural network (RNN) that simulates the attentional bottleneck and function of sequential, top-down, recurrent control of the visual cortex. Adding S3TA to the MIL attention mechanism is beneficial to make up for the insufficiency of the hard attention mechanism, and can also obtain the advantages of human image classification. The second method proposed in this thesis is to improve the MIL attention mechanism from the channel domain perspective. This thesis uses a squeeze and excitation (SE) attention mechanism module to improve the MIL attention mechanism. The SE module represents the interdependence of each channel in the convolutional layer, so that the network can recalibrate the weights of each channel domain, that is to say, the SE module can emphasize important information and suppress invalid features in the channel domain. Because the depth of all network models in this thesis is shallow and the original SE module performs poorly in shallow layers, this thesis improves the SE module. The improved SE module can not only insert the network model at any depth, but also retains the advantages of simplicity, flexibility and efficiency. The experimental results

show that the two new models have higher classification accuracy than the original MIL attention mechanism model, and the combined model performs better. This thesis provides a more effective solution for reducing the cost of WSI labeling.

Keywords: Attention Mechanism, Multiple Instance Learning, Colon Cancer, Recurrent Neural Network, Channel Domain

目录

第一章 综述

本章将介绍研究的内容及其意义。首先，本章介绍课题研究的背景。然后，本章介绍计算机诊断的国内外研究现状。接着，本章介绍一些常用的组织病理学图片分类的方法。最后，本章介绍本文的整体结构。

§1.1 课题的研究背景与意义

癌症至今被人们认为是不治之症。很多病人在癌症初期时并不自知，因此癌症患者大都会错过最佳治疗时间。迄今为止，在恶性肿瘤中，结肠癌已经成为最严重威胁人类生命的癌症之一。在 2018 年的全球癌症统计报告 ①中显示，结肠癌占所有恶性肿瘤新发病人数的 10.2%；其中死亡人数大约为 86.1 万，占据所有因为肿瘤致死人数的 9.2%；其死亡率在所有癌症中排名第二，发病率排名第三。目前为止，临床医学技术对结肠癌的诊断非常麻烦。而且因为这种病症与结肠炎很类似，所以很多患者在前期很难发现癌变。从结肠癌的相关数据中可以看出，癌症严重威胁着人类健康。据有关资料显示，如果患者能够较早发现结肠癌并且进行治疗，患者就拥有 90%的 5 年生存率。但是，若患者未能及时发现并治疗，则其只有 14%的 5 年生存率。因此我们有必要加强对结肠癌之类的癌症的预防和筛查。

组织病理学图像是癌症诊断的金标准。高效准确的诊断对患者及时发现和后续治疗至关重要。到目前为止，计算机辅助诊断还没有在病理学领域得到广泛应用，目前解决的任务只是冰山一角。目前我国的医学影像数据年增长率巨大，大约为每年 30%。但是每年医学影像科的医生数量的增长率却不到 5%。两者之间的巨大差距导致大量的临床医学图像诊断需要延后很长时间，有些患者因此错过黄金治疗时间。于是很多研究者想到使用计算机辅助诊断来弥补两者之间的巨大差距。

医学中应用计算机辅助诊断相关技术可以追溯到 20 世纪 50 年代。在 1959 年，美国研究员 Ledley 等人[54]第一次提出了计算机辅助诊断的数学模型，并且将该模型成功引入临床医学。他们使用该模型诊断了一组肺癌病例，这便是医学中使用计算机辅助诊断的开篇。在 1966 年，Ledley 正式提出了计算机辅助诊断的概念。20 世纪 80 年代初，计算机辅助诊断系统得到进一步发展，其中应用于中医领域的专家系统最为引人注目。计算机辅助诊断的过程包括患者数据和检查

① http://www.mango-group.it/fileadmin/user_upload/pdf_assemblea2021/GADDUCCI_1.pdf，2018 年

数据的收集、医学信息的定量处理、统计分析和最终诊断。当时比较流行的计算机辅助诊断模型有贝叶斯定理和极大似然模型等。20 世纪 60 年代以后，计算机辅助诊断的研究陷入了低谷。一方面，人们对计算机辅助诊断的期望很高，希望在计算机的帮助下实现自动诊断。另一方面，由于缺乏相应的理论算法和原理分析，计算机辅助诊断的发展仍然受到限制。这种内外双重困境直到上世纪 80 年代和 90 年代才得到改善。当时凭借计算机技术和各种数学、统计学的飞速发展，一些发达国家在医学影像领域取得了重大突破。

近年来，人工智能正在如火如荼的发展。它的发展恰好带动了医学图像辅助诊断的发展。在国际象棋、将棋、围棋等游戏中，人工智能已经能够击败人类冠军，并且能够实现自动驾驶。在这种情况下，人工智能在医学诊断领域以惊人的速度发展，尤其是影像诊断领域。人工智能是计算机科学的一个重要分支，它能够根据输入计算机的环境因素，最大程度的提高某件事情的成功率。除此之外，人工智能可以从之前的经验中学习新的东西，做出合理决策，并且快速回应。人工智能包含四个主要的组成部分：专家系统、启发式问题解决、自然语言处理和计算机视觉。当前人工智能在计算机视觉中的应用主要建立在神经网络之上。神经网络的结构类似于大脑神经突触联接结构，它是一种能够进行信息处理的数学模型。现在，几乎所有主流计算机视觉网络模型都以卷积神经网络为基础。卷积神经网络是一种深度学习模型，常用于视觉图像分析。

深度学习是机器学习的重要组成部分，是一种以神经网络为基础，对数据进行表征学习的算法。到目前为止，深度学习已经有很多种模型框架，如深度神经网络、卷积神经网络和深度置信网络以及循环神经网络等等。这些模型被大量应用于计算机视觉、语音识别、自然语言处理、音频识别与生物信息学等领域。深度学习最早可以追溯到 1980 年福岛邦彦提出的新感知机。1989 年，Yann LeCun 等人[57]将 1974 年提出的反向传播算法应用于深度神经网络。这一网络被应用于手写邮政编码的识别。虽然这个算法可以成功执行，但是由于计算量巨大和当时计算机低下的算力的影响，该网络的训练时间长达 3 天，所以没人将其投入实际应用中。导致训练缓慢的因素还有很多，比如梯度消失问题。在 2007 年前后，Ruslan Salakhutdinov 等人[55]提出了一种前馈神经网络中进行有效训练的算法，并在实验中证明了这一算法能够有效提高有监督学习的执行速度。硬件的进步也是深度学习重新获得关注的重要因素。高性能图形处理器的出现极大地提高了数值和矩阵运算的速度，使得机器学习算法的运行时间得到了显著的缩短。人们适用深度学习技术来处理医学图像问题有两大优势。第一，准确率、效率、可靠性很高，并且这些特点和优势还在不断提升。第二，深度学习模型是可以复用、可移植、可延续的，而人类医生会因为各自经验、状态等因素造成不同的诊断结果。

本文为了使用深度学习技术进行医学图像诊断，对最新的深度学习分类网络、多示例学习模型和注意力机制做了大量研究之后，力求设计出一个可以使用无标注的全视野数字切片(Whole Slide Image, WSI)进行分类的网络并且提高其分类准确率。

§1.2 国内外研究现状

计算机辅助诊断(Computer Aided Diagnosis, CAD)是计算机和医学两大领域学科交叉形成的一门学科。计算机辅助诊断又称计算机辅助检测(Computer Aided Detection, CAD)，其具体定义为通过影像学、医学图像处理技术以及其他可能的生理、生化手段，结合计算机的分析计算，辅助发现病灶，提高诊断的准确率。可以看出计算机辅助诊断技术的定义非常广泛，平时常说的计算机辅助诊断技术实际上大多数是指医学影像学的计算机辅助技术。而后者只需要根据医学影像进行判断是否存在病灶区域，无需进行进一步的医学诊断。计算机辅助诊断系统的广泛应用有助于提高医生诊断的精确性，所以计算机辅助诊断技术常常被称为医生的第三只眼。图像分类在所有计算机视觉领域都属于基础任务，在医学图像辅助诊断中尤为重要。

§1.2.1 传统机器学习的医学图像分类

从人工智能风靡全球开始，机器学习就一直在计算机视觉领域大放异彩。即使在当下深度学习大火的时代，依旧可以看到传统机器学习的身影，比如很多图像分类的研究工作中，很多研究人员都会在分类器中使用支持向量机(Support Vector Machine, SVM)进行分类。

在使用机器学习进行病理学图像分类的任务中，主要步骤是先对图像进行特征提取，然后使用各种分类器进行分类。这样相比于人工分类的效率和精度都有显著提高。这种基于传统机器学习的模型有很多。比如 Rathore S 等人[56]针对图片的纹理、几何、颜色和 Haralick 特征进行特征提取，然后集成了旋转增强分类器、支持向量机和集成分类器对结肠癌进行分类，最终达到 98%的分类准确率。

虽然机器学习在很大程度上提高了癌症病理学图像分类的速度和自动化水平，但是传统机器学习的特征提取部分却不如人意。特征提取部分就是计算机对图像的某些有用特征进行保留和放大并且弱化无用特征，最终将这些特征转化为有利于计算机分析的形式。但是传统机器学习特征提取算法的鲁棒性和泛化能力比较差，所以无法满足未来数据更新较快的计算机医疗辅助诊断的需求。

§1.2.2 基于常规深度学习的医学图像分类

为了解决上述问题，深度学习技术被广泛应用于现代医学图像研究中。深度学习拥有出色的特征提取能力，这使其在医学图像分析和医学图像诊断领域中表现极为突出。深度神经网络是一种通过组合多层神经元网络的端到端的网络模型。这种网络通过多次的前向传播和反向传播对网络中的参数进行迭代更新，最终得到一个几乎不会再改变的网络模型。其中卷积神经网络(Convolutional Neural Networks, CNN)通过共享参数的方式对图片进行特征提取，最终得到的模型相比于全连接神经网络更加具有鲁棒性，并且精确度也更高。所以几乎所有和图像有关的网络模型都会使用卷积神经网络。

卷积神经网络是计算机视觉领域最具代表性和创造性的算法之一。众所周知，在计算机视觉领域中，ImageNet 竞赛就相当于在体育领域中的奥运会。在 2012 年时，Alex Krizhevsky 凭借卷积神经网络在这项竞赛中一举夺魁，成功将之前的图片分类的错误率从 26%降到了 15%。自此之后，卷积神经网络便开始飞速崛起。很多大型互联网公司都看中卷积神经网络的优越性能和扩展性，并开始围绕卷积神经网络进行研究和使用。比如 Instagram 将卷积神经网络应用于搜索引擎的算法中，Facebook 公司使用卷积神经网络来自动标记，Google 也使用了卷积神经网络来进行图片搜索。

从上文可知，卷积神经网络很强大，本文主要使用卷积神经网络进行病理学图像分类。图像分类是指：把一张图片给机器，机器根据这张图片进行运算，最终得出这张图片所属类别。图像分类对于人类而言是一件非常简单的事情，但是对于机器来说却是困难的。因为机器没有人脑那么多的神经元，所以无法像人一样思考。于是很多研究者想到用机器模拟人脑的神经元进行运算，初代的神经网络就这样应运而生，也就是我们常说的全连接神经网络。虽然全连接神经网络理论上很好，但是因为参数众多，所以训练困难。全连接神经网络应用于输入维度较低的任务中，它的表现还可以。但是它被应用到图像分类这样的高维度输入的任务中就显得力不从心了。这种情况就是卷积神经网络大展身手的时候。

将大脑的视觉神经系统应用于计算机领域是在 1998 年， Yann LeCun [50]提出了现代卷积神经网络的雏形 LeNet-5，该模型还将反向传播算法应用到了网络的训练中，但是由于初代卷积神经网络并不成熟而且当时计算机硬件的计算能力跟不上，所以算法的表现还是不如人意。这种情况一直持续到 2012 年 AlexNet 出现，卷积神经网络才算真正开始崭露头角。在 2014 年 ImageNet 竞赛中，Hinton 研究小组在 AlexNet 的基础上引入了 dropout，该方法用来应对过拟合问题。该模型又一次大幅下降了分类错误率。

通常来说，卷积神经网络有三个组成部分：卷积层、池化层和全连接层。卷积层是卷积神经网络的基础，它由大量的卷积核组成。卷积层的主要作用就是对图片进行特征提取。卷积神经网络和原始的神经网络不同的地方在于神经元之间不再是一一连接，转而变成了卷积核区域对图片进行全部区域扫描的连接形式。也就是说卷积层有两种基本特征：权值共享和局部连接。卷积层的核心操作是卷积操作，也就是将图片某个区域的多个值映射为一个单一值。这就是卷积层的局部连接特征。而卷积操作过程中所使用的卷积核在一次前向传播时都是不变的，换句话说就是所有局部区域映射为单一值的时候，卷积核是共享的。这就是卷积层的权值共享特征。这里所说的卷积操作和数学中的卷积操作完全是两码事，但是都有相乘步骤所以会以此命名。卷积操作可以表示为：

$$f_l^m(\alpha,\beta)=\sum_j\sum_{p,q} i_j(p,q)\cdot\lambda_l^m(u,v), \qquad (1-1)$$

其中$i_j(p,q)$表示第i张图片的第j个通道的第p行第q列的参数，同理$\lambda_l^m(u,v)$表示在第l层的第m个卷积核的第u行第v列的参数，$f_l^m(\alpha,\beta)$表示经过第l层的第m个卷积核进行卷积运算之后，所得到的特征图在第α行第β列的参数。

池化层又称为下采样层，也是卷积神经网络的一个重要组成部分。下采样过程的作用是过滤无用信息并且保留有用信息。这个过程就类似于化学中的半透膜过滤液体。池化处理带来的好处就是减少特征尺寸，顺便降低后续的计算成本。因为池化层的作用是对特征进行过滤，所以直接提升了整个网络的鲁棒性，使得网络更加健壮不容易受到噪声干扰，这也有利于防止过拟合。常用的池化操作有：随机池化、均值池化和最大池化。

在普通的图像分类任务中，人们通常会把一幅分辨率不大的图片放入卷积神经网络中进行特征提取并分类。通过数字化病理学图像获得的全视野数字切片(Whole Slide Image, WSI)的分辨率是非常高的。如果将整个 WSI 放入卷积神经网络中[3-5]，那么整个网络的参数量将变得非常巨大。最终整个网络的鲁棒性和准确率都会大幅下降。因此，研究人员用需要其他手段去处理 WSI，使其可以放入卷积神经网络中训练。最常见的做法是将 WSI 裁剪为若干很小的图像块，并且将这些图像块全都进行标注。这样，人们就能将图像块和对应的标签放入卷积神经网络进行分类训练[6-14]。

这种分割 WSI 并且进行全部标记的方法又产生了另一个问题，那就是对大量的图像块进行标注是非常费时费力的。众所周知，医生的时间都是非常宝贵的。标注大量的图像块对于他们来说就是简单的机械式劳动，这大大占用了他们处理其他医学问题的时间。有人想到雇佣工人进行标注，这种想法甚至催生了人工智能扶贫相关行业的诞生。这种做法存在一个问题，普通人想要对分割出来的图像

块进行标注是需要时间进行培训的。并且工人的工资和标注准确率都是大问题。据了解，一名熟练的标注工人一天所能标注的 WSI 数量在 200 张左右，效率并不是很高。综上所述，需要一种可以消除或者大大降低标注成本的手段来解决常规深度学习网络模型不能解决的问题。

§1.2.3 基于多示例学习的医学图像分类

针对上述需求，很多研究者们将目光投放到了多示例学习(Multiple Instance Learning, MIL)上。多示例学习不需要对图像块进行逐个标注，仅仅需要对每个 WSI 进行标注。对一个 WSI 进行裁剪获得的图像块作为一个包，这个包中的所有图像块的标签都和原 WSI 相同。最后将这个包放入多示例学习模型中进行训练即可。用这种方法就达到了大幅度降低标注成本的目的。

在图像分类等典型深度学习问题中，研究者都会认为每个图像只能属于一个类别。然而，在许多实际应用中，一张图片会有多个类别的示例在其中。但是最后这张图片只会被分类到一个类别中去。在 1997 年左右，Dietterich 和 Maron 等人[52]将这样的学习模式称为多示例学习。在 2014 年，Oquab 等人[58]又把这种方式称为从弱注释中学习数据。在 2017 年，Quellec 等人[59]发现弱标记的问题数据在医学成像中尤为明显，例如，计算病理学、乳房 X 光线照相技术或 CT 肺筛查。这些病理学图像中通常获取的单个标签或感兴趣区域(Region Of Interest, ROI)大多数都是粗略给出的。多示例学习用来处理整个包的图片，一个包中的所有示例都有一个相同的标签。因此，多示例学习的主要目标是学习预测包标签的模型，例如医学诊断。在 2012 年，刘国庆等人[60]进行了一个发现关键示例的额外挑战。这些关键示例就是用来决定包类别的主要决定性因素。在医学领域里面，因为一些法律问题和临床实践的用处，这项工作的作用显得非常巨大。为了解决一个包分类不同的首要任务，研究人员提出了一些方法，例如在 2015 年，Cheplygina 等人[61]提出利用包之间的相似性，将示例嵌入到更紧凑的低维表示中去。之后再将这种低纬度张量放入到 2003 年由 Andrews 等人[62]提出的包级别分类器中完成分类，该分类方法还结合了 2000 年 Ramon 等人[63]提出的示例级的响应分类器。但是只有最后一种方法能够提供可解释的结果。然而，在 2015 年，Kandemir 和 Hamprecht 的文献[64]显示此类方法的示例级准确性较低。并且 Cheplygina 等人[61]在 2015 年表示大量示例级别的多示例学习方法之间存在很多争议。这些问题令人质疑多示例学习模型是否可以用于解释最终结果。

本文将使用一种在 2018 年由 Maximilian Ilse 等人[15]提出的方法。该方法旨在提高多示例学习方法的可解释性和灵活性。该方法创建多示例学习模型时，对包标签使用伯努利分布，并通过优化对数似然函数来训练它。该方法已经证明了

应用对称函数的基本定理可以提供一个为包标签概率建模的一般程序，也就是包评分函数。该程序由三个步骤组成，首先将示例转换为低维嵌入的转化，然后置换不变聚合函数，最后转换为包概率。该方法提出使用神经网络，即卷积层和全连接层，来参数化所有转换，这增加了该方法的灵活性。并且该方法通过优化不受约束的目标函数来进行端到端的方式训练模型。最后同样重要的是，该方法替换了广泛使用的置换不变算子，例如最大值算子和加权平均算子，其中权重由两层神经网络给出。值得注意的是，注意力权重使网络能够找到关键示例，这些示例可以进一步用于突出显示每个示例的重要性。但是这种方法的鲁棒性较低，并且也不能对每张图像块内部进行评估。

本文的创意来自于 2020 的 CVPR 会议中，Noriaki Hashimoto 等人[16]提出将多示例学习注意力机制和域对抗(Domain Adversarial, DA)正则化还有多尺度技术(Multi-Scale, MS)融合，融合模型的准确率和鲁棒性获得了一定提升。本文认为这种方法很有创造性，但是该方法需要进行优化。这种方法最大的弊端就是为了融合三种方法，使得整个模型的参数量急剧提升。原多示例学习方法的输入尺寸为27×27，但是因为需要使用域对抗正则化的方法来提高模型鲁棒性，融合模型的输入尺寸直接变为224×224的输入尺寸。这种做法导致参数剧增。其次，融合模型的特征提取部分是由多个尺度的特征提取器分别训练之后进行拆分，然后再融合所有训练好的特征提取器，最终放入一个多示例学习注意力机制的分类器中。从上述对该模型的描述可以看出，该模型是一个分两阶段，非端到端的训练模型。虽然模型的准确率和鲁棒性有所提升，但是本文认为大部分的功劳在于参数过度增加。为了大幅降低参数量和模型分两阶段训练的复杂性，该模型非常有必要被优化。关于优化该模型的方案，本文使用了一种循环神经网络注意力机制来完美解决上述问题。

§1.2.4 基于注意力机制的图像分类

注意力机制已经广泛应用于很多序列输入的网络模型中，比如机器翻译、视频和标记、图片分类和标记、文本分类等任务。多示例学习注意力机制也是一种硬注意力机制。可以看出注意力机制的应用已经相当广泛。本文使用注意力机制是一种基于循环神经网络的注意力机制和一种通道域注意力机制。

本文用到的循环神经网络注意力机制是一种软的、连续的、空间的、自上而下的注意力机制(Soft, Sequential, Spatial, Top-down Attention mechanism, S3TA)[18]。本文使用的通道域注意力机制是一种被称为挤压和激发注意力机制(Squeeze-and-Excitation, SE)[37]。第一种方法从灵长类视觉系统中汲取灵感。该方法拥有捕捉视觉皮层的一些功能，即注意力瓶颈和顺序，自上而下控制。该方法在

ImageNet 图像上对模型进行对抗训练，表明它具有最先进的鲁棒性。该模型可以通过增加展开模型的步数，来更好地防御和对抗更强的攻击，这就形成了计算进攻方和防守方之间的竞赛。第二种方式则从通道域角度出发对多示例学习注意力机制进行改进。该方法相比于第一种方法更佳简单灵活且实用有效。

§1.3 论文研究内容

本文研究内容是优化原有的多示例学习注意力机制模型，并且使用改进的模型对医学图像进行分类识别。主要研究方向有两个。第一，改进多尺度技术[16]的两阶段训练，使其成为一个端到端的模型进行训练。然后改进因为添加模块而造成的参数量暴增问题。最后提高该模型的鲁棒性和分类准确率。第二，本文通过改进 SE 模块，使其可以嵌入浅层神经网络中并且对多示例学习注意力机制进行优化。具体研究内容如下。

本文的第一项研究工作是在多示例学习分类器的基础上，对特征提取部分进行优化。首先用一般的卷积神经网络对带有包标签的医学图像进行特征提取。然后将输出的特征图通过 S3TA 进行优化。最后将 S3TA 输出的特征向量组放入多示例学习注意力模型进行分类。在最终的分类器选择上，本文使用 sigmoid 分类器。对上述模型中重要模块进行替换，用来验证最终选定的模块都是最有效的模块。原始的 S3TA 模型中使用的循环核是 LSTM，本文将其替换为 GRU。将 S3TA 模型在步数上进行扩展，选择一个有效并且时间代价不大的扩展步数。最后将该模型的超参数进行微调，选出最好用的超参数组合。最后本文将统计新模型的参数量。因为文献[16]中模型的输入和本文输入尺寸不同，所以没有多少对比性。本文将新模型的参数量和原多示例学习注意力机制的参数量进行对比，可以看出改进模型的优势。

本文的第二项研究工作是将 S3TA 模块的成功经验应用于 SE 模块进行改进。SE 模块之所以在浅层性能差强人意，是因为在 SE 模块在压缩阶段直接使用了全局平均池化导致平面位置信息丢失。为了防止平面位置信息丢失，S3TA 在 keys 和 values 张量中分别加入了一个固定不变的空间基。所以，本文仿照这个处理方案，在 SE 模块使用全局平均池化压缩之前，将特征张量和空间基在通道维度进行合并。最后通过实验说明这种方法的优势。

本文是一种应用型计算机医疗辅助诊断的研究，不仅在理论层面优于原有模型，而且实际的效果也可以在实验中得到验证。通过相关的对比试验，本文的两个模型明显优于原有模型和其他改进模型。本文的第一种模型具有很强的鲁棒性，自身的循环步数可以灵活改变来增强模型的鲁棒性。第二种模型更加简单灵活，而且适用范围更广。

§1.4 论文组织结构

本文的主要内容如下所示，主要包含 6 个章节：

第一章，综述。本章首先介绍当前医疗环境和对计算机辅助诊断的迫切需求，也就是研究的背景和意义。然后本章对国内外研究现状进行阐述，并且着重分析了原有方法和现阶段方法的优势和劣势。接着本章对需要用的模块进行简要描述。最后本章对研究目标和内容以及组织结构进行详细描述。

第二章，基于深度学习的医学图像分类方法。本章内容详细介绍了本文所涉及到的主要网络模块。首先，本章大概介绍深度学习在医学图像分类中的应用。然后，本章介绍了循环和通道域注意力机制模型。

第三章，多示例学习模型与基于 S3TA 的改进模型。本章具体介绍了多示例学习注意力机制与 S3TA 模块的融合模型如何搭建。首先，本章对 S3TA 模型和多示例学习注意力机制进行了详细描述。然后，本章介绍融合模型的具体的细节。

第四章，结肠癌分类的应用系统。为了使实验结果更加客观，本章采用原始多示例学习注意力机制使用的结肠癌公开数据集。首先，本章对结肠癌进行介绍。然后，本章对具体数据集的输入尺寸和数量进行介绍。最后，本章对实验环境进行介绍并且分析实验结果。

第五章，基于通道域注意力对多示例学习的改进和应用。首先，本章对通道域注意力机制中的 SE 模块进行改进。然后，本章将改进的 SE 模块与多示例学习注意力机制相融合。最后，本章通过实验证明改进的 SE 模块的优势。

第六章，总结与展望。首先，本章对本文所取得的研究成果进行总结。然后，本章分析新模型的不足之处，并且给出未来的发展方向。

第二章 基于深度学习的医学图像分类方法

本章将介绍本文的相关模型和一些基本知识。首先，本章介绍医学图像分类问题。然后，本章将重点介绍多示例学习、循环和通道域注意力机制。因为循环神经网络是循环注意力机制的基础，所以本章也会介绍循环神经网络。

§2.1 医学图像分类问题

将数字病理学应用于临床诊断可以说是病理学研究中最具颠覆性的技术之一。随着患者组织样本的数字化，数字病理学应运而生，特别是数字化 WSI 的使用。这些数字化 WSI 可以在全球范围内分享，并且用于诊断、教学和研究目的。相关研究表明数字诊断和基于切片的诊断之间存在相关性[38][39]。目前，欧洲最大的单点病理学服务机构之一的 NHS Greater Glasgow and Clyde 已经开始进行全面数字化的改进。随着数字病理学的应用越来越广泛，组织形态的自动图像分析有可能在病理学中进一步确立自己的地位，并最终减少病理学家的工作量，减少报告的周转时间，并使临床实践标准化。例如，可以自动量化已知或新的生物标志物和组织病理学特征。此外，深度学习技术可用于识别标本内的形态模式，并且用于诊断和分类目的[40]。

深度学习已经成功应用于处理 WSI，这有可能创造新的临床诊断方法。并且该方法在准确性、可重复性和客观性方面超越当前的临床方法，同时也为各种病理研究提供新的见解。然而，WSI 是数千兆字节的图像，典型分辨率为100,000 × 100,000像素，具有很高的形态差异，并且通常包含各种类型的伪影。这些条件排除了传统深度学习技术的直接应用。相反，从业者面临着两个艰难挑战。一方面，图像的视觉理解受到形态变化、伪影和数据集规模较小的阻碍，另一方面，当前硬件无法从具有如此高分辨率的图像中学习解决，因此需要对图像进行某种形式的降维。这两个问题有时被称 what 和 where 的问题[41][42]。大多数 WSI 是使用明场照明捕获的，例如用临床常规苏木精和伊红(Hematoxylin and Eosin, H&E)染色的载玻片。与更多定制标记试剂相比，H&E 染色 WSI 具有更广泛的实用性。目前这种模式对深度学习的研究人员更具吸引力。H&E 染色组织非常适合表征组织样本内的形态，这使得其在临床实践中的可以长期使用。然而，H&E 染色的载玻片缺乏与细胞相关的原位分子数据。相比之下，这可以通过免疫标记进行蛋白质可视化。可以使用多重免疫荧光 (Immunofluorescence, IF)观察多种细胞类型的标记及其蛋白质表达，这为癌症研究，特别是免疫肿瘤学提供了有价值的信

息。然而，使用深度学习技术对 IF 标记切片的分析受到 IF 和 WSI 数据集的有限可用性的阻碍。数据稀缺的潜在原因包括试剂费用和使用荧光扫描仪的费用，以及巨大的 IF 和 WSI 大小，有时每张图像可能超过 10 GB。与采用监督深度学习技术的许多其他领域不同[43][44]，标记数据在数字病理学中更难获得，从而限制了监督方法的实用性。尽管近年来有更广泛的数据发布，但大部分的工作仍然使用专有的 WSI 数据集 ①。目前市场上有多种来自不同供应商的 WSI 扫描仪，具有明场和荧光成像的能力。每个扫描仪使用不同的压缩类型和大小、照明、目标和分辨率来捕获图像，并且还以不同的专有文件格式输出图像。缺乏通用图像格式可能会延迟大型数据集的管理。放射学领域通过采用 DICOM 开源文件格式克服了这个问题，并且允许访问和查询大型图像数据集[45][46]。数字病理学尚未广泛采用单一的开源文件格式，尽管有很多研究正朝着这个目标不断推进[47][48]。为了在临床上可迁移，深度学习算法必须适用于大量患者群体，并概括图像伪影和染色中的颜色变异性。伪影可以在整个样品制备工作流程以及成像过程中引入。这些可能包括缺血时间、固定时间、切片伪影、染色试剂可变性以及来自不均匀照明、聚焦、图像平铺和荧光沉积和渗出的成像伪影。通过训练，人脑可以熟练地忽略伪影和染色变异性，并磨练准确诊断所需的视觉信息。为了促进深度学习模型中的类似结果，通常可以遵循两种方法。第一个涉及显式去除伪影，例如，使用图像过滤器，以及颜色可变性的归一化。相比之下，第二种方法采用了一种不太直接的策略，用通常是合成生成的数据来增强数据，这些数据捕获了伪影和染色的代表性可变性，使他们的学习成为训练过程中不可或缺的一部分。这两种方法都已成功地用于纠正批次效应或来自不同诊所的存档临床样本的差异，尽管这一发现并不普遍。在 WSI 上训练深度学习模型的大多数成功方法不使用整个图像作为输入，而是仅提取和使用少量图像块。图像块通常是正方形区域，尺寸范围大致从32×32像素到$10{,}000 \times 10{,}000$像素不等。其中大多数方法使用大约256×256像素的图像块。这种降低 WSI 高维的方法可以看作是人工引导的特征选择。选择图像块的方式也是 WSI 分析的关键研究领域之一。现有方法可以根据它们是否使用注释以及在哪个级别进行分组。

由于所有提取的训练图像块都具有类标签，因此图像块级别注释可实现强大的监督学习网络模型。通常，基于图像块的注释源自像素级注释，这需要专家注释所有像素。例如，给定包含癌组织的 WSI，病理学家需要定位和注释所有癌细胞。一种基于图像块的学习的简单方法将利用所有平铺的即非重叠的图像块。然而，这种简单性是以过多的计算和内存开销为代价的，以及许多其他问题，例如不平衡的类和缓慢的训练。考虑到在大多数情况下，图像块比原始 WSI 小得多，

① https://arxiv.org/pdf/1805.06983.pdf，2018 年

随机抽样图像块可能会导致更高的类别不平衡。因此，必须引导抽样。指导采样过程的一种方法是使用图像块级别注释。例如，在乳腺癌转移检测方面。大多数成功的方法还采用硬负挖掘，这是一个迭代过程，将误报添加到训练数据集以进行进一步训练。由于图像块级别注释的可用性，在这种情况下可以识别误报。由于实际限制，在大多数情况下，标记是在 WSI 级别上完成的，而不是单个图像块。尽管标签的粒度较低，但许多基于深度学习的方法已经证明了这是一个非常有前景的方法。技术各不相同，通常采用多示例学习、无监督学习、强化学习和迁移学习或其组合的形式。直观地说，目标通常是识别可以集体或独立预测整个切片标签的图像块。可以使用基于图像过滤器的预处理来减少需要分析的图像块数量。多项研究还采用 Otsu、滞后或其他类型的阈值作为识别 WSI 内组织的自动方法。其他操作，例如对比度归一化、形态学操作和特定问题的图像块评分系统，也可以用来进一步减少候选图像块的数量，甚至可以实现自动图像块定位。然而，想要验证每个图像块是否确实具有与切片相同的标签通常需要特定领域的专业知识，最佳图像的过滤过程甚至需要一些人类直觉。为了避免潜在的人为偏见，大多数方法采用无监督或多示例学习，或两者结合。Tellez 等人[53]研究了将图像块压缩到低维潜在空间中的无监督表示学习方法，然后提出可以直接在压缩的 WSI 上训练卷积神经网络。其他研究者使用传统的降维技术，例如主成分分析，以及在 ImageNet 上预训练的卷积神经网络来降低图像块的维数。还有的研究者使用 k-means 聚类并通过弱监督方式训练卷积神经网络来找到最具辨别力的图像块集群。一些研究者提出了一种期望最大化算法，该算法能够选择越来越多的判别图像块，同时在每次迭代中，卷积神经网络被多训练 2 个轮次。如果当作为卷积神经网络的输入时，预测更接近切片级别标签，则认为图像块更具区分性。最有希望的新兴方向之一旨在将选择图像块的过程纳入视觉理解。Courtiol 等人 ①通过在卷积层之后添加1 × 1卷积层来修改预训练的 ResNet-50，以此获得图像块级别的预测。该算法在顶部添加了一个 MinMax 层，然后是完全连接的层来预测切片级别标签。MinMax 层是一种注意力机制，它提供了对两个类中最具辨别力的块进行选择性训练的能力。最近的工作采用了注意力模型。直观地说，注意力模型最初与随机猜测图像块选择差不多，但会逐渐选择有助于更具辨别力的图像块。在许多情况下，训练发生在较低级别，例如图像块级别，但最终目标位于较高级别，例如切片级别。例如，在癌症诊断的情况下，可以训练卷积神经网络来识别切片中是否存在癌细胞。然而，需要某种类型的聚合操作来推断 WSI 是否包含癌细胞。这可以采取对一些或所有图像块预测进行最大或平均操作的形式。在其他情况下，可以使用由卷积神经网络提取的特征和更高级别可用的基本

① https://arxiv.org/pdf/1802.02212.pdf，2018 年

事实来使用和训练传统机器学习模型或循环神经网络。由于通过对较小区域的独立分析来分析大型输入图像，因此出现了基于图像块的分析的主要限制。特别是，这些方法本质上无法捕获分布在大于图像块大小的尺度上的信息。例如，虽然可以从单个斑块中提取细胞特征，但只有在分析更大的区域时才能捕获更高层次的结构信息，例如肿瘤的形状或延伸。对图像块之间的空间相关性进行显式建模作为一种潜在的解决方案已经被提出。然而，这个想法只在少数邻域中进行了测试，并且需要图像块级别的注释。一种不同的方法涉及从多个放大级别图像中提取图像块。最后，最近的工作试图通过使用更大的图像块大小来改善与基于图像块的分析相关的一些上述问题。计算机视觉的目的是创建能够视觉理解的算法解决方案。应用范围可以从对象识别和检测到图像字幕和场景分解。在过去十年中，计算机视觉的大多数领域都取得了显着进步，其中大部分受到基于神经网络的学习算法进步的影响。这些方法学的成功是现在建立的深度学习领域的一部分，可以归因于多种原因，特征提取阶段的转变通常被描述为主导因素。在过去的十年中，大多数方法都集中在寻找从图像中显式提取特征以供随后使用的模型的方法。通过使用卷积神经网络使这一过程自动化已被证明可以为现有的问题定制更多的判别特征。深度学习成功的一些关键因素包括硬件和软件的进步，以及数据可用性的提高。

随着数字病理学的出现，数千兆字节图像的分析称为深度学习的一个新挑战。构建能够理解 WSI 的深度学习模型是该领域的新挑战。当图像块级别标签可用时，在许多情况下图像块采样与硬负挖掘相结合可以训练出匹配甚至超过病理学家的准确度的深度学习模型。对于许多带有图像块级别注释的医学数据集，深度学习模型似乎表现出色，并且随着竞赛的引入，例如Camelyon16和Camelyon17，这种类型的深度学习在性能和可解释性方面一再证明其成功。因此，从千兆像素图像和图像块级别注释中进行基于图像块的学习似乎是最接近临床应用的方法。然而，在许多情况下，只能获得较低粒度的标记。因为它非常费力和昂贵，所以仅仅依赖这种方法是不可行的。此外，图像块级别的监督可能会限制深度学习模型的潜力，因为模型只能与提供的注释一样好。为了使用切片或患者级别的标签，当前的方法集中在 where 问题上。在更难的问题上，例如预后估计，以及其他低维潜在映射，Tellez 等人[53]的工作就显得很重要。关于 where 问题，一般有两种做法。第一个使用一种元学习，为了优化 where 问题，首先要优化 what 问题。第二种方法试图在端到端设置中同时优化 what 和 where 的问题。这是通过让卷积神经网络转发一组图像块并参与其中,或者通过在每个时间步定位并关注单个图像块来完成的。深度学习已经在与数字病理学相关的广泛医学问题中展示了其潜力。然而，对详细注释的需求限制了强监督技术的适用性。弱监督、无监督、

强化和迁移学习等其他技术可以用于处理大量数据集时对详细注释的需求。这种非强监督学习的兴起为 WSI 分析开辟了新的道路。

§2.2 多示例学习模型

在过去的几年中，从示例中学习是机器学习中最繁荣的领域之一。根据训练数据的模糊性，该领域的研究大致可以分为三种学习框架，即监督学习、无监督学习和强化学习。监督学习尝试学习具有正确标记的示例，其中训练示例具有已知标签，因此歧义最小。无监督学习尝试学习基础示例源的结构，其中训练示例没有已知标签，因此歧义最大。强化学习试图学习从状态到动作的映射，其中示例没有标签但具有延迟奖励，可以被视为延迟标签，因此模糊性介于监督学习和无监督学习之间。

多示例学习一词是由 Dietterich 等人[52]在研究药物活性预测问题时提出的。在多示例学习中，训练集由许多包组成，每个包包含许多示例。如果包至少包含一个阳性示例，则该包被标记为阳性；否则它被标记为负包。与所有训练示例都具有已知标签的监督学习不同，在多示例学习中，训练示例的标签是未知的；与所有训练示例都没有已知标签的无监督学习不同，在多示例学习中，训练包的标签是已知的；与训练示例的标签被延迟的强化学习不同，在多示例学习中没有任何延迟。相关研究已经表明，忽略多示例问题特征的学习算法表现的并不好，例如流行的决策树和神经网络。虽然多示例问题广泛存在，但因为以前的学习框架所解决的问题是独一无二的，所以多示例学习依然被认为是一种新的学习框架。在 1990 年代中期，Dietterich 等人[52]研究了药物活性预测的问题。它们的目标是通过分析已知分子的集合，赋予学习系统预测新分子是否有资格制造某种药物的能力。大多数药物都是小分子，通过与酶和细胞表面受体等较大的蛋白质分子结合而起作用。对于有资格制造药物的分子，其低能量形状之一可以与目标区域紧密结合；而对于不具备制造药物资格的分子，其低能量形状都不能与目标区域紧密结合。药物活性预测的主要难点在于每个分子可以有许多可供选择的低能形状，但目前生物化学家只知道一个分子是否有资格制造药物，而不知道它的哪些替代低能量形状对资格有响应。一个直观的解决方案是利用监督学习算法，将好分子的所有低能量形状视为正训练示例，而将坏分子的所有低能量形状视为负训练示例。然而，由于正训练示例噪声高，这种方法几乎无法奏效，这是因为好分子可能有数百个低能量形状，但可能只有其中一个是真正好的形状。为了解决这个问题，Dietterich 等人[52]将每个分子视为一个包，将分子的替代低能形状视为包中的示例，从而定义了多示例学习。Dietterich 等人[52]提出了三轴平行矩形(Axis-Parallel Rectangle, APR)算法来解决药物活性预测问题，该算法试图搜索由特征结

合构建的合适的轴平行矩形。GFS elim-count APR 算法首先从正包中识别出一个覆盖所有示例的 APR。然后，它通过贪婪地从负包中消除示例来逐渐缩小 APR。对于来自本包的每个示例，该算法计算必须从 APR 中排除的正包示例的最小数量，以便将相关示例从负包中排除。该算法迭代地选择从负包中消除最容易消除的示例，直到消除所有此类示例。然后，算法通过贪婪特征选择确定相关特征的边界，从而获得最终的 APR。GFS kde APR 算法和 GFS elim-count APR 算法的区别主要在于前者不只是计算必须排除的正包中的示例数才能从负包中排除示例。相反，GFS kde APR 会考虑初始 APR 所覆盖的不同正包中的示例数，并使用成本函数来控制从负包中消除示例的过程，以便每个正包在 APR 中保留至少一个示例。Iterateddiscrim APR 算法采用贪心反向拟合算法来识别一个 APR，该 APR 覆盖每个正包中的至少一个示例。然后，它利用这个 APR 来选择最具辨别力的特征。最后，利用核密度估计来通过扩展 APR 的范围来帮助改进泛化性，从而使来自正包的新示例很有可能落入 APR。

一般来说，监督学习器的重点是区分示例。因为所有训练示例都在监督场景中进行标记。但是在多示例学习中，即使可行，也很难区分训练示例，因为它们都没有被标记。此外，如果将包标签简单地视为其示例的标签，即认为正包只包含正示例，而负包只包含负示例，那么即使每个训练示例现在持有一个标签，学习任务都可能非常困难。这是因为 Dietterich 等人[52]所指出的正包噪声可能非常高。因此，是否可以区分训练示例是监督学习和多示例学习之间的主要区别。事实上，目前几乎所有的多示例学习算法都是根据一个通用规则从监督学习算法修改而来的，即将监督学习算法的重点从对示例的判别转移到对包的判别上。大多数多示例学习算法的作者都隐含地使用了该规则，他们已经确定了在不同类型的监督学习器上实现该规则的有效方法。在将多示例学习技术应用于实际任务时，必须考虑两个重要问题。首先是选择合适的多示例学习算法。第二个是设计一种合适的方法，将现实世界的问题抽象为多示例表示，即确定什么是包，包中的示例是什么。在这里，将此类方法称为包生成器。从某种意义上说，包生成器的设计比多示例学习算法的选择更重要，因为如果使用合适的包生成器，学习任务可能很容易，而如果使用较差的包生成器，学习任务可能会非常困难。多示例学习也已应用于机器人控制。为了使机器人能够在大规模环境中导航，通常需要完成地标匹配。也就是说，机器人应该能够根据在机器人当前位置获取的数据识别它是否在给定地标附近。一种常见的方法是将视觉图像与在地标位置获取的数据进行匹配。但由于图像可能会随着地标位置周围的小幅移动而发生显着变化，因此这种方法遇到了困难。更好的方法是将这个问题转化为几何图案的学习。Goldman 等人[65]提出了一种在线不可知论学习算法，通过将问题简化为学习大

量变量的析取问题来学习离散常量维几何模式的类，这实际上是一种多示例学习算法，可以学习轴平行的矩形。阻碍多示例学习发展的最严重的问题是只有一种普遍使用的现实世界基准数据，即 Musk 数据集。虽然一些应用数据已经在一些文献中使用，但由于某些原因，它们很难作为基准。但由于 COREL 数据库包含大量图像，通常只使用数据库的一部分，而不同研究人员使用的部分通常不同；Chevaleyre 和 Zucker 以及 Alphonse 和 Matwin 使用了 Mutagenesis 数据，但这些数据通常用于测试 ILP 学习器而不是多示例学习器。尽管有一些人工数据集，但它们很难被广泛使用，因为它们是为多示例学习的扩展而设计的，例如多示例回归和广义多示例学习。此外，人工数据集的意义可能小于现实世界的意义。

Dietterich 等人[52]认为迭代判别 APR 算法的性能可能是 Musk 数据的上限，但这种性能水平已经被几种算法超过。事实上，目前 Musk 数据的准确度非常高，很难预测新算法会做得更好。即使一些新算法可以做到这一点，这也可能不是一个好消息，因为这些算法很有可能对这些特定数据产生过拟合。为了为测试新算法和比较不同算法提供公平的基础，迫切需要更多的数据集。在提出多示例学习的概念时，Dietterich 等人[52]提出了一个开放性问题，即如何为流行的机器学习算法设计多示例修改。这个悬而未决的问题极大地推动了这一领域的发展。事实上，几乎所有流行的机器学习算法的多示例版本都是在过去几年中开发出来的。现在似乎是时候提出新的具有挑战性的问题来刺激新算法的设计了。以下可能是一个不错的候选者：进行有效地标记看不见的包中的示例。实际上，在多示例学习的原始定义中，任务只是从训练集中学习一些概念，以正确标记看不见的包。但在大多数应用程序中，如果包中的示例能够正确标记会更有帮助。目前包括 APR 算法、Diverse Density 和 EM-DD 在内的几种算法都可以识别出包含真实正例的区域，这可能为解决上述问题提供依据。很明显，在这个方向上还有很多工作要做。许多研究人员试图将新问题引入多示例学习的研究范围。多示例回归显然值得研究，因为它的潜力是显而易见的，至少 Dietterich 等人[52]认为这样。如果可以产生实值输出，则可以预测不同分子的结合强度，这对药物设计很有价值。利用多示例范式来连接命题和关系学习也很值得探索，因为这可能会带来新一代强大的学习算法。广义多示例学习的价值应从应用程序中检验。如果可以识别出一些现实世界的广义多示例问题，而不是诸如门钥匙问题之类的故事，那么广义多示例学习就值得研究。确实，关注应用不仅可以帮助人们确定多示例学习扩展的价值，还可以帮助人们获取数据集和激发多示例学习的问题。

多示例学习已成功应用于组织病理学图像的分类问题中。当没有为每个提取的图像块进行注释时，多示例学习框架在组织病理学图像分类中很有用。本文中的主要思想就是使用多示例学习框架来自动识别多个感兴趣区域。

§2.3 循环与通道域注意力机制

所谓注意力机制就是分配权重给输入。近年来，注意力机制被广泛应用于自然语言处理、计算机视觉和语音识别等领域，是深度学习最值得研究的技术之一。在计算机视觉领域的注意力机制的基本思想就让网络自行学会注意，从而忽略掉无用信息并且放大有用信息。

计算机视觉领域中，按照注意力机制的可微行可以分为：硬注意力机制和注意力机制。硬注意力机制一种 0/1 问题，具体一点就是针对图片的某块区域关注或者不关注。如果你是一名计算机视觉研究者，那么你都有意或者无意间都会使用一种硬注意力机制：图像裁剪。也就是说，硬注意力机制是不可微的，它不会注意到一个图像的内部像素，它会将整个注意力机制作为整体来进行关注。相对的软注意力机制就是一个可微的注意力机制，它对一个区域的所有像素点都会施加权重来衡量哪些像素需要注意，哪些像素不需要。可以对关注的区域进行分类：空间域、通道域等，虽然还有其他划分方法，但是这里并不是重点。空间域让人联想到三维空间，但是这里空间域实际上指的是平面空间。空间域注意力机制的经典算法是在 2015 年的 NIPS 上发表的[51]，该文使用空间域注意力机制将原图片中的关键信息映射到另一个空间中。该算法的目的是为了改变池化层过于暴力的过滤手段，因为池化操作可能会导致丢失重要信息。通道域注意力机制是指在图片的通道维度上进行注意，经典算法就是 2017 年 CVPR 上发表的文献[37]。该算法通过对特征图进行挤压、激励得到各个通道维度上的权重，并将这些权重施加到原特征图进行注意。还有其他很多注意力机制，本文主要使用一种基于循环神经网络的注意力机制和一种通道注意力机制分别对多示例学习注意力机制进行改进。神经网络是强大的学习模型，可在广泛的监督和无监督机器学习任务中取得最先进的结果。它们特别适合机器感知任务，其中原始的底层特征不能单独解释。这种成功归功于他们学习层次表示的能力，这与依赖手工设计特征的传统方法不同。在过去的几年里，存储变得更加实惠，数据集变得更大，并行计算领域也取得了长足的进步。在大型数据集的设置中，简单的线性模型往往会欠拟合，并且经常不能充分利用计算资源。深度学习方法，特别是基于深度置信网络的方法，是通过堆叠受限玻尔兹曼机贪婪地构建的，以及利用视觉信息的局部依赖性的卷积神经网络，已经在许多重要应用中展示了创纪录的结果。然而，尽管它们很强大，标准的神经网络还是有局限性的。最值得注意的是，它们依赖于训练和测试示例之间独立性的假设。在处理完每个示例后，网络的整个状态都会丢失。此外，标准网络通常依赖于作为固定长度向量的示例。因此，希望将这些强大的学习工具扩展到对具有时间或顺序结构和可变长度输入和输出的数据进行建模，

尤其是在神经网络已经是最先进技术的许多领域中。循环神经网络(Recurrent Neural Network, RNN)是连续模型，能够选择性地跨序列步骤传递信息，同时一次处理一个元素的序列数据。因此，他们可以对由不独立的元素序列组成的输入和输出进行建模。此外，循环神经网络可以同时在多个尺度上对顺序和时间依赖性进行建模。

刚入门的学者不明白为什么要建立顺序性模型，这是一个值得深究的问题。支持向量机、逻辑回归和前馈网络已被证明非常有用，而且无需建模时间。可以说，正是独立性假设促进了机器学习的进展。此外，许多模型通过将每个输入与一些直接的前任和后继连接起来，隐含地使用了时间序列，也为机器学习模型提供关于每个感兴趣区域的窗口。不幸的是，尽管独立性假设很有用，但它排除了对具有长期依赖关系任务的建模的可能性。例如，使用长度为 5 的有限长度上下文窗口训练的模型永远无法训练回答简单问题：六个时间步前看到的数据点是什么。对于呼叫中心自动化等实际应用，这种有限的系统可能会学习路由呼叫，但永远无法完全成功地参与扩展对话。自从最早的人工智能概念出现以来，研究人员就一直在寻求构建能够及时与人类互动的系统。Alan Turing[49]提出了一种模仿游戏，通过令人信服地参与对话的能力来判断机器的智能。除了对话系统，具有经济重要性的现代交互系统还包括自动驾驶汽车和机器人手术等。如果没有明确的顺序或时间模型，分类器或回归器的任何组合似乎都不太可能拼凑在一起以提供此功能。循环神经网络并不是唯一能够表示时间依赖性的模型。

为什么连续模型，即人工神经网络，表现更好是合理的。首先，循环神经网络可以捕捉长期的时间依赖性。但是，任何时间步的隐藏状态都可以包含来自几乎任意长的信息上下文窗口。这是可能的，因为可以在节点的隐藏层中表示的不同状态的数量随着层中节点的数量呈指数增长。即使每个节点只取二进制值，网络也可以表示2^N个状态，其中N是隐藏层中的节点数。当每个节点的值为实数时，网络可以表示更多不同的状态。虽然网络的潜在表达能力随着节点数量呈指数增长，但推理和训练的复杂性最多呈二次增长。具有非线性激活的有限大小循环神经网络是一个丰富的模型家族，能够进行几乎任意的计算。一个众所周知的结果是，具有 sigmoid 激活函数的有限大小的循环神经网络可以模拟通用图灵机。循环神经网络执行任意计算的能力证明了它们的表达能力。在过去的 30 年中，循环神经网络已经从主要用于认知建模和计算神经科学的模型发展为用于从序列进行大规模监督学习的强大而实用的工具。这一进步归功于模型架构、训练算法和并行计算方面的进步。循环网络特别有趣，因为它们克服了传统机器学习方法对输入和输出数据施加的许多限制。对于循环网络，连续示例之间独立性的假设被打破，因此也打破了固定维度输入和输出的假设。尽管近年来 LSTM 和 BRNN

在许多任务上的准确性都创下了记录，但值得注意的是，这些进步来自新颖的架构，而不是来自根本上新颖的算法。神经网络提供了广泛的可转移和可组合技术。新的激活函数、训练程序、初始化程序等通常可以跨网络和任务转移。随着此类技术数量的增加，测试所有组合的变得不切实际。可以合理地推断，神经网络研究人员正在探索模型架构和配置的空间。正如刚才所说，这些研究可以从探索模型空间的自动化程序中受益。其次，当构建旨在执行更复杂任务的系统时，将受益于改进的适应功能。为此，在可能的情况下，谨慎的做法是先在具有已建立基准的数据集上使用经典前馈网络单独测试技术，然后再将它们应用于评估标准不太可靠的设置中的循环网络。最后，循环神经网络在自然语言任务上的迅速成功让人们相信，这项工作可以扩展到更长的文本应用中。

本文处理的是图像类的数据，那么就必须对循环网络进行改进。常用的方法就是卷积循环神经网络，也就是在图像输入循环神经网络之前将其先通过卷积神经网络进行特征提取和降维。这样图像中用于分类的有用部分被保留了下来，也不会增加循环神经网络的参数量。

§2.4 本章小结

本章对本文中涉及的主要研究内容进行了描述。首先，本章描述了医学图像分类问题和深度学习在其中的应用。然后，本章着重介绍了多示例学习模型，并且描述了多示例学习的历史和发展中遇到的困难和解决方法。最后，本章对循环注意力机制和通道域注意力机制进行了着重描述，并且详细介绍了循环神经网络。

第三章 多示例学习模型与基于 S3TA 的改进模型

本章将具体介绍多示例学习与 S3TA 的融合模型。首先，本章介绍对多示例学习进行改进的研究动机。然后，本章将描述改进的思路方法。接着，本章将结合具体的网络模型图对改进模型进行详细描述。

§3.1 改进多示例学习的研究动机

计算机辅助诊断(Computer-Aided Diagnosis, CAD)已成为医学图像分析的主要研究课题之一。计算机辅助诊断的研究内容有医学图像的分割、识别、检测和去噪。这些研究可以有效地帮助研究人员获得更准确的信息[1]。在深度学习算法和显卡(Graphic Processing Unit, GPU)不断发展的推动下，计算机辅助诊断也进入了一个新的研究阶段[2]。在使用深度学习研究医学图像的任务中，分类任务是最基础和最重要的任务之一。

在图像分类任务中，人们通常将小分辨率图像放入卷积神经网络(Convolutional Neural Network, CNN)中进行特征提取和分类。但是通过数字化整个病理组织切片获得的全视野数字切片(Whole Slide Image, WSI)的分辨率非常高。因为显存的限制，所以无法将整个 WSI 输入卷积神经网络[3-5]。因此研究人员需要用其他方法来处理 WSI，以便可以在卷积神经网络中对其进行训练。最常见的方法是将 WSI 切割成大量的图像块，并对这些图像块进行标记。这样就可以将图像块和相应的标签放入卷积神经网络中进行分类训练[6-14]。这种方式带来了另一个问题，就是对大量的图像块进行标注非常耗时费力。研究人员希望降低标记图像块的成本。为了解决这个问题，许多研究人员将目光投向了多示例学习(Multiple-Instance Learning, MIL)。多示例学习不需要一一标注图像块，只需要标注每个 WSI。裁剪一个 WSI 得到的图像块被分配到一个包中，这个包中所有图像块的标签与原始 WSI 的标签相同。最后，将这个包放入多示例学习网络进行训练。这样就达到了大大降低标注成本的目的。

§3.2 多示例学习的问题和改进方法

直接使用包标签作为包中所有图像块的标签进行训练会带来三个问题[16]。首先，阳性区域和阴性区域的图像块都会出现在阳性包中，常规卷积神经网络无法区分。多示例学习注意力机制[15]被提出来解决这个问题。其次，WSI 染色环境

的变化和着色污染会干扰最终的分类。最后，不同放大倍率下的病理组织特征不同，单一放大倍率的 WSI 切割的图像块无法捕捉不同放大倍率下的组织特征。在 2020 年的 CVPR 上，Noriaki Hashimot 等人[16]提出了使用多示例学习注意力机制、域对抗(Domain Adversarial, DA)归一化和多尺度(Multi-Scale, MS)技术[16]的融合模型来解决后两个问题，这也是本文的创意来源。

现在来分析一下这三种解决方案。首先，虽然多示例学习注意力机制成功地解决了无法区分图像块类别的问题，但该方法本身只是一种硬注意力机制，无法评估每个图像块的内部权重。虽然多尺度技术可以对评估图像块内部权重起到一些效果，但多尺度技术放大的区域不一定有用。其次，添加域对抗模块可以抵抗染色环境差异，但不能有效防御染色污染的干扰。域对抗使用不当会导致特征提取器欠拟合。最后，域对抗和多尺度技术的加入导致模型参数量急剧增加，并且训练被分为两个阶段。这些变化让人难以忍受。

本章提出使用一种受灵长类视觉系统启发的软的、连续的、平面空间的、自上而下的注意力机制(Soft, Sequential, Spatial, Top-down Attention mechanism, S3TA)来解决上述问题。该方法在文献[17]中提出，并在文献[18]中应用于 ImageNet 分类。灵长类动物眼中的中央凹可能与视觉系统中强烈的注意力瓶颈密切相关[19][20]。中央凹以不同的空间分辨率对视觉输入场的不同区域进行采样[21]。人类不会将图像视为静态场景，而是在一系列扫视或者注视中探索图像，并在此过程中收集并整合信息[22]。据推测，这会让人类大脑分析不同的分类错误。这些错误在性质上与深度神经网络的分类错误不同[23]。S3TA 具有视觉皮层的一些功能，即注意力瓶颈和顺序、自上而下的控制。这些功能可以完美解决上述问题。

首先，只需在多示例学习注意力机制中加入 S3TA 模块即可进行端到端的训练。这种方式简化了 MIL-DA-MS 模型的训练过程，并且极大地减少了参数量。S3TA 是一种软注意力机制，这种性质弥补了多示例学习注意力机制的缺陷。S3TA 的注意力瓶颈功能类似于灵长类动物视觉系统中的中央凹。它可以用不同的分辨率对不同的区域进行采样，并且在自上而下的控制模式下，网络可以自行找到感兴趣的区域。S3TA 可以通过增加循环步数来增强模型的鲁棒性。S3TA 的鲁棒性已被证明是迄今为止最好的，可用于防御各种 WSI 染色噪声。

§3.3 S3TA 和多示例学习网络结构

本实验中，由于每个图像块的尺寸很小，因此特征提取部分仅仅采用两层卷积神经网络。网络的主体部分采用两种网络模型：多示例神经网络和 S3TA 网络模型。本节将分别介绍两种网络模型的具体构造。

这里将训练集表示为$\{X_n, Y_n\}_{n=1}^{N}$，其中N表示训练集包的总数，X_n是第n个包

的图像块集合，以及$Y_n \in \{0,1\}$，它是一个标量代表第n个包标签。

§3.3.1 S3TA 模型

当前的神经网络模型范式一定程度上受到了人类和灵长类动物视觉系统的启发[28]。早期的网络模型直接建立了这种联系，并且有一系列工作将这种神经网络模型的激活与大脑中的神经活动联系起来[29]。模型和生物视觉系统之间的这些相似之处主要适用于早期视觉处理，特别是在时间有限的场景中发生的前馈处理[30]。这已经在一些文献中进行了讨论[32][33]。然而，前馈神经网络和灵长类视觉系统之间存在一些重要差异。灵长类动物的眼睛有一个中央凹，它以不同的空间分辨率对视觉输入场的不同区域进行采样[31]。此外该系统具有很强的注意力瓶颈，这可能与中央凹紧密相连，这已在许多不同的工作中进行了研究。视觉皮层有许多反馈和自上而下的循环连接[34]，它不是纯粹的前馈。此外，人类不会将图像视为静态场景，而是在一系列扫视或注视中探索图像的，并且在此过程中收集和整合信息[35]。据推测，这会使得人类大脑分析不同的分类错误，这些错误与深度神经网络的分类错误在性质上是不同的[36]。

本章使用一种软的、连续的、平面空间的、自上而下的注意力机制(Soft, Sequential, Spatial, Top-down Attention mechanism, S3TA)[18]。该注意力机制的灵感源于灵长类动物的视觉系统。虽然这不是一个生物学上合理的模型，但这个模型确实获得了视觉皮层的一些功能，即注意力瓶颈和连续、自上而下的控制。该注意力机制已经在 ImageNet 训练集上对模型进行对抗性训练，证明它对对抗性攻击具有最先进的鲁棒性。该注意力机制通过增加展开模型的步数就能够更好地防御更强的噪声干扰，这使得噪声和该注意力机制的防御性之间展开了计算竞赛。

该注意力机制来自于[17]提出的用于强化学习的模型，并且已经应用于 ImageNet 数据集的图像分类。该模型顺序查询输入，在每个时间步内积极参与分析相关的空间信息，以改进其对正确标签的预测。该模型的两个关键组成部分是模型的顺序性和自上而下的注意力瓶颈，并且已经被证明这两者都有助于其抵御对抗性攻击。

首先简要描述一下模型的重要组成部分。该模型首先将输入图像放入视觉网络也就是卷积神经网络。本文对所有时间步都使用相同的输入图像，因此卷积神经网络的输出只需要计算一次。然后将生成的输出张量沿通道维度拆分以生成 keys 张量和 values 张量。将这两个张量分别和一个固定的空间基张量相连接。该空间基张量使用傅里叶表示法对空间位置进行编码。这个空间基张量很重要，因为注意力瓶颈会在平面空间上求和，这就导致这些张量的空间结构消失。所以使用这个空间基张量可以用来传递空间位置信息。该模型为几个计算步骤展开自上

而下的控制器，在每个步骤中处理输入并通过控制器处理答案向量以产生输出或者说是下一个状态。自上而下的控制器是一个循环核，其先前的状态通过查询网络和多层感知机解码为一个或多个查询向量。每个查询向量具有与 keys 张量加上空间基中的通道数相同的通道数。该模型在每个空间位置获取查询向量与 keys 和空间基张量之间的内积，从而得到注意力 logits 的单个通道图。再将该图通过空间 softmax 函数来生成该查询的注意力图。然后将得到的注意力图与值张量和空间基础逐点相乘。将得到的张量在平面空间维度上求和，以产生一个答案向量，每个查询都能得到一个答案向量。这些答案向量作为当前时间步的输出，再将其输入到下一个循环核中。如果有多个答案向量，则将它们连接起来。最后一个循环核的输出会作为后面多示例学习注意力机制模块最前面的全连接层的输入。

该模型有几个注意点。首先注意力瓶颈使得模型的决策取决于图像的潜在范围。这可能是由于每个时间步的注意力图的形状，以及这些图在时间步之间可能发生很大变化。其次，注意力图会将所有值通道一起和单个通道相乘，这一事实限制了这些通道的内容在空间上是连贯的。在常规的卷积神经网络架构中，最后一个块输出是通过在每个通道上独立完成的平均池化来读取的，这使得网络在信息到达最后一层时会丢失空间结构。注意机制的自上而下性质使得查询向量是来自循环核的状态，而不是来自输入。因此，模型可以根据其内部状态主动选择相关信息，而不仅仅是根据输入选择。这点使得模型在查询图像和生成输出时会考虑到自身的不确定性。该模型的循环特性使得在该模型在不改变参数数量的情况下，通过增加时间步数来增加计算能力，这有助于提高模型的鲁棒性。

本文在多示例学习注意力机制中添加了调整和修改过的 S3TA，用来取代 MIL-DA-MS 模型。S3TA 使用类似于文献[25][26]中键值对和查询方式的注意力机制。但是，S3TA 没有使用自注意力机制，它们的查询直接来自输入。S3TA 使用的是循环核自顶向下地生成查询向量。由于病理图像的特征比 ImageNet 少，本文修改了 S3TA 的输入大小和空间基大小。由于 LSTM 参数较多，本文将 LSTM 替换为 GRU。

S3TA 模块的结构如图 3-1 所示。输入图像通过特征提取器生成一个 keys 和一个 values 张量。本文用 keys 和 values 张量在通道维度上连接到固定的、预定义的空间基。来自上一个时间步的 GRU 状态通过查询网络生成查询向量。这个查询网络是一个简单的两层感知器。将 query 向量和 keys 张量在每个空间位置进行内积运算，得到通道维度为 1 的 map。然后对整个 map 进行 softmax 操作，生成注意力图。在通道域维度广播注意力图后，逐点乘以 values 张量，将结果进行空间求和，得到答案向量。这个答案向量是 GRU 在当前时间步的输入。这里将完整的问答系统称为注意力头。每个时间步可以有多个注意力头。因此，可以

在每个时间步生成多个查询向量和对应的答案向量。在图 3-1 中，在一个时间步中只绘制了一个注意力头。

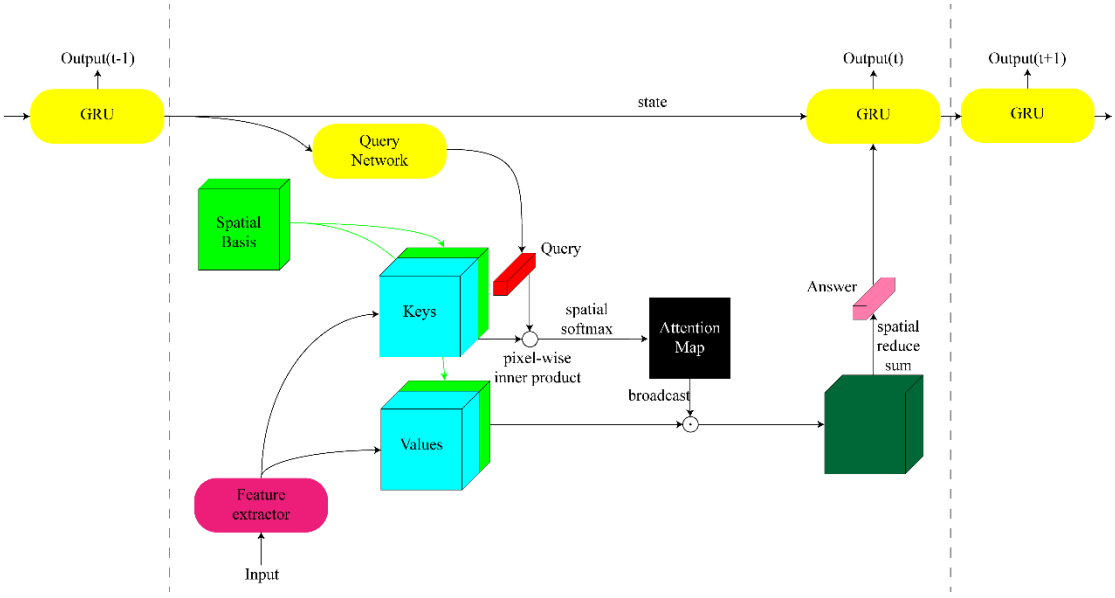

图 3-1 S3TA 模型结构

空间基$S^{h\times w\times c}$使用空间求和后可以保存空间位置信息，其中h,w,c分别为高度、宽度和通道数。在空间基中，每个位置i,j的通道将对空间位置信息进行编码。本研究使用傅立叶空间基。给定两个频率u,v，位置i,j的一个通道的傅里叶空间基公式如下：

$$S_{i,j,(u,v)} = f_1\left(\frac{\pi u i}{h}\right) f_2\left(\frac{\pi v j}{w}\right), \qquad (3-1)$$

其中f_1和f_2可以是sin或cos函数，它们固定在一个通道中。因此，两个频率可以产生 4 个通道。在本研究中,u和v各取一个固定值以减少参数量。空间基在模型中始终是恒定的。

§3.3.2 多示例学习注意力机制模型

多示例学习是一种弱监督学习问题，每个包中的所有示例共享一个包标签。多示例学习已成功应用于组织病理学图像的分类[13][27]。多示例学习注意机制是在文献[15]中提出的。该注意机制通过分配权重来确定每个图像块提取的特征的重要性。通过这种方式，网络模型可以自行评估图像块所在的类别区域的可能性。文献[15]中应用于结肠癌数据集的多示例学习注意力机制的实验结果将作为本研

究的对比基准。

由上文可知，S3TA 的输出是一组特征向量$h = \{h_i^{l\times 1}\}_{i=1}^{b}$，其中$b$是包中的图像块个数，$l$是每个特征向量的维度。多示例学习注意力机制的公式如下：

$$z = \Sigma_{i=1}^{b} a_i h_i, \tag{3-2}$$

其中a_i的公式如下：

$$a_i = \frac{\exp(w^T \tanh(Vh_i))}{\Sigma_{j=1}^{b} \exp\left(w^T \tanh\left(Vh_j\right)\right)}, \tag{3-3}$$

其中$V^{s\times b}$和$w^{s\times 1}$分别是可训练的参数矩阵和向量，其中s是超参数。h经过多示例学习注意模块后，得到一个向量$z^{l\times 1}$，z进入分类器得到最终结果。从上述公式可以，向量$a^{b\times 1}$决定了所有特征向量的重要性。

§3.4 S3TA 模块对多示例学习的改进模型

本文提出的卷积神经网络如图 3-2 所示。它由 4 个主要部分组成。第一部分是特征提取器，由卷积层和池化层组成的两个模块组成。第二部分和第三部分分别是 S3TA 和多示例学习注意力机制模块，最后一部分是全连接层分类器。

本文搭建的最终的网络模型如图 3-2 所示。首先输入的数据集都是以包为单位，通俗来说，一个包就是一个批次。与很多网络模型不同的是，这里每个包的数量或者说是批次数量不是固定的，这取决于数据集的预处理工作。数据集预处理并不是本文的工作，这里不过多介绍。根据多示例学习的定义，每个包里的所有示例的标签都是相同的，如果一个包里的所有示例的实际标签都为阴性，则这个包的标签就是阴性；反之而言，一个包里只要有一个示例的实际标签为阳性，则不管这个包里其余示例的实际标签是否为阴性或者阳性，这个包的标签就是阳性。在实际模型中，1 表示阳性，而 0 则表示阴性。每个包输入模型后，首先被放入特征提取器。因为输入网络的每个图像块的大小为27×27，这对卷积神经网络而言是一个比较小的输入尺寸，所以这里不需要使用一个超大的特征提取模块进行提取。本文在特征提取模块仅仅使用了两个卷积层，每个卷积层之后都有一个池化层。在这里就不得不对比一下文献[16]中的 MIL-DA-MS 模型，该模型中的特征提取部分为了配合后面的域对抗模块和多尺度策略，直接将原本的27×27输入尺寸变为224×224的输入大小。并且该模型的特征提取器采用了庞大的 VGG16 模型，再加上多尺度策略，这使得整个网络模型变得庞大而复杂。WSI 中的特征和 ImageNet 数据集中的特征完全不是一个数量级的，所以不需要用如此庞大的特征提取器进行特征提取。

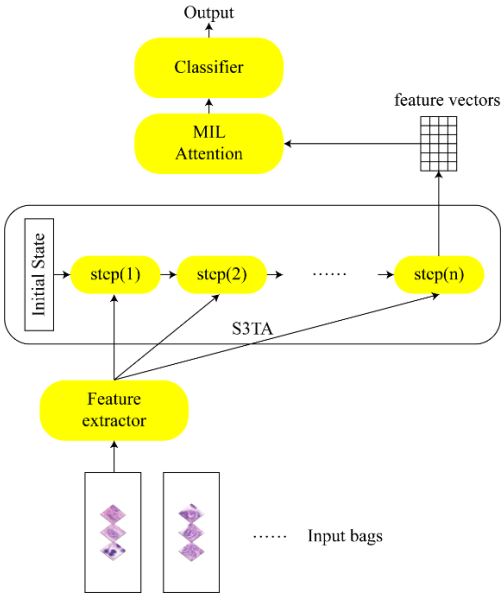

图 3-2 MIL-S3TA-n 网络模型

通过特征提取得到一个$5 \times 5 \times 48$的特征张量。将这个特征张量送入 S3TA 模块进行一次软注意，这恰好弥补了多示例学习注意力机制模型只有硬注意力机制的缺陷。注意，S3TA 是一个循环神经网络，需要一个有时间维度的特征张量输入。这里因为每个时间维度上的输入都是相同的，所以只需要对每个图像块进行一次特征提取，再将其复制多份放入 S3TA 中即可。这里所说的复制的份数也就是上文所说的在时间步上进行展开的步数。理论上来说，展开的步数越多，鲁棒性越高，训练时间也越长。S3TA 的这种特性恰好可以代替 MIL-DA-MS 模型中的域对抗模块的作用，而且所消耗的参数量远远低于前者，效果却优于前者。该模块也是原多示例学习注意力机制模型中所没有的。每个图像块经过 S3TA 处理之后得到一个128×1的特征向量，将其输入到全连接神经网络中。因为这里的全连接神经网络是常用模块，所以并没有在图 3-2 中显示。但是这里需要注意的是，原来的多示例学习注意力机制模型里面经过特征提取之后的特征张量会被直接展开送入后面的全连接神经网络中，所以这里产生的参数量相当巨大。而经过 S3TA 处理之后的特征向量的维度只有 128，所以在送入全连接神经网络之后所产生的参数量比原多示例学习注意力机制模型的参数量大大减少。实验统计的参数量显示新模型的参数量比原多示例学习注意力机制模型的参数量下降了一半。这也进一步说明了本文提出的模型的优越性。

每个包经过全连接神经网络得到一组特征向量。这组特征向量经过多示例学习注意力机制模块之后，从一组向量被压缩为一个特征向量。该向量进入最终的包分类器得到最终的分类结果。这里最终的分类函数采用 sigmoid 函数。

§3.5 本章小结

首先，本章介绍对多示例学习注意力机制进行改进的研究动机和灵感来源。然后，本章对新模型使用的各个模块进行详细描述，并且对 S3TA 模块和多示例学习注意力机制模块的所有细节和性质都做了充分说明。最后，本章对新模型的运行过程进行了详细描述。

第四章 结肠癌分类的应用系统

本章将对结肠癌分类应用进行介绍。首先，本章对结肠癌分类任务进行介绍。然后，本章介绍了结肠癌数据集。接着，本章介绍了实验环境。最后，本章着重分析了实验结果和影响因素。

§4.1 结肠癌分类任务

结肠癌是发生于结肠的消化道常见恶性肿瘤。结肠癌也称为结肠直肠癌。它影响由结肠和直肠组成的大肠。人体内的细胞通常以受控方式分裂和生长。当癌症发展时，细胞会发生变化并以不受控制的方式生长。大多数结肠癌是由被称为息肉的癌前生长物发展而来的。但并非所有的息肉都会发展成癌症。如果医生发现息肉，可以将它们移除以防止它们变成癌症。癌细胞可能会留在肠道中，也可能会扩散到身体的其他部位，例如肝脏或肺部。结肠癌是可以被治疗和治愈的。如果在最早阶段被诊断出来，几乎每个人都能幸免于结肠癌。然而，如果任由结肠癌发展，生存率将显著下降。早期诊断确实可以挽救生命。自 1970 年代以来，死于结肠癌的人数一直在下降。这可能是由于早期诊断和更好的治疗导致的。因为组织病理学图像是癌症诊断的金标准，所以人们想要在早期发现和诊断结肠癌，就需要对组织病理学图像进行研究。

§4.2 结肠癌数据集

结肠癌的 WSI 被苏木精和伊红(hematoxylin and eosin, H&E)染色。实验直接使用已经被预处理过的结肠癌数据集，也就是文献[15]中所使用的结肠癌数据集 ①。原数据集包含 100 张 H&E 染色图像。预处理后的数据集有 99 个包，其中 48 个是阴性包，51 个是阳性包，其中每个包的图像块数量各不相同，而且数量相差过大，所以这里不给出相关统计性描述。每个图像块的大小为27×27。如果一个包中含有一个或多个上皮细胞核，则给予阳性标记，否则将给予阴性标记。因为结肠癌起源于上皮细胞[24]，标记上皮细胞和临床实验信息高度相关。

① https://github.com/utayao/Atten_Deep_MIL，2018 年

§4.3 结肠癌分类实验环境的搭建

实验中所有的模型使用的 GPU 是 RTX 2070，CPU 是 Core i7。用于构建模型的深度学习框架是 Tensorflow2.1。这里将介绍实验中的超参数设置。模型采用 Adam 优化器，学习率为 0.0001，其中参数β_1和β_2分别为 0.9 和 0.999。损失函数采用交叉熵损失函数，准确率是预测正确的包的数量占所有用于测试的包数的比例。模型中所有可训练参数均采用L_2正则化，其中参数λ的大小为 0.0005。S3TA 中 keys 和 values 张量的通道数分别为 32 和 16。S3TA 的输出维度为 128。S3TA 和多示例学习模块之间有两个全连接层，都有 512 个神经元，并且都使用 dropout 层，其中参数设为 0.5。多示例学习注意力机制中的超参数s设置为 128。所有实验都使用 5 折交叉验证。值得注意的是，很多研究人员经常在深度学习框架的训练中使用交叉验证来观察模型的准确率和损失函数值的变化，这里的交叉验证结果是不能作为最终测试集上的结果的。所以很多人会误以本文的交叉验证和这里的交叉验证结果是一回事。这里着重强调 5 折交叉验证。

首先将整个数据集以包为单位进行打乱，然后平均分为 5 份。对于其中一折交叉验证来说，就是用其中一份为测试集，剩下的为训练集，对一个独立的没有训练过的模型进行训练，所得结果作为该折交叉验证的结果。其他折的交叉验证均是如此。也就是说，5 折交叉验证就将 5 份数据集轮流作为测试集来训练 5 个独立的模型。所得的交叉验证结果可以作为最终结果。

§4.4 结肠癌分类实验结果与分析

在本节中，本文的实验结果说明了 S3TA 在结肠癌数据集上对多示例学习注意力机制改进的合理性。为方便起见，本文将 S3TA-n表示为具有n个时间步长的 S3TA 模型，并对 S3TA-2、S3TA-4 和 S3TA-8 进行实验。本文将这些模型与原始注意力机制进行比较，所得实验结果如表 4-1 所示。

表 4-1 原多示例学习注意力机制模型和改进模型对比

Method	Accuracy	Precision	Recall
Attention-based MIL	0.904 ± 0.011	0.953 ± 0.014	0.855 ± 0.017
MIL-S3TA-2	0.906 ± 0.086	0.880 ± 0.121	0.933 ± 0.088
MIL-S3TA-4	0.885 ± 0.071	0.900 ± 0.095	0.889 ± 0.109
MIL-S3TA-8	**0.936** ± 0.039	0.922 ± 0.075	**0.960** ± 0.080

注意表 4-1 中第一个模型 Attention-based MIL 为文献[15]中给出的数据，本文自己构建该模型时得到的精确率数据略低于该数据，Accuracy、Precision、Recall 分别为精确率、查准率、查全率，该标准均为基础内容，此处不过多说明。从表 4-1 可以看出，提出的模型在展开 2 步时，准确率与原始模型几乎相同。当模型展开 4 步时，虽然平均准确率有所下降，但中位数准确率与前两个模型相似。当模型展开 8 步时，平均准确率达到 93.6%，远高于前三个模型。S3TA 的展开步数越多，精度越高。这里已经证明了扩展达到步数达到 8 步即可获得巨大性能提升但是训练时间已经比原多示例学习注意力机制要高出一些，考虑到继续扩展步数会使时间增加，因此不再增加步数继续测试，这里扩展 8 步即为性价比最高的模型。

在文献[19]和文献[20]中，两个 S3TA 模块中使用的循环核都是 LSTM。但是，上述所有模型中使用的循环和都是 GRU。为了证明用 GRU 代替 LSTM 的行为是正确的。在表 4-2 中，实验结果比较了两个模型的效果。

表 4-2 GRU 和 LSTM 在改进模型的效果对比

Method	Accuracy	Precision	Recall
MIL-S3TA-8 with LSTM	0.875 ± 0.091	0.887 ± 0.111	0.879 ± 0.098
MIL-S3TA-8 with GRU	**0.936** ± 0.039	0.922 ± 0.075	**0.960** ± 0.080

出现这种情况的原因是 LSTM 中的参数太多。对于输入为特征较少的医学图片并且图片尺寸仅为27×27的模型，LSTM 过大且容易造成过拟合。LSTM 更适合 ImageNet 等具有更多特征和更大图像尺寸的数据集。原多示例学习注意力机制模型的参数总数为 961,109。这里一定有人认为改进模型比前者使用了更多的可训练参数。事实上，新模型只使用了 502,557 个可训练参数。后者使用的参数量几乎是前者的一半。造成这种现象的原因是 S3TA 将特征图变成了低维特征向量，而不是将特征图展平直接放入全连接神经网络。虽然训练参数有所下降，但训练时间会随着时间步数的增加而增加。原始的多示例学习注意力机制模型训练 100 个 epoch 只需要 2.19 分钟，而 MIL-S3TA-8 模型需要 2.98 分钟。考虑到本文模型的性能提升，稍微增加训练时间对研究人员来说是完全可以接受的。实验中有很多参数设置会影响实验结果。学习率对模型训练的影响更大。本文尝试将学习率分别设置为 0.001 和 0.00001。但是得到了两个非常糟糕的结果。因为学习率太高，模型的结果会一直徘徊在最优解附近，而学习率太低会导致收敛太慢。因此，本文将学习率设置为 0.0001 是合理的。另一个非常重要的参数设置是空间基的尺寸。在实验中，我们尝试将新模型的空间基的大小设置为27×27，

与每个图像块的大小相同，但是所得结果非常糟糕。本文分析这是由于参数过多造成的。经过多次尝试，本文最终将空间基的大小设置为5×5。以上所有模型的空间基均采用该尺度。

§4.5 本章小结

首先，本章对结肠癌和数据集做了简要描述。然后，本章使用该数据集对第三章提出的新模型进行训练和测试。从实验结果可以可知，第三章提出的新模型比原始多示例学习注意力机制的分类准确率更高，新模型可以通过更换循环核和超参数来提高分类准确率。

第五章 基于通道域注意力对多示例学习的改进和应用

本章将从通道域角度对原始多示例学习注意力机制进行改进。虽然第三章的新模型已经对原多示例学习注意力机制[15]做出了巨大改进，但是本章对新模型进行分析后认为第三章的新模型仍然有上升空间。首先，本章分析了 S3TA 模型的缺陷。然后，本章分析了通道域注意力机制，并且将其与多示例学习注意力机制相结合。最后，本章通过实验证明该模型的分类准确率有所提高。

§5.1 改进通道域注意力机制的研究动机

原多示例学习注意力模块属于硬注意力机制，从三维空间的角度理解它属于平面域上的注意力机制。所以对多示例学习注意力机制的改进可以加入软注意力机制进行改进，比如前文加入的改进的 S3TA 模块就是一个很好的软注意力机制。但是 S3TA 模型的主要用途是针对平面域的注意力机制，而且 S3TA 模型过于复杂。因此本节提出使用一种通道域上的注意力机制对多示例学习注意力机制进行改进。

通道域注意力机制相比于平面域注意力机制较少。这是因为平面域属于二维空间，其中的像素点数量远远超过通道数量。所以对通道域进行优化相比于平面域更加简单。说起通道域注意力机制，就不得不提到大名鼎鼎的 Squeeze-and-Excitation (SE)模块。以 SE 模块为基础的 SENet 在 2017 年的 ImageNet 分类竞赛中获得第一名的成绩。并且将 top-5 的错误率降低到了 2.251%，比 2016 年提升了约 25%。可以看出 SE 模块对性能的提升很有帮助，而且提升性能的成本很低。下面来看一下 SE 模块的运作原理。

一般的卷积层使用多个卷积核对整个图像进行卷积操作，每个卷积核的初始参数都是随机且互不关联的。然后由卷积核生成的对应特征图的各个通道域之间也具有这种性质，而且无论后面训练多少次，都不会产生变化。所以文献[37]针对这一问题提出了 SENet。SENet 就是由一系列的 SE 模块组成。SE 模块通过显示地建模对卷积层中各个通道的相互依赖关系进行表示，从而让网络可以对各个通道域的权重进行重新校准，也就是说通过 SE 模块可以在通道域中对重要信息进行强调并且抑制无效特征。

虽然 SE 模块的功能很强大，但是其工作原理却很简单。SE 模块对特征张量的处理流程可以分为两个步骤。第一步就是对原有特征张量进行挤压，也就是通过聚合特征图来生成一个通道描述符。描述符的功能就是对每个特征通道进行总

结并且允许网络其他层可以使用这些总结信息，这一步可以通过使用全局平均池化的操作实现。第二步继续对生成的通道描述符进行激发，目的就是让网络根据通道描述符进行总结，最后自行挑选出相对有用信息，这一步可以由常规的全连接层完成。

通过上述描述可以看出 SE 模块的结构非常简单。众所周知，创造新的卷积神经网络模型非常困难，这是因为不仅需要对众多超参数进行选择，而且每个层的配置也是大问题。对比之下，SE 模块简单高效且成本小的优点就非常突出。SE 模块可以简单地堆叠来构建 SENet，也可以直接嵌入其他模块中。所以本文想到使用 SE 模块对多示例学习注意力机制进行改进。

§5.2 SE 模块的问题和改进方法

SE 模块在网络的不同深度的作用是不同的。所处位置较深的 SE 模块通常能取得很大成效，这是因为深层卷积层的卷积核的数量非常大，所以生成的特征张量的通道数也非常多。SE 模块能获取到的通道间的相互依赖关系的信息也就越大。但是在层数较浅的 SE 模块效果就不尽如人意，原因和上面相反，特征张量在前几层的通道数非常少，所以效果自然很差。原本的多示例学习注意力机制也只有两层卷积层，整个模型和 VGG 和 ResNet 等著名网络相比，层数算是少的可怜。所以如何在层数如此之少的多示例学习注意力机制模型中使用 SE 模块就成了重要问题。

有两种方案可以解决这个问题。第一种方案就是增加原模型的网络深度，尤其是特征提取器部分的深度。第二种方案就是对 SE 模块进行改进。第一种方案看起来非常简单粗暴。虽然本文使用的数据集的输入尺寸和网络规模都很小，不适合大规模加深网络，但是为了更有说服力，本文也会适当添加几层。我们想要大幅提升 SE 模块性能，还是要从第二种方案入手。

我们将 S3TA 模块和 SE 模块进行了反复对比研究发现，SE 模块之所以在浅层性能差强人意，是因为在 SE 模块在压缩阶段直接使用了全局平均池化导致平面位置信息丢失。为了防止平面位置信息丢失，S3TA 在 keys 和 values 张量中分别加入了一个固定不变的空间基。所以，我们仿照这个处理方案，在 SE 模块使用全局平均池化压缩之前，将特征张量和空间基在通道维度进行合并。

这个做法有三个好处。第一，SE 模块可以在压缩全局信息时保存平面位置信息。第二，因为空间基都是在整个模型训练之前就已经计算好的，所以不占用训练时间。第三，最重要的是 SE 模块可以通过不加深网络层数的形式就增加了通道数量。这不仅没有增加前向传播的计算时间，而且增加的训练时间也很少。下面具体介绍一下原有的 SE 模块及其改进模型。

§ 5.3 SE 模块及其改进模块

本节将详细介绍原 SE 模块及其改进模块。

§ 5.3.1 SE 模块

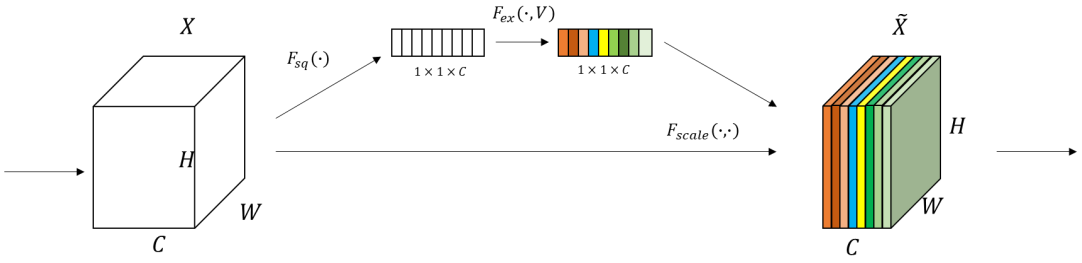

图 5-1 SE 模块

SE 模块如图 5-1 所示。SE 模块的输入为经过卷积神经网络处理的特征图X，输出是一个新的特征图$\tilde{X}$，两张特征图的尺寸均为$H \times W \times C$。其中H, W, C分别为高度和宽度还有通道数。首先是压缩操作，其目的是获取各个通道的全局感受野的特征信息。这一步通常由全局池化(Global Pooling)完成，这里的全局池化一般采用全局平均池化(Global Average Pooling, GAP)，也就是求每个通道c的特征图的平均值[37]：

$$z_c = F_{sq}(\boldsymbol{x}_c) = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\boldsymbol{x}_c(i,j), c \in \{1,2,\dots,C\}, \quad (5-1)$$

其中$\boldsymbol{x}_c$表示特征张量X的第c个通道的特征图，z_c表示$\boldsymbol{x}_c$经过压缩操作之后得到的全局感受野的平均值，所有的z_c构成特征向量$\boldsymbol{z}$。然后进入到激发操作，这一步的目的是通过$\boldsymbol{z}$学习所有通道的相关性。这一步需要够灵活和简单，而且必须真实有用。根据这些需求，这里采用两个全连接(Fully Connected, FC)层构成的门机制完成，其中门控单元$\boldsymbol{s}$的计算公式如下[37]：

$$\boldsymbol{s} = F_{ex}(\boldsymbol{z}, V) = \sigma\big(g(\boldsymbol{z}, V)\big) = \sigma\big(V_2\delta(V_1\boldsymbol{z})\big), \quad (5-2)$$

其中δ表示 ReLU 激活函数，$V_1 \in \mathbb{R}^{\frac{C}{r}\times C}, V_2 \in \mathbb{R}^{C\times\frac{C}{r}}$分别是两个全连接层的权值矩阵，$r$是隐藏层的神经元个数一般为$\frac{C}{2}$。$\sigma$表示 Sigmoid 函数。最后将门控单元$\boldsymbol{s}$和特征张量$X$作为输入执行$F_{scale}$放缩操作，计算公式如下[37]：

$$\tilde{\boldsymbol{x}}_c = F_{scale}(\boldsymbol{x}_c, s_c) = s_c \cdot \boldsymbol{x}_c, \quad (5-3)$$

其中$\tilde{\boldsymbol{x}}_c$为新特征张量$\tilde{X}$在第c个通道中的特征图。以上就是 SE 模块的全部流

程。从中可以看出该注意力机制就是在通道维度上学习一组权值。

§ 5.3.1 SE 改进模块

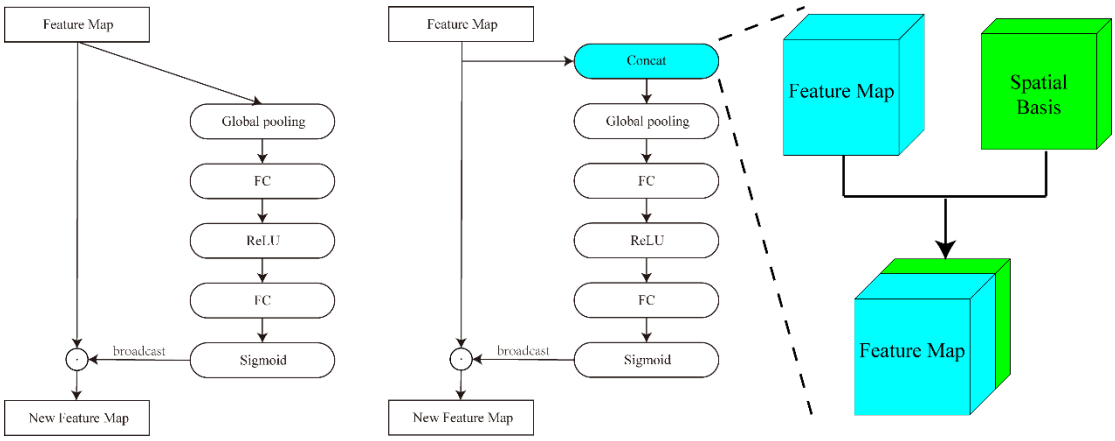

图 5-2 SE 模块(左)和 SE 改进模块(右)

SE 改进前后的模型图 5-2 所示，左侧是原 SE 模块，右侧是 SE 的改进模块，其中⊙表示哈达玛积，也就是两个尺寸相同的张量对应元素相乘。在执行哈达玛积之前 Sigmoid 函数之后需要对$1 \times 1 \times C$的特征向量在平面维度做一次广播(broadcast)操作，也就是在整个平面都是使用一个相同的特征值。可以看出整个模块唯一改动的地方就是在特征图进入全局平均池化之前执行了一次连接操作，将原特征图和一个空间基在通道域上相连接。这里的空间基和 S3TA 中使用的空间基都一样，都是在模型训练之前就已经计算好的。可以看出 SE 改进后的模型依然非常简单，对原 SE 模块的运算并不会产生多大影响。下面本文会用上面两个模块分别进行实验，并且对比两者在浅层卷积神经网络中的作用，并且会尝试对加入 SE 模块的新模型进行加深，对比这几个模型的效果。

§ 5.4 SE 模块及其改进模块对多示例学习的改进模型

本节将介绍 SE 模块及其改进模块对原多示例学习注意力机制模型的改进。前文中使用 S3TA 对原多示例学习注意力机制进行改进时，是通过在原多示例学习注意力机制模型的特征提取器和全连接层之间插入改动过得 S3TA 模块，无法像 SE 模块这样灵活地插入各种地方。原本特征提取部分是由两层卷积神经网络和两层最大池化层来处理的，所以对于 SE 模块这样的通道模块来说是非常浅的深度。这种浅层的卷积神经网络中嵌入一个 SE 模块不但浪费训练时间和参数量，而且可能出现负面效果。所以本节将设计出几个新模型用来进行对比。

§5.4.1 添加 SE 模块或改进模块的多示例学习注意力机制

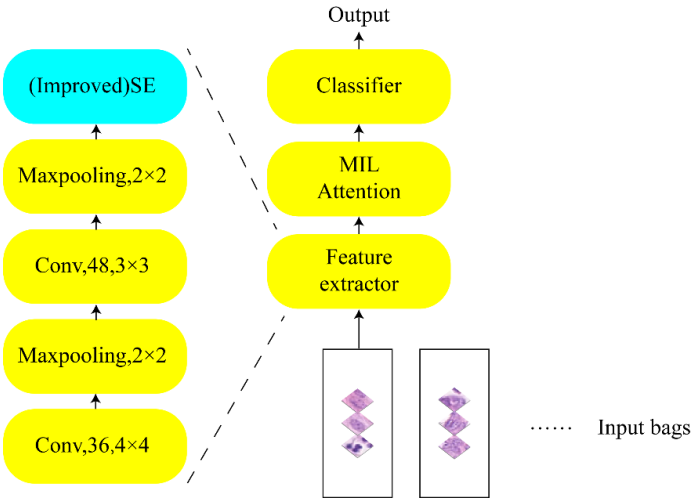

图 5-3 添加 SE 模块或改进模块的新模型

在原多示例学习注意力机制模型中单独添加 SE 模块或者改进模块的模型如图 5-3 所示。前两个卷积层和两个最大池化层都没有改变，也就是在全连接层和原来的特征提取器之间分别添加了 SE 模块或者 SE 改进模块，和上文中 S3TA 对多示例学习注意力机制模型的改进相同。这么做的目的是为了可以和原有模型进行对比时更有说服力。

§5.4.1 添加 SE-ResNet 或改进模块的多示例学习注意力机制

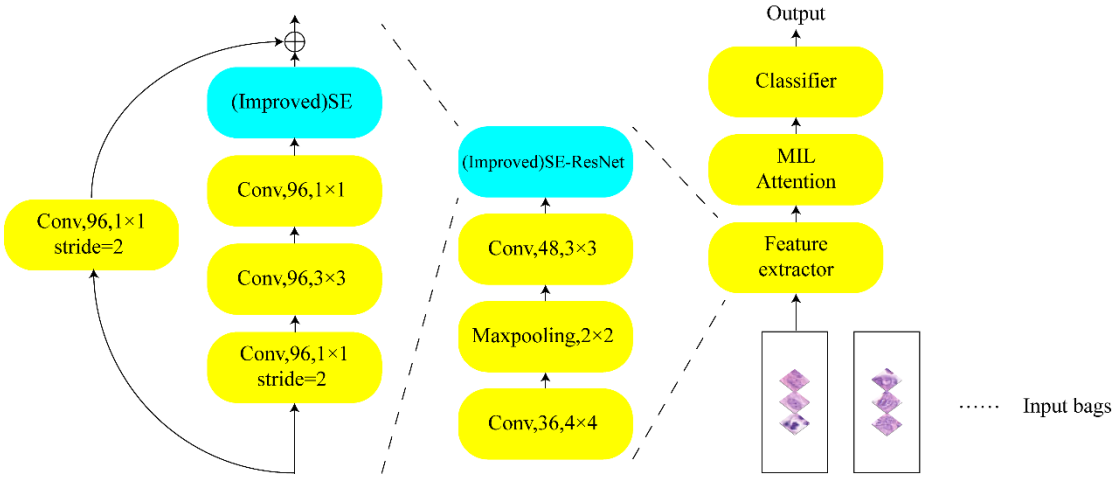

图 5-4 添加 SE-ResNet 或改进 SE-ResNet 的新模型

为了验证加深网络深度来获得更多的方案是否有效，本文也设计了添加两个分别添加 SE-ResNet 和改进 SE-ResNet 的模型。该模型为了将通道数增加，

先把原特征提取器最后一个最大池化层去掉，然后添加一个 ResNet 模块。该模块把第一个卷积层的卷积核设为前一层的两倍，且根据 ResNet 的设计原则，将步长增加为 2。在 ResNet 模块的短路支路的卷积层中也是如此。在 ResNet 模块中的残差块之后添加 SE 模块或者改进的 SE 模块，构成了最后的 SE-ResNet 和改进的 SE-ResNet 模块，最终的模型图如图 5-4 所示。下面本文将使用上述几个模型进行实验。

§5.5 结肠癌分类实验和分析

本章实验所用的数据集、实验环境、训练参数和模型中未改变的模块的超参数都与上一章相同。通过对上述模型进行训练，我们得出的实验结果如表 5-1 所示。

表 5-1 实验结果

Method	Accuracy	Precision	Recall
Attention-based MIL	0.904 ± 0.011	0.953 ± 0.014	0.855 ± 0.017
MIL-SE	0.872 ± 0.069	0.861 ± 0.081	0.895 ± 0.097
MIL-Improved SE	**0.915** ± 0.071	0.947 ± 0.054	0.912 ± 0.078
MIL-SE-ResNet	0.654 ± 0.754	0.556 ± 0.538	0.723 ± 0.680
MIL-Improved SE-ResNet	0.734 ± 0.576	0.821 ± 0.482	0.617 ± 0.682

从表 5-1 可以看出，所有的模型中只有 MIL-Improved SE 比原模型的结果要高。MIL-SE 的结果比原模型结果要略低，本文认为这是由于 SE 模块在深度学习模型前几层添加造成的负面影响。MIL-SE-ResNet 和 MIL-Improved SE-ResNet 模型的表现就是灾难级的，这两种模型是根据第一种加深网络深度来增加通道数的方案设计出来的。由此证明，在一个小数据集且网络深度很小的模型中随便添加 SE 模块或者加深网络深度之后添加 SE 模块都是不可行的。因为小数据集的特征很少，所以很容易就产生过拟合的现象。MIL-Improved SE 模型却能在只嵌入在浅层网络中就能够表现的比原模型更好，这是因为它没有通过加深网络深度的方式来增加通道数，而是通过添加空间基这种训练前就计算好的张量来增加通道数。而且空间基也保存了特征图的平面位置信息，一举两得。我们进一步尝试了在 MIL-Improved SE 的改进 SE 模块之后添加第三章的 S3TA-8 模块，实验结果如表 5-2 所示。

表 5-2 融合改进的 S3TA 和 SE 模块实验结果

Method	Accuracy	Precision	Recall
MIL-S3TA-8	0.936 ± 0.039	0.922 ± 0.075	0.960 ± 0.080
MIL-Improved SE -S3TA-8	**0.942** ± 0.049	0.923 ±0.095	0.961 ± 0.048

MIL-Improved SE-S3TA-8 模型的实验结果如表 5-2 所示。由上表可知实验的准确率虽然增加，但是涨动幅度没有单独添加改进的 SE 模块大。本文认为这再次证明了小数据集的训练模型增加参数量会带来过拟合的后果。

§5.6 本章小结

本章主要对 SE 模块进行改进,并且将其成功添加到多示例学习注意力机制模型中。虽然本章对 SE 模块只进行了一个很小的改动，但是改进的 SE 模块使得多示例学习注意力机制的分类准确率提升明显，并且改进的 SE 模块的适用范围很大。虽然 S3TA 模块也使得多示例学习注意力机制的分类准确率有所提升，但是 S3TA 的适用范围很小。S3TA 只能在特征提取器和全连接层之间加入。SE 模块和改进的 SE 模块都具有简单、灵活和实用有效的特点，两者都能在网络的不同层次之间任意插入。根据插入位置深度的不同，改进的 SE 模块可以选择合适的空间基通道数。改进的 SE 模块插入的位置越靠前，空间基的通道数越多。本章对 SE 模块的改进不仅提升了多示例学习注意力机制的分类准确率，还使得改进的 SE 模块在其他网络模型中的适用范围更大。

第六章 总结与展望

病理学图像是临床诊断的黄金标准，也是计算机辅助诊断的主要研究对象。近年来，因为传统病理学图像分类的研究工作日趋完善，所以很多研究人员转向了其他研究领域。弱监督学习一定会使病理学图像分类的研究重新焕发活力。本章将对病理学图像分类任务的研究工作进行总结，并且给出未来相关研究的发展方向。

§6.1 工作总结

常规深度学习网络进行病理学图像分类时，所有图像块都带有标签。虽然常规深度学习的分类准确率很高，但是这种方法有一个严重的问题，那就是标记图像块的成本过高。很多研究人员认为多示例学习方法可以解决该问题，但是这种方法依然存在三个问题。

首先，阳性包中既包含阳性区域的图像块，又包含阴性区域的图像块。常规网络模型无法判断阳性包中的图像块是否属于阳性区域，这导致分类准确率大幅下降。很多研究人员使用多示例学习注意力机制来解决这个问题。其次，WSI 染色环境的变化和着色污染也会干扰最终的分类准确率。最后，在不同放大倍率下，WSI 中的病理组织特征是不同的。在单一放大倍率下，被切割的 WSI 图像块无法捕捉不同放大倍率下的组织特征。在 2020 年的 CVPR 上，Noriaki Hashimot 等人[16]提出使用多示例学习注意力机制、多尺度技术和域对抗归一化相结合的方法来解决后面两个问题。

本文的第一项研究工作是通过对该方法的一些缺陷进行分析，然后提出改进方法，最后进行实验。该方法在原有的多示例学习注意力机制中插入 S3TA 模块，该方法有以下优点。首先，虽然多示例学习注意力机制成功地解决了无法区分图像块类别的问题，但该方法本身只是一种硬注意力机制，它无法评估每个图像块的内部权重。虽然多尺度技术可以对评估图像块内部权重起到一些效果，但是多尺度技术放大的区域不一定有用。S3TA 本身就是一个软注意力模块，它可以对每个图像块内部进行软注意。这一点弥补了多示例学习注意力机制无法对图像块内部进行权重评估的缺陷。其次，虽然多尺度技术能够帮助网络模型观察到不同倍率下的病理学组织特征，但是该技术使得整个模型的参数量剧增。更糟糕的是，整个模型的训练流程被拆分为两个阶段。S3TA 具有视觉皮层的一些功能，即注意力瓶颈和顺序、自上而下的控制。灵长类动物眼中的中央凹可能与视觉系统中

的注意力瓶颈密切相关。中央凹以不同的空间分辨率对视觉输入场的不同区域进行采样。人类不会将图像视为静态场景，而是在一系列扫视或者注视中探索图像，并在此过程中收集并且整合信息。这种性质帮助 S3TA 完美模拟了医生经常进行缩放观察病理学图像的行为。因为多尺度技术只能观察几个固定倍率下的图像，所以 S3TA 比多尺度技术更加灵活。MIL-S3TA 模型只需要进行端到端的训练。因为多尺度技术需要进行 2 阶段训练，所以 MIL-S3TA 模型训练更加简单。最后，虽然模型中使用域对抗模块可以对染色环境差异进行预防，但是对病理学图像其他的染色噪声却无能为力。因为 S3TA 模型[18]已经被证明拥有最好的鲁棒性，所以在对抗噪声方面 S3TA 和域对抗模块相比是有过之而无不及。并且 S3TA 模块不会产生欠拟合的情况。虽然 MIL-S3TA 模型在多示例学习注意力机制中插入了一个新模块，但整个模型的参数量相比于原模型下降了一半左右。之所以参数量不增反减，是因为 S3TA 模型在特征提取之后对特征张量进行了降维处理。从实验结果可知，MIL-S3TA 模型相比于原多示例学习注意力机制有着更高的分类准确率。该模型还可以根据不同的任务选择合适的循环核，如 GRU 和 LSTM。总之，该模型大大改善了原多示例学习注意力机制的性能。

本文的第二项研究工作是从通道域维度对原多示例学习注意力机制进行优化，并且通过实验证明改进的 SE 模块具有以下优点。首先，原 SE 模块在网络的浅层中加入效果一般，甚至可能起到负面效果。本文分析了 SE 模块产生该缺陷额的原因，并且在 SE 模块中加入了空间基。改进的 SE 模块可以嵌入到深度学习模型的浅层中。其次，改进的 SE 模块保持了简单灵活且实用有效的特性。研究者可以调整空间基通道数来适应网络模型中的各种层次。该模块插入的层次越靠前，空间基通道数越高。最后，虽然 S3TA 模型内部有对通道域的改进，但是该模块只能放在特征提取器和全连接层之间。所以改进的 SE 模块与 S3TA 相比，SE 模块的适用范围更大。

§6.2 研究展望

本文的研究工作仍然有很多不足之处，未来研究人员可以改进这些不足之处来继续提升相关模型的性能。首先，实验所用数据集较小，而且数据波动较大。所以未来研究人员可以寻找一些规模更大的数据集进行训练。其次，因为本文的实验并没有对数据集进行专门的预处理，所以未来研究人员可以在数据清洗方面专门改进。除此之外，因为数据集和输入尺寸较小的问题，实验中所有的模型都无法使用大型的特征提取器。未来研究人员在使用大型数据集时可以加入大型特征提取器进行优化，这种模型可以获得更好的性能。因为数据集太小，本文中改进的 SE 模块并没有改变空间基通道数。未来研究人员在使用大型网络数据集时

可以改变空间基通道数。

综上所述，改进的 S3TA 和 SE 模块与多示例学习注意力机制结合的新模型克服了大量多示例学习的缺陷。这意味着研究人员可以更有信心地使用没有图像块标签的 WSI。从长远来看，如果相关研究工作继续进展，医生将减少大量标记图像块的时间，并且可以做更多有意义的工作。

参考文献

[1] Cheng H D, Shan J, Ju W, *et al*. Automated breast cancer detection and classification using ultrasound images: A survey[J]. Pattern recognition, 2010, 43(1): 299-317

[2] Eklund A, Dufort P, Forsberg D, *et al*. Medical image processing on the GPU–Past, present and future[J]. Medical image analysis, 2013, 17(8): 1073-1094

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105

[4] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778

[5] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258

[6] Cireşan D C, Giusti A, Gambardella L M, *et al*. Mitosis detection in breast cancer histology images with deep neural networks[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2013: 411-418

[7] Cruz-Roa A, Basavanhally A, González F, *et al*. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks[C]. Medical Imaging 2014: Digital Pathology. International Society for Optics and Photonics, 2014, 9041: 904103

[8] Mousavi H S, Monga V, Rao G, *et al*. Automated discrimination of lower and higher grade gliomas based on histopathological image analysis[J]. Journal of pathology informatics, 2015, 6

[9] Xu Y, Jia Z, Ai Y, *et al*. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP). IEEE, 2015: 947-951

[10] Hou L, Samaras D, Kurc T M, *et al*. Patch-based convolutional neural network for whole slide tissue image classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2424-2433

[11] Bejnordi B E, Veta M, Van Diest P J, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer[J]. Jama, 2017, 318(22): 2199-2210

[12] 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

[13] Bandi P, Geessink O, Manson Q, *et al*. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge[J]. IEEE transactions on medical imaging, 2018, 38(2): 550-560

[14] Graham S, Shaban M, Qaiser T, *et al*. Classification of lung cancer histology images using patch-level summary statistics[C]. Medical Imaging 2018: Digital Pathology. International Society for Optics and Photonics, 2018, 10581: 1058119

[15] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning[C]. International Conference on Machine Learning. PMLR, 2018: 2127-2136

[16] Hashimoto N, Fukushima D, Koga R, *et al*. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3852-3861

[17] Mott A, Zoran D, Chrzanowski M, *et al*. Towards interpretable reinforcement learning using attention augmented agents[J]. Advances in Neural Information Processing Systems, 2019, 32

[18] Zoran D, Chrzanowski M, Huang P S, *et al*. Towards robust image classification using sequential attention models[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9483-9492

[19] Roelfsema P R, Lamme V A F, Spekreijse H. Object-based attention in the primary visual cortex of the macaque monkey[J]. Nature, 1998, 395(6700): 376-381

[20] Baldauf D, Desimone R. Neural mechanisms of object-based attention[J]. Science, 2014, 344(6182): 424-427

[21] Van Essen D C, Anderson C H. Information processing strategies and pathways in the primate visual system[J]. An introduction to neural and electronic networks, 1995, 2: 45-76

[22] Liversedge S P, Findlay J M. Saccadic eye movements and cognition[J]. Trends in cognitive sciences, 2000, 4(1): 6-14

[23] Eckstein M P, Koehler K, Welbourne L E, *et al*. Humans, but not deep neural networks, often miss giant targets in scenes[J]. Current Biology, 2017, 27(18): 2827-2832

[24] Ricci-Vitiani L, Lombardi D G, Pilozzi E, *et al*. Identification and expansion of human colon-cancer-initiating cells[J]. Nature, 2007, 445(7123): 111-115

[25] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C]. Advances in Neural Information Processing Systems. 2017: 5998-6008

[26] Parmar N, Vaswani A, Uszkoreit J, *et al*. Image transformer[C]. International Conference on Machine Learning. PMLR, 2018: 4055-4064

[27] Lee M C H, Oktay O, Schuh A, *et al*. Image-and-spatial transformer networks for structure-guided image registration[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 337-345

[28] Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex[J]. Nature neuroscience, 1999, 2(11): 1019-1025

[29] Cadieu C F, Hong H, Yamins D L K, *et al*. Deep neural networks rival the representation of primate IT cortex for core visual object recognition[J]. PLoS computational biology, 2014, 10

[30] Elsayed G, Shankar S, Cheung B, *et al*. Adversarial examples that fool both computer vision

and time-limited humans[J]. Advances in neural information processing systems, 2018, 31

[31] Van Essen D C, Anderson C H. Information processing strategies and pathways in the primate visual system[J]. An introduction to neural and electronic networks, 1995, 2: 45-76

[32] Roelfsema P R, Lamme V A F, Spekreijse H. Object-based attention in the primary visual cortex of the macaque monkey[J]. Nature, 1998, 395(6700): 376-381

[33] Baldauf D, Desimone R. Neural mechanisms of object-based attention[J]. Science, 2014, 344(6182): 424-427

[34] Bower J M. 20 years of computational neuroscience[M]. New York: Springer, 2013

[35] Liversedge S P, Findlay J M. Saccadic eye movements and cognition[J]. Trends in cognitive sciences, 2000, 4(1): 6-14

[36] Eckstein M P, Koehler K, Welbourne L E, *et al*. Humans, but not deep neural networks, often miss giant targets in scenes[J]. Current Biology, 2017, 27(18): 2827-2832

[37] 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

[38] Snead D R J, Tsang Y W, Meskiri A, *et al*. Validation of digital pathology imaging for primary histopathological diagnosis[J]. Histopathology, 2016, 68(7): 1063-1072

[39] Pantanowitz L, Sinard J H, Henricks W H, *et al*. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center[J]. Archives of Pathology and Laboratory Medicine, 2013, 137(12): 1710-1722

[40] Hanna M G, Parwani A, Sirintrapun S J. Whole slide imaging: technology and applications[J]. Advances in Anatomic Pathology, 2020, 27(4): 251-259

[41] BenTaieb A, Hamarneh G. Predicting cancer with a recurrent visual attention model for histopathology images[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 129-137

[42] Tellez D, Litjens G, van der Laak J, *et al*. Neural image compression for gigapixel histopathology image analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 567-578

[43] Schlag I, Arandjelovic O. Ancient Roman coin recognition in the wild using deep learning based recognition of artistically depicted face profiles[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 2898-2906

[44] Cooper J, Arandjelovic O. Visually understanding rather than merely matching ancient coin images[C]. Proceedings of the INNS Conference on Big Data and Deep Learning. Sestri Levante. 2019

[45] Bennett W, Smith K, Jarosz Q, *et al*. Reengineering workflow for curation of DICOM datasets[J]. Journal of digital imaging, 2018, 31(6): 783-791

[46] Kahn C E, Carrino J A, Flynn M J, *et al*. DICOM and radiology: past, present, and future[J]. Journal of the American College of Radiology, 2007, 4(9): 652-657

[47] Herrmann M D, Clunie D A, Fedorov A, *et al*. Implementing the DICOM standard for digital pathology[J]. Journal of pathology informatics, 2018, 9

[48] Clunie D A. Dual-personality DICOM-TIFF for whole slide images: a migration technique for legacy software[J]. Journal of Pathology Informatics, 2019, 10

[49] Turing A M. Computing Machinery and Intelligence[J]. Creative Computing, 1980, 6(1): 44-53

[50] LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[51] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28

[52] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial intelligence, 1997, 89(1-2): 31-71

[53] Tellez D, Litjens G, van der Laak J, *et al*. Neural image compression for gigapixel histopathology image analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 567-578.

[54] Ledley R S, Lusted L B. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason[J]. Science, 1959, 130(3366): 9-21

[55] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]. Proceedings of the 24th International Conference on Machine Learning. 2007: 791-798

[56] Rathore S, Hussain M, Iftikhar M A, *et al*. Ensemble classification of colon biopsy images based on information rich hybrid features[J]. Computers in biology and medicine, 2014, 47: 76-92

[57] LeCun Y, Boser B, Denker J S, *et al*. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551

[58] Oquab M, Bottou L, Laptev I, *et al*. Learning and transferring mid-level image representations using convolutional neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1717-1724

[59] Quellec G, Cazuguel G, Cochener B, *et al*. Multiple-instance learning for medical image and video analysis[J]. IEEE reviews in biomedical engineering, 2017, 10: 213-234

[60] Liu G, Wu J, Zhou Z H. Key instance detection in multi-instance learning[C]. Asian Conference on Machine Learning. PMLR, 2012: 253-268

[61] Cheplygina V, Tax D M J, Loog M. Multiple instance learning with bag dissimilarities[J]. Pattern recognition, 2015, 48(1): 264-275

[62] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning[J]. Advances in neural information processing systems, 2002, 15

[63] Ramon J, De Raedt L. Multi instance neural networks[C]. Proceedings of the ICML-2000

workshop on attribute-value and relational learning. 2000: 53-60

[64] Kandemir M, Hamprecht F A. Computer-aided diagnosis from weak supervision: A benchmarking study[J]. Computerized medical imaging and graphics, 2015, 42: 44-50

[65] Goldman S A, Kwek S S, Scott S D. Agnostic learning of geometric patterns[J]. Journal of Computer and System Sciences, 2001, 62(1): 123-151

致谢

不知不觉三年时光已经过去，在桂电的时光里，我已经从当初的研究小白，到如今大致明白如何做研究。这其中有很多对我帮助很多的人。感谢罗老师和蓝老师给我研究指明方向。当遇到困难时，实验室的小伙伴总是会一起解决问题。这其中的艰难我们都一起承受，我很感谢你们对我的帮助。

虽然很多时候会有迷茫，但是有大家相互交流可以让我敞开心扉面对一切。虽然老师有时会很严格，但是我知道都是为我们将来不在社会上吃亏打预防针。虽然很多时候实验不会成功，但是可以为下一次实验的成功总结经验。虽然有时和同学之间有不愉快，但是最终我们还是朋友。

最后，感谢罗老师和蓝老师的对我的教导，感谢热心帮助过我的同学。

作者在攻读硕士期间的主要研究成果

论文

[1] Multiple-instance CNN Improved by S3TA for Colon Cancer Classification with Unannotated Histopathological Images[C].2021 11th International Conference on Intelligent Control and Information Processing (ICICIP). IEEE, 2021: 444-448, DOI: 10.1109/ICICIP53388.2021.9642206 (第一作者，已录用)

参与项目

[1] 基于深度神经网络的图像描述关键技术研究(桂科 ZY20198016)，中央引导地方科技发展专项，2020.09-2023.08，在研，参与。

[2] 智能医学图像分析(2019GXNSFFA245014)，广西杰出青年科学基金项目，2020.1-2024.01，在研，参与。

参与标准

[1] 可视养生箱，企业标准，Q/450323 ZH 003-2021，参与。

[2] 掌静脉智能锁，企业标准，Q/450323 SM 012-2020，参与。