基于 U-Net 的胃肠道图像分割

2021023206　徐力

2022 年 6 月 12 日

机器学习课程作业

摘要

分析了肠胃道图像分割竞赛数据的特点，包括掩码类别的占比、病例扫描次数分布、不同器官类别的共现性。通过观察掩码的重叠，制定了多标签分类的任务类型。分析了在扫描切片维度的掩码分布变化。修改了 U-Net 以适用于竞赛数据集，设计了不同参数和下采样深度的两种 U-Net 变形结构，对比了两种结构的性能，并选择了结构较简单且泛化性能较优的 4 层 U-Net 结构。通过应用弹性畸变和高斯模糊，充实了训练数据，极大地提高了 U-Net 的泛化性能，改善了过拟合的问题。探究了几种损失函数，探究了其可用性，比较了不同损失函数的优缺点，通过实验确定了目前最佳比例的混合损失函数，并分析了平均准确率较低的可能的原因。提出了仍待解决的问题，给出了未来的改进思路。

关键词: 医学图像分割;U-Net; 弹性畸变; 损失函数

I. 引言

医学图像分割是医学图像处理与分析领域的复杂而关键的步骤，其目的是将医学图像中具有某些特殊含义的部分分割出来，并提取相关特征，为临床诊疗和病理学研究提供可靠的依据，辅助医生作出更为准确的诊断。医学图像具有复杂性，在分割过程中需要解决不均匀及个体差异等一系列问题，所以一般的图像分割方法难以直接应用于医学图像分割[1]。

UW-Madison 胃肠道图像分割 (UW-Madison GI Tract Image Segmentation) 是一场研究型代码竞赛[2]，要求在医学扫描中跟踪健康肠胃器官的准确位置以改善肠胃癌的放射治疗。在治疗当中，借助集成磁共振成像和线性加速器系统（也称为 MR-Linacs）等技术，为了指向肿瘤施加高剂量辐射，同时避开肠和胃，放射肿瘤学家必须手动勾勒出胃和肠的位置。这项工作非常耗时，因为肿瘤、肠和胃的位置每天都在变化，使得治疗时间大大延长。

在比赛中，需要创建模型以在 MRI 扫描中自动分割肠和胃。在给出的数据集中，不同的患者在不同日子里进行了多次 MRI 扫描，每次扫描包括多张断层扫描。

II. U-Net 框架

U-Net[3] 的构建基于“全卷积网络”。它由收缩路径（左侧）和扩展路径（右侧）组成。收缩路径遵循卷积网络的典型架构。它由两个 3x3 卷积（未填充卷积）的重复应用组成，每个卷积后跟一个整流线性单元 (ReLU) 和一个 2x2 最大池化操作，步幅为 2，用于下采样。在每个下采样步骤中，我们将特征通道的数量加倍。扩展路径中的每一步都包括对特征图进行上采样，然后是将特征通道数量减半的 2x2 卷积（“上卷积”），与收缩路径中相应裁剪的特征图的连接，以及两个 3x3 卷积，每个后跟一个 ReLU。由于在每个卷积中都会丢失边界像素，因此收缩路径中的图像需要经过裁剪，再与扩展路径中相应的图像拼接。在最后一层，使用 1x1 卷积将每个 64 分量特征向量映射到所需数量的类。该网络总共有 23 个卷积层，如图 1 所示。

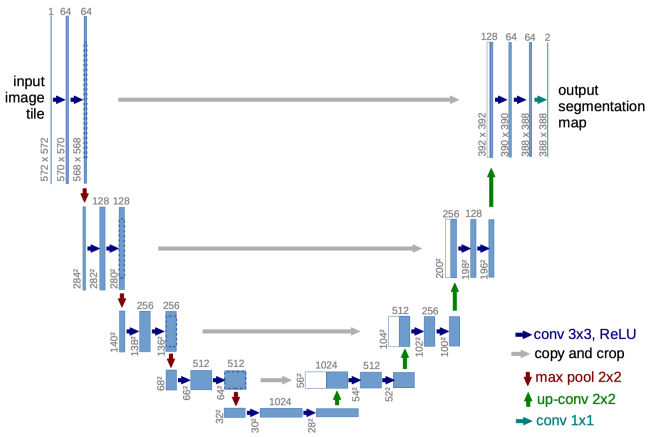

图 1: U-Net 结构

在上采样部分，上采样的特征图包含了上下文信息，并与来自收缩路径的高分辨率特征结合，将上下文信息传播到更高分辨率的层。分割图只包含输入图像中完整上下文可用的像素，为了预测图像边界区域中的像素，通过镜像输入图像来推断缺失的上下文，如图 2 所示。

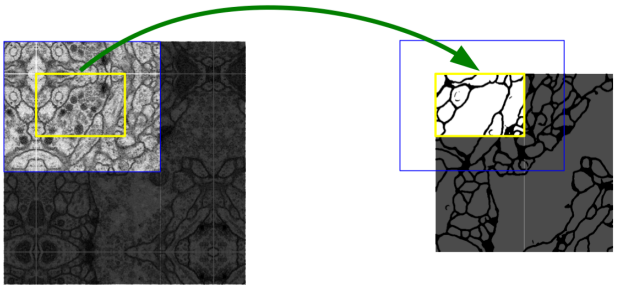

图 2: 镜像拼接原图，以补充边缘像素缺失的上下文

生物医学图像分割中，可用的训练数据通常较少。因此，需要对可用的训练图像应用数据增强，如弹性变形等。这允许网络学习对此类变形的不变性。同时在生物医学当中，变形曾经是组织中最常见的变化，并且可以有效地模拟真实的变形。通过数据增强，模型学习数据的不变性特征，从而防止过拟合。

III. 赛题数据分析

竞赛给出 16 位灰度 PNG 图像，参赛者需要在图像中分割器官细胞。训练注释以 RLE 编码掩码的形式提供。

比赛中有多个病例，每个病例的扫描分为多组，每组由扫描发生的日期标识，扫描的当天产生多张扫描切片。一部分病例按照扫描日期的先后划分为训练集和测试集；另外一部分病例的全部数据处于训练集或测试集中。因此，模型不仅需要针对见过的病例标注其器官位置，也需要标注完全没见过的病例。

训练集中，每个病例的扫描次数（天数）在 1 到 6 之间。

表 1. 数据集的统计数据

注释数量（图片数量）	病例数量	无掩码图片数量	有掩码图片数量
38496	85	21906	16590

表 2. 扫描次数统计

扫描次数	1 次	2 次	3 次	4 次	5 次	6 次	合计
病例数量	9	8	45	2	20	1	85

表 3. 胃、大肠和小肠的共现; 其中对角线上的数字表示该器官的总出现次数

次数	胃	大肠	小肠
胃	8627	6181	3361
大肠	6181	14085	10982
小肠	3361	10982	11201

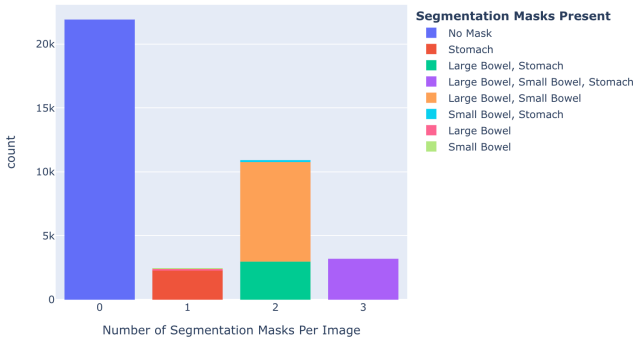

图 3: 每张图片的分割掩码数量

如表 1 所示，数据集中一共有 38496 张图片，其中没有掩码的图片有 21906 张，占 56.9046%；有至少一类掩码的图片有 16590 张，占 43.0954%。这些扫描图片来自一共 85 个病例。

每个病例的扫描次数（天数）在 1 次（天）到 6 次（天）之间，如表 2 所示。其中，最多的病例（45 例）是扫描了 3 次，其次（20 例）是扫描了 5 次。按照比赛规则，部分病例的不同阶段的扫描被分别划分到了训练集和测试集，这可能是扫描次数分布不规律的原因。

进一步统计当图片的掩码分别存在 0、1、2、3 类时对应的各器官种类的数量，如图 3 所示。没有注释的情况可在表 1 中观察，下面分别观察图片存在 1、2、3 种掩码的情况：

A. 存在一种掩码的图片

存在且仅存在一种掩码的图片一共有 2468 张，占 6.41%。在这些注释当中，大多数的注释是关于胃。其中，胃的注释有 2286 个，占 92.6%；大肠的注释有 123 个，占 4.98%；小肠的注释有 59 个，仅占 2.39%。

B. 存在两种掩码的图片

存在且仅存在两种掩码的图片一共有 10921 张，占 28.37%. 在这些注释当中，大多数的注释的组合是关于大肠和小肠。其中，大肠-小肠的注释有 7781 个，占 71.3%；大肠-胃的注释有 2980 个，占 27.3%；小肠-胃的注释有 160 个，仅占 1.47%。

C. 存在三种掩码的图片

存在三种掩码的图片一共有 3201 张，占 8.32%.

如表 3 所示。所有注释一共有 33913 条，其中最多的种类是大肠（14085 条），其次是小肠（11201 条）、胃（8627 条），分别占 41.53%，33.03%，25.44%. 一次共现在这里表示为一对注释在同一张扫描图片上同时出现。胃-大肠、大肠-小肠、小肠-胃的共现分别是 6181、10982、3361。可见，断层扫描图片中，最常同时出现器官的是大肠和小肠，最少同时出现的是小肠和胃。造成的原因可能是身体结构导致的器官分布规律：大肠通常分布在胃的下方，而小肠分布在大肠的下方。

数据集中的图片尺寸主要有 (266,266) 和 (310,360) 两种，此外还有 (276,276) 和 (234,234)，如表 4 所示。

表 4. 不同尺寸的图片数量

(266,266)	(310,360)	(276,276)	(234,234)	合计
25920	11232	1200	144	38496

数据集注释采用游程编码（RLE, run-length encoding）格式。每个像素可以属于背景，或大肠、小肠、胃当中的一类或多类。典型掩码标签的可视化如图 4 所示。部分器官分割注释是有重叠的，如图 5 所示，有一部分大肠的注释完全处于小肠的注释当中。因此本竞赛是逐个像素的多标签多分类任务。

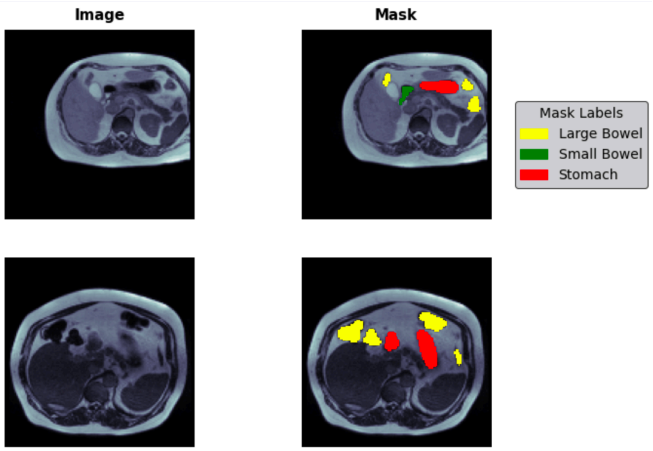

图 4: 典型的掩码标签注释

为了进一步探究将竞赛作为单标签多分类处理会造成的损失，统计不同种类和不同大小的掩码重叠发生的次数，如图 12 所示。胃-大肠-小肠重叠掩码的面积可达到 300 以上；大肠-小肠重叠掩码的面积可达到 600 以上。因此，掩码应表示为 $W \times H \times 3$。其中 3 个通道分别表示不同的掩码类型。

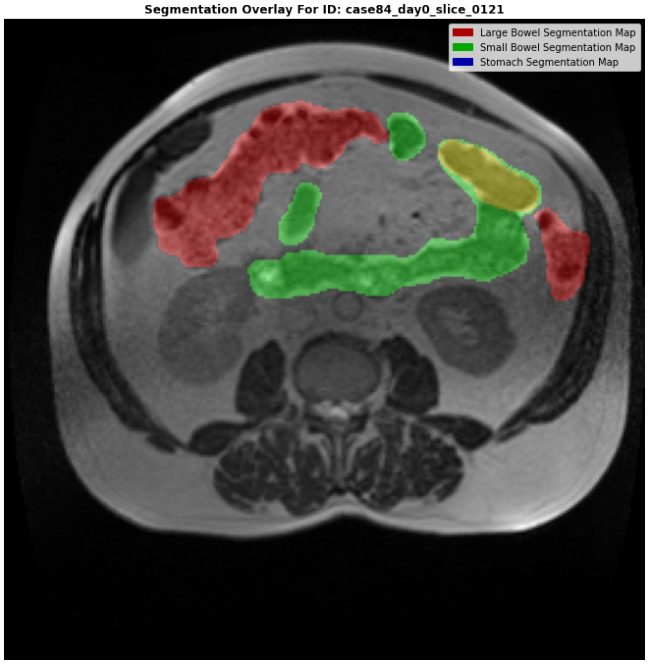

图 5: 重叠的掩码标签注释

对于给定的病例和扫描天数为单次扫描。单次扫描得到一系列连续的断层切片，得到的切片数量有两种情况：144 张或 80 张。一共 259 次扫描是 144 张，15 次扫描是 80 张。统计所有不同位置的切片包含器官掩码的数量，分布近似钟形曲线，如图 6 所示。观察得到，胃、大肠、小肠在扫描切片上分布达到峰值所在的位置分别为 67、100、102，对应的最大注释数量分别为 196、240、235 个。大肠和小肠对峰值以左的切片位置有所偏置，而胃有向右的偏置。小肠、胃的扫描切片位置相对大肠较为集中。可见，器官扫描切片数量峰值的位置排序符合器官的大致高低位置排序（胃 > 大肠 > 小肠）；而峰值的注释数量排序符合各器官总注释数量的排序（大肠 > 小肠 > 胃）。

IV. 实验

A. 评价指标

竞赛根据平均骰子系数 (Dice coefficient) 和 3D 豪斯多夫距离 (Hausdorff distance) 评估结果。

平均骰子系数是每张图片的 Dice 系数的平均值：

$$SC_{Dice}$$

3D 豪斯多夫距离 (Hausdorff distance) 的计算：首先使用扫描切片构建 3D 图像，每个切片的深度为 1，计算预测物体到基准物体的豪斯多夫距离并归一化，取每

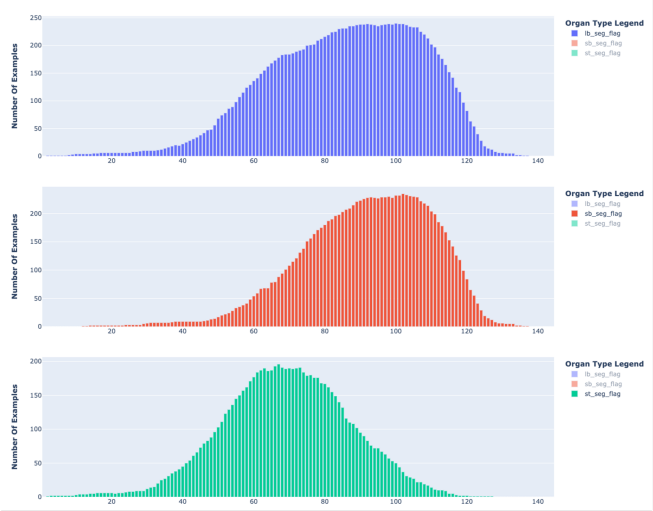

图 6: 分割掩码重叠面积的分布

张图片的平均，最后转化为分数：

$$SC_{Haus}$$

最终分数是两个分数的加权和，计算如下：

$$SC_{Final} = 0.4 * SC_{Dice} + 0.6 * SC_{Haus}$$

1) 骰子系数: 设 X 是预测像素的集合，Y 是基准像素的集合，则 Dice 系数计算如下：

$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$

Dice 系数用于度量两个集合之间的一致性。当 X 和 Y 都为空时，Dice 系数定义为 0。

2) 豪斯多夫距离: 竞赛中的豪斯多夫距离指的是有向豪斯多夫距离，即一个点集中的点到另一个点集中的点的最短距离的最大值。设 X 是预测物体像素的集合，Y 是基准物体像素的集合，则 3D 豪斯多夫距离计算如下：

$$\sup_{x \in X} \inf_{y \in Y} d(x, y)$$

其中 sup 和 inf 分别表示上确界和下确界。

B. 实验设置

1) 训练集划分: 实验采用 5 折交叉验证，需要将样本随机分成数量大致相等的五个数据集。划分的过程还需要满足一些要求：第一，每个病例作为一个组，且每个组中的所有数据必须处于同一折当中。否则，若一个病例的部分数据处于训练集且另一部分处于验证集，则会出现数据泄漏的问题。第二，每折数据应包含所有的分类，即背景、胃、大肠和小肠，且每折数据包含的各类别应尽量保持等比例，以防止数据不平衡的问题。

使用 keras 中的 StratifiedGroupKFold 方法，将图片中分别有 0、1、2、3 个类别作为标签，并按病例分组，即可划分 5 折交叉验证数据集并满足以上两点要求。

2) 模型结构: 对 U-Net 进行修改以适应数据集：第一，U-Net 原输入张量为 $572 \times 572 \times 1$，为了适应肠胃的多标签分类，输入张量修改为 $128 \times 128 \times 3$，每个通道表示一个类别的掩码。

第二，每个 3×3 卷积包含 padding 操作，不改变图像尺寸，从而保持了收缩路径和扩展路径尺寸的一一对应，在拼接操作时，高分辨率图像不需要裁剪，从而保留了高分辨率图像的边缘信息。在 unet++ 等文献中，同样是这种做法。

第三，比较了两种结构：收缩路径下采样深度为 4 或 5，最深层神经元的感受野分别为 125 和 253，而图像的输入大小为 128×128，略大于 4 层 unet 的最大下采样感受野，小于 5 层 unet 的最大下采样感受野，如表 5 所示。

表 5. 4 层和 5 层 unet 的每层下采样前的理论感受野

4 层下采样			**5 层下采样**		
层	输入大小	感受野	层	输入大小	感受野
0	128	5×5	0	128	5×5
1	64	13×13	1	64	13×13
2	32	29×29	2	32	29×29
3	16	61×61	3	16	61×61
4	8	125×125	4	8	125×125
/	/	/	5	4	253×253

3) 训练设置: 以二元交叉熵损失函数作为优化目标，使用 Adam 算法作为优化器。度量函数为骰子系数、交并比系数和准确率。实验在 i5-8400 CPU 和 RTX3080 GPU 的 Linux 服务器上进行。

C. 初步实验结果分析

如图 7 所示，初步的实验取得了 50% 以上的验证集上 Dice 系数。由于设置了早停，在连续 5 轮的验证集损失没有低于第 6 轮的损失之后，训练停止。由此可见，过拟合是亟待解决的问题。

模型预测的可视化如图 9 所示。可以观察到，对于已经判断正确的器官，模型预测的边界倾向于收缩，即将一些器官的边缘像素分类为背景。同时，模型可能将背景误分类为器官。

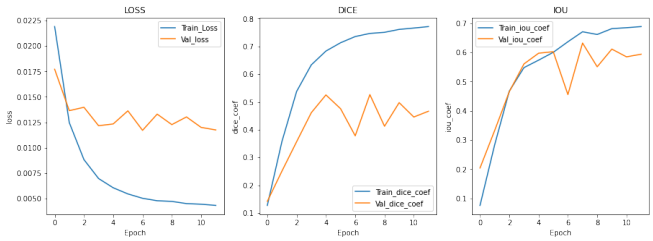

图 7: 损失函数和度量函数的优化过程

D. 数据增强

为了解决模型的过拟合的问题，对训练集进行数据增强。弹性畸变[5] 是一种适用于生物细胞组织的数据增强方法，这有利于模型学习在这种畸变当中的不变性，而畸变是器官组织很常见的变化。同时加入高斯噪声从而突出图像的低频特征。数据增强的效果如图 8 所示。

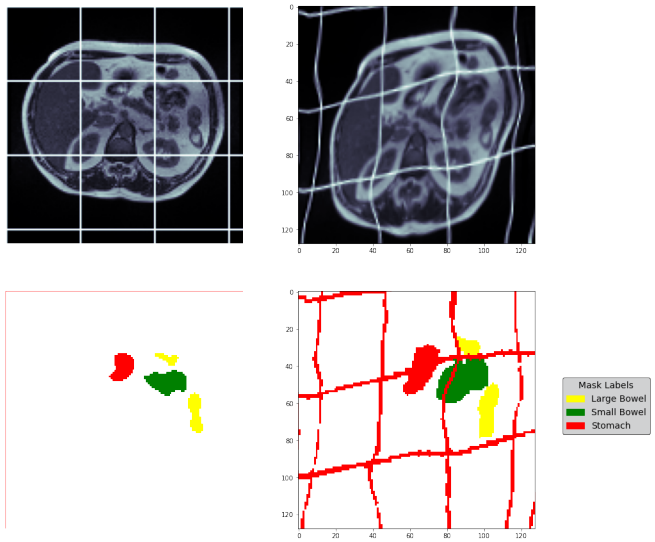

图 8: 数据增强效果；左上：原图（加入栅格以更明显地查看畸变程度）；右上：加入数据增强的图像；左下：大肠、小肠和胃的掩码图像；右下：加入数据增强的掩码图像

数据增强后的损失函数、Dice 系数和 IOU 系数的优化过程如图 9 所示。验证集上损失在第 19 轮达到极小值，并直到第 34 轮没有明显的降低或提升，因此在 34 轮进入早停。可见，在 20 轮训练之前没有出现过明显的拟合问题。此外，与未经过数据增强相比较，最优 Dice 系数也从 0.5 提升到 0.7006，最优交并比则从 0.6 提升到 0.7699. 并且，训练集的交并比也有所提升。

E. 4 层 unet 和 5 层 unet 的比较

为了探究增加下采样层数是否会提升模型的性能，分别构建了 4 层 unet 和 5 层 unet，采取相同的数据增强

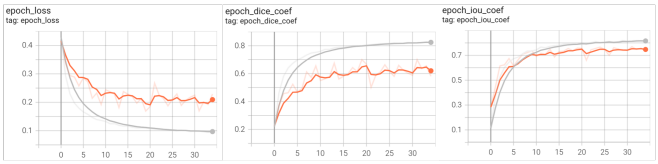

图 9: 数据增强后的损失函数、Dice、IOU 优化过程（曲线有 60% 的平滑）

方法和学习率等超参数，并对比它们的性能。4 层 unet 和 5 层 unet 的参数数量分别为 2.50 兆和 5.72 兆。训练结果如图 10 所示。

可见，4 层和 5 层 unet 的 loss、Dice 系数、IOU 无明显差异，验证集上最优 Dice 系数均处在 0.7 附近，验证集上最优交并比均在 0.75 附近。然而，在 15 轮等待的早停中，4 层 unet 的最优损失出现在第 27 轮且训练了 39 轮（最大训练轮数为 40 轮）；而 5 层 unet 的最优损失出现在第 19 轮且训练了 34 轮。因此可以认为，参数更少的 4 层 unet 的泛化性能优于参数更多的 5 层 unet，而 5 层 unet 并没有学习到比 4 层 unet 更多的特征。

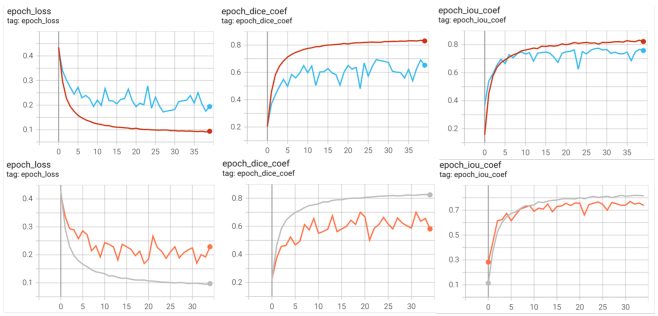

图 10: 4 层 unet 和 5 层 unet 的损失函数、Dice、IOU 优化过程比较；上方：4 层 unet；下方：5 层 unet

F. 不同损失函数的比较

比较不同的目标损失函数，并对比它们的性能。实验中比较的损失函数有：

1) 二元交叉熵损失函数: 逐像素的二元交叉熵损失函数有着梯度平滑的优点，但是它的优化易受类别不平衡的影响。定义为：

$$\mathcal{L}_{BCE}(Y,P) = -\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N}(y_{n,c}\log p_{n,c}+(1-y_{n,c})\log(1-p_{n,c}))$$

2) Dice 损失函数: Dice 损失函数仅考虑真实掩码和预测掩码的交面积和总面积之比，避免了背景像素占绝大部分比例而导致的类别不平衡问题。但是它有着梯

度不平滑的缺点。

$$\mathcal{L}_{Dice}(Y,P)=\sum_{c=1}^{C}(\frac{2Y_cP_c}{sum(Y_c)+sum(P_c)})$$

3) soft-dice 损失函数: soft-dice 损失函数[6] 考虑在真实类别像素上，对该像素的预测接近 1 的程度。soft-dice 损失函数失去了 dice 系数的几何意义，它同样能解决类别不平衡问题。

$$\mathcal{L}_{SD}(Y,P)=-\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N}(\frac{2y_{n,c}p_{n,c}}{y_{n,c}^2+p_{n,c}^2})$$

4) 二元交叉熵损失函数 +Dice 损失函数: 合并二元交叉熵损失函数和 Dice 损失函数，得到混合损失，它既有梯度平滑的优点，也能解决类别不平衡问题。

$$\mathcal{L}_{BCEDice}(Y,P)=\mathcal{L}_{BCE}(Y,P)+\mathcal{L}_{Dice}(Y,P)$$

5) 二元交叉熵损失函数 +Dice 损失函数/2: 实验结果表明，二元交叉熵损失函数和 0.5 的 Dice 损失函数的混合是较优的组合。

$$\mathcal{L}_{BCEDice_{2:1}}(Y,P)=\mathcal{L}_{BCE}(Y,P)+0.5\times\mathcal{L}_{Dice}(Y,P)$$

表 6 给出了各损失函数比较的结果。首先，DICE 损失函数无法优化，在训练过程中该损失函数不能下降。单独将 soft-dice 作为损失函数，性能最差，但是平均准确率的结果很好，可能是因为 soft-dice 损失函数目标隐含了提高平均准确率的目标。单独使用二元交叉熵损失函数的性能不如混合损失函数，原因可能是遭到了类别不平衡的影响。最后，2:1 的 BCE-DICE 损失函数略微优于 1:1 的 BCE-DICE 损失函数。然而，混合损失函数的评价准确率在训练过程中均不稳定，这可能是损失函数和平均准确率不相关导致的。

V. 结论

分析了肠胃道图像分割竞赛数据的特点，包括掩码类别的占比、病例扫描次数分布、不同器官类别的共现性。通过观察掩码的重叠，制定了多标签分类的任务类型。分析了在扫描切片维度的掩码分布变化。修改了 U-Net 以适用于竞赛数据集，设计了不同参数和下采样深度的两种 U-Net 变形结构，对比了两种结构的性能，并选择了结构较简单且泛化性能较优的 4 层 U-Net 结构。通过应用弹性畸变和高斯模糊，充实了训练数据，极大地提高了 U-Net 的泛化性能，改善了过拟合的问题。探究了几种损失函数，探究了其可用性，比较了不同损失函数的优缺点，通过实验确定了目前最佳比例的混合损失函数，并分析了平均准确率较低的可能的原因。

表 6. 各损失函数的结果比较；其中轮数表示损失函数极小值所在的最优轮数；DICE、IOU、ACC 表示优化过程中的最优值，其下方括号表示该指标在最优轮的数值

损失函数	轮数	DICE	IOU	ACC
BCE	25	0.6786 (0.4241)	0.7305 (0.3747)	0.6786 (0.4241)
$DICE$	/	/	/	/
SD	37	0.448 (0.4458)	0.5678 (0.5309)	0.9683 (0.9548)
$BCEDICE$	36	0.6957 (0.6957)	0.7665 (0.7631)	0.8628 (0.1534)
$BCEDICE_{2:1}$	19	0.7006 (0.7006)	0.7699 (0.7576)	0.9287 (0.6449)

目前，仍有以下问题需要解决，并提出相应的改进思路：

第一，在 Dice 系数和 IOU 系数提升的同时，平均准确率变得十分不稳定，且在 loss 最优时准确率较低，在 loss 非最优时准确率却能达到极大值。说明损失函数和准确率没有相关性。同时，soft-dice 损失函数却有着与准确率一致的特点。目前对于背后原因的理论分析不足，对于损失函数或需要更多的改进和取舍。

第二，U-Net 应用于胃肠道图像分割竞赛，仅仅将本具有连续性的断层扫描的图片随机打乱进行训练和预测，遗失了大量纵坐标上的连续信息。例如，图 11 中将一些背景像素预测成了器官，但是，如果在与该层相邻的图片上，模型能正确识别背景，则在连续的切片上可以做信息的互相补充，从而纠正错误并正确识别。同时，器官属于三维实心物体，若模型将连续的扫描图片作为整体判断，则可以减少将中间像素预测为背景像素的情况。

第三，通过观察模型预测结果，发现除错分类以外，模型还经常漏分类，以及对于正确的分类，预测的掩码的边界相对正确的掩码边界有所收缩，即预测过于保守。然而，对于原始数据的观察发现，有时候模型预测的掩码的收缩反而是正确的，因为专家标注有时过于随意，标注的掩码超出了真实掩码的边界许多。

未来的工作将：关注将切片扫描作为 3D 物体切割问题考虑，研究相邻切片间的信息补充从而提高预测精度的方法；改进损失函数，使模型更多关注平均准确率；最后，找改进模型，解决预测边界过于保守的问题，或进行数据清洗，解决专家标注过于激进的问题。

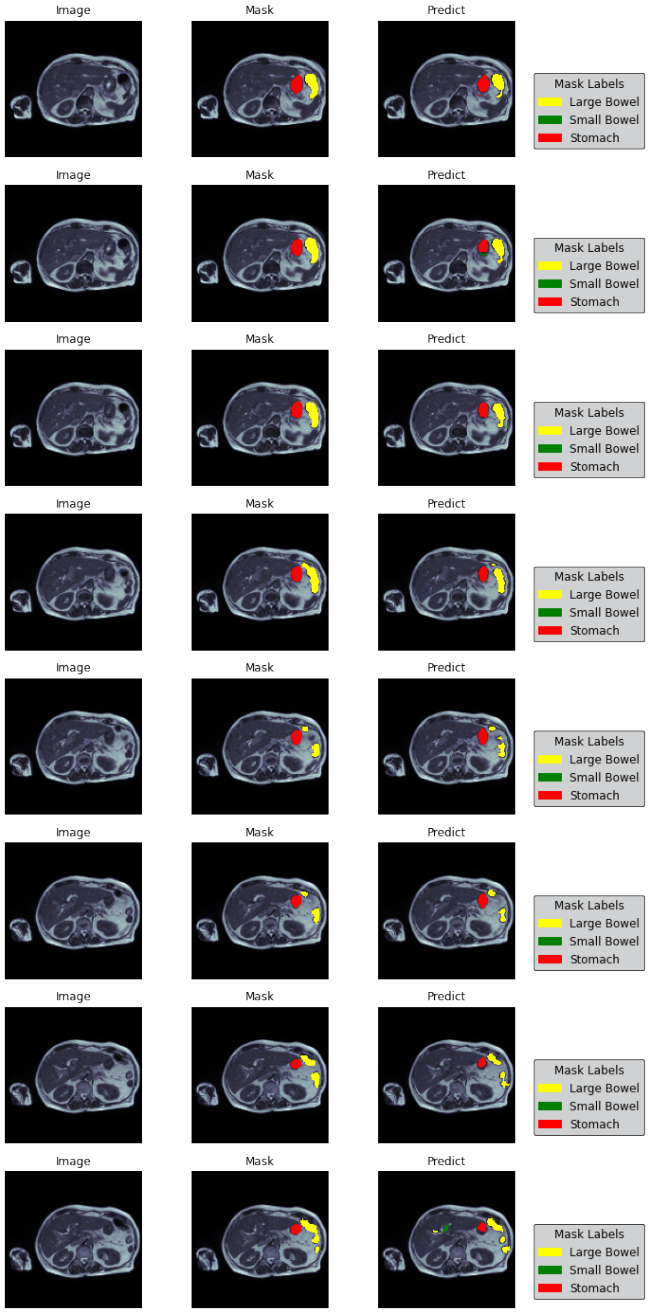

图 11: 原图、真实掩码和预测掩码的可视化

参考文献

[1] 田娟秀, et al. 医学图像分析深度学习方法研究与挑战. 自动化学报, 2018, 44.3: 401-424.

[2] https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation

[3] RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015. p. 234-241.

[4] LIN, Tsung-Yi, et al. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2980-2988.

[5] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., 2003, pp. 958-963, doi: 10.1109/ICDAR.2003.1227801.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," in IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856-1867, June 2020, doi: 10.1109/TMI.2019.2959609.

图 12: 分割掩码重叠面积的分布