项目技术指标与测试报告

一 概述

1.1 项目模型概述

本次项目构建的是，一个可针对口腔根管治疗后CT图像或X光图像进行根管牙自动定位与评分的软件系统，整个系统包括检测和评分两个主要模型：检测模型与评分模型。jpg或jpng格式的口腔CT图像或x光图像进入系统后，首先经过检测模型，得到以矩形框框定的根管牙，被框定的部分将进一步被切取并进入评分模型，评分模型将根据根管填充情况对其进行评分，系统最终输出框定出根管牙并给出评分后的jpg或jpng格式图像。

1.2 主要需测试模块

1.2.1 检测模型单独功能与性能测试；

1.2.2 评分模型单独功能与性能测试；

1.2.3 集成系统的功能与性能测试；

二 项目标准与技术指标

2.1 项目预期应用

集成后系统通过检测与验证后，将应用与医院或医疗机构口腔科，输入根管治疗后口腔CT图像或x光图像，以系统输出的评分结果来评价主治医生本次治疗的效果及医生水平，节省人力，便于医院进行人才管理。

2.2 应用标准要求

2.2.1 检测标准：

输出检测框应符合根管标注准则，以牙为单位进行标注，每个根管框准确包含同一颗牙的所有根管，且检测框不允许过大而包含过多除根管及根管牙之外的其余信息；

2.2.2 评分标准：

本次项目使用的数据均由尚善口腔提供，评分标准由尚善口腔专家团队拟定，因此输出的评分结果应符合该评分标准，最终系统评分与专家团队评分之间的差别表征了系统的误差。

2.3 本项目技术指标

2.3.1 精确度

2.3.2 速度

2.3.3 占用内存（计算机本身内存和显存）

三 项目测试方案

3.1 检测模型测试

3.1.1 测试目标

使用验证集val\_txt中的图像对模型进行测试，本次使用的验证集共318图像。检测模型测试包括两个方面，第一个方面是确定一个评估标准，对预测结果的精确度进行测试，另一方面是评估超参数对网络的影响，从而筛选出最优的超参数。本模型相关超参数主要有四个:

nms控制预测框的密集程度。网络会尽量地去预测标注框，可能导致同一个目标有多个框，如果全部呈现出来会使得视觉效果较差，因此控制这个参数以删除过于密集的框。nms=1表示不允许标注框之间有重叠，这可能导致相临近的两个目标只有一个能被预测出，nms=0表示呈现网络预测出的所有标注框，使得肉眼看上去非常密集。在[0,1]之间进行测试，找到最适合的nms;

score\_threshold是网络确信度阈值，每个框被预测出来时，网络对于这个框的预测结果都会有一个确信程度，控制这个阈值，将确信程度低于该阈值的框全部删除，可提高预测结果的准确率；

detection\_scale是控制预测框大小的变量，由于训练时对图像进行的变换导致标注框略微比真实框大，网络将学习到这一特性并表现为最终预测框偏大，控制这一变量可使其略微缩小；

iou\_threshold：iou表示网络标注框与人工标注框的交并比，将预测框与人工标注框对比，框重叠率越高表示预测效果越好，重叠程度过低则将这个预测框记为错误预测框。iou\_threshold作为一个阈值决定我们本次测试将哪些预测框作为真实预测框，哪些作为错误预测框。

调整这四个参数，选出最佳值。

最佳参数确定后，使用这一组参数来对模型的精确度进行测试。检测结果的精确度由average precision(ap)来表征，这一参数的含义及计算方法将在后面给出。测试过程中记录处理每一张图片的时长，即为模型的运行速度，同时查看并记录模型占用的内存和显存。

3.1.2 测试步骤

①首先进行最佳超参数的筛选。四个参数中，控制其中三个不变，另外一个参数即为待优化参数，使待优化参数在某一个范围内变化，每取一个值都能获得一组数据的检测结果，这些标注框中，实际应该被预测出，并且预网络测结果也将其预测出的则为true\_positive，网络预测将其标出但实际不应该被预测出的则为false\_positive，这里决定一个框是否该被网络预测出，就需用到前面的iou\_threshold参数，预测框与人工标注框大于此参数规定值时，表示应该被预测出，反之则表示不该被预测出。定义一个recall变量，recall=true\_positive/应该被预测出的总量，用于表示在全部应该被预测出的标注框中，成功被预测出比例，recall越大，网络更不会错过有嫌疑的对象；定义另一个变量precision，precision=true\_positive/(true\_positive+false\_positive)，表示所有被预测出的框中，预测正确的比例，precision越大，网络预测准确度越高。前面提到的ap即是recall和precision通过一个函数综合计算出的结果。这三个参数将随着待优化参数的变化而变化，我们将这一变化表现为折线图，选择它们处于最优值时对应的参数为本模型的最优参数。

②选定最优参数后，测试得到的ap值反映了模型的精确度，ap越大，精确度越高。

3.1.3 测试记录

①测试人员

②测试时间

③测试环境

④测试工具

⑤测试结果

1. 参数nms最优值选择：将nms选为待优化参数，在[0,1]之间每隔0.05取一个值进行测试，得到recall、precision、ap值随nms值变化折线图，根据结果不断缩小最优值区间，最终选出综合最优值。nms在（0,1）之间变化得到折线图如下图3.1.1，多次实验证明当nms取值在[0.65,0.7]之间时，这三个值可综合达到最优，模型最终选取nms=0.7;

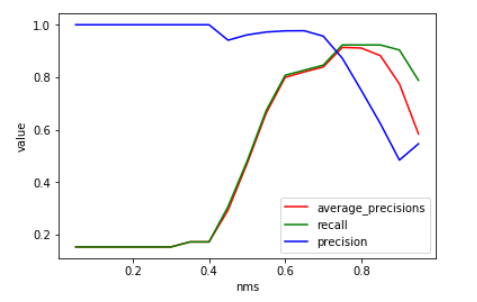


图 3.1 .1 nms取值区间(0,1)

1. 参数score\_threshold最优值选择：将score\_threshold选为待优化参数，在[0,1]之间每隔0.05取一个值进行测试，得到recall、precision、ap值随score\_threshold值变化折线图，根据结果不断缩小最优值区间，最终选出综合最优值。score\_threshold在(0,1)之间变化得到折线图如下图3.1.2。多次实验证明当score\_threshold取值在[0.5,0.6]之间时，这三个值可综合达到最优，模型最终选取score\_threshold=0.5;

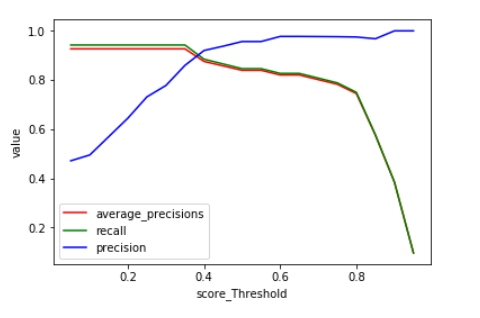


图3.1.2 score\_threshold取值区间为(0,1)

1. 参数iou\_threshold最优值选择：将iou\_threshold选为待优化参数，在[0,1]之间每隔0.1取一个值进行测试，得到recall、precision、ap值随iou\_threshold值变化折线图，根据结果不断缩小最优值区间，最终选出综合最优值。iou\_threshold在(0,1)之间变化得到折线图如下图3.1.3。多次实验证明当iou\_threshold取值在[0.2,0.3]之间时，这三个值可综合达到最优，模型最终选取iou\_threshold=0.3;但整体来看，iou\_threshold值的选择在一定范围内对折线图变化没有太大影响。

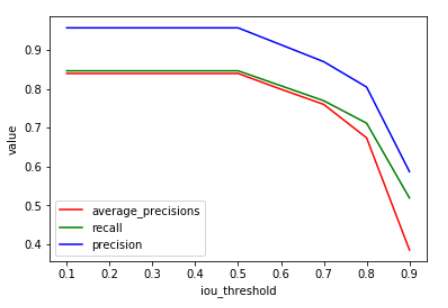


图3.1.3 iou\_threshold取值区间为(0,1)

1. 参数detection\_scale最优值选择：将detection\_scale选为待优化参数，在[0,1]之间每隔0.05取一个值进行测试，得到recall、precision、ap值随detection\_scale值变化折线图，根据结果不断缩小最优值区间，最终选出综合最优值。detection\_scale取值在[0,1]之间时折线图如下图3.1.4，多次实验证明detection\_scale取值在[0.7,1]之间时，这三个值可综合达到最优，综合来看detection\_scale取值对折线图变化影响不大，模型最终选取detection\_scale=1，相当于不缩小检测框。

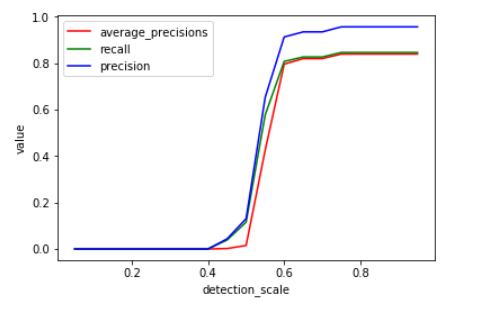


图3.1.4 detection\_scale取值范围为(0,1)

1. 最优参数选定后，利用该组参数处理测试集数据，本次测试选用318张图像中的49张，测试过程中得到precision和recall随测试过程变化，得到折线图如下3.1.5，从折线图可见其变化趋势符合上述计算公式，

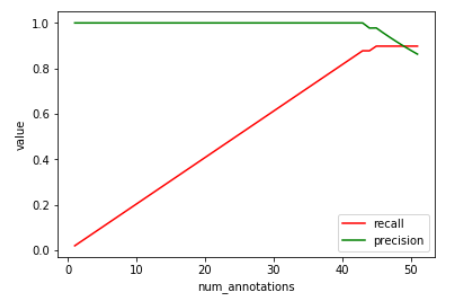


图3.1.5

1. 精确度测试：测试49张图像，得到最终的recall和precision值，利用\_compute\_ap函数计算出average\_precision，此函数的计算原理：......。得到最终ap值如图3.1.6，其中ap以字典形式呈现，0表示根管检测，1表示全牙检测，结果表明此检测模型根管检测精确度为0.8393，全牙检测精确度为0.8975。

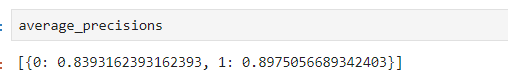


图3.1.6 检测模型精确度

1. 速度测试：记录测试49张图像的开始和结束时间，将结束时间减去开始时间并除以图像数量，得到处理每一张图像的平均时间，结果如下图3.1.7所示，每张图像处理时间约0.42s。

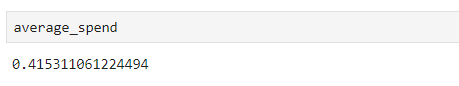


图3.1.7 检测模型速度

1. 占用内存：查看内存使用情况，模型本身占用内存140M，运行时将其加载入显存，再载入图像，查看发现模型和每一张图像及其中间处理结果共占用显存873M，占用内存3.2%/64G，约2.05G。

3.1.4 测试结果评价

从上面的测试结果可知，根管检测的平均精确度约为84%，大部分模型预测结果与人工标注结果相符合，但由于人工标注由不同人完成，标注结果略有不同，而且有部分根管漏标等情况，因此检测模型测试还出现了部分特殊结果。现在将测试结果可视化，将模型标注与人工标注的根管部分在同一张图像上体现出来（这里为了使视觉效果更好只显示根管部分标注框，不显示全牙标注框），人工标注用蓝色框表示，模型标注用绿色框表示。

下图3.1.8为人工标注与系统预测结果均较好且重合率较高的图像举例：

下图3.1.9为人工正确标注，但模型预测时误将非根管标注出（第一个单独绿色框），即增加了false\_positive，同时应该被预测出的框没有被预测出（最后一个单独蓝色框），减少了true\_positive。这两种情况分别会降低precision和recall，导致ap下降。

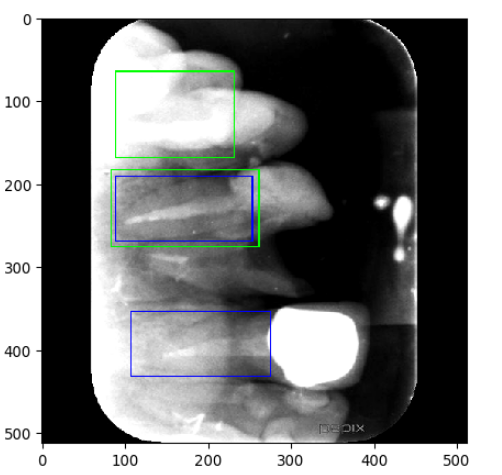


图3.1.9

下图3.1.10为人工标注漏标，但模型将其正确预测出的特殊情况，这种情况实际上是模型正确而人工标注错误，但在精确度计算时以人工标注为标准，将这种情况视作增加了false\_positive，导致precision和recall均下降，最终引起ap下降：

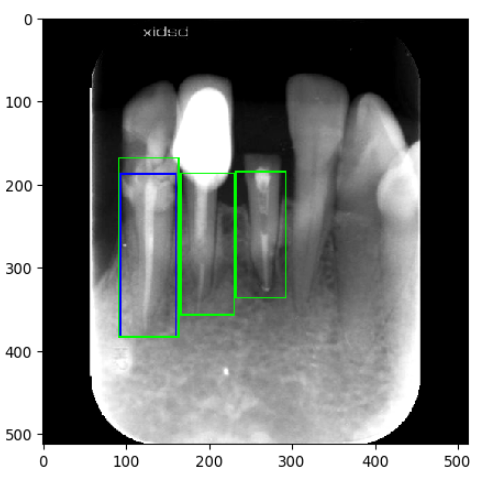


图3.1.10

综合上述情况来看，由于人工标注的失误存在，实际上检测模型的精确度应该会高于84%；

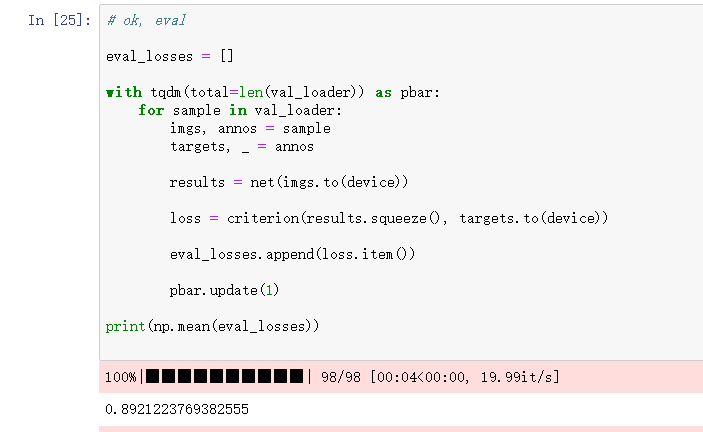
3.2 评分模型测试

3.2.1 测试目标

评分模型训练结束后，利用98张（？）测试集图像对其进行测试，主要计算模型预测分数与专家评分之间的绝对误差的平均值loss，这个值越小，说明模型预测越准确。

3.2.2 测试步骤

将检测模型输出的根管部分检测框切取出后，进入评分模型，输出结果是进行归一化处理后的评分模型（评分模型归一化处理的方法）。首先将测试集中所有图像的专家评分结果的均值mean和标准差std求出，利用(score-mean)/std公式将其归一化，利用L1Loss损失函数，求出预测评分与专家评分之间的绝对误差平均值，这一平均值大小表征了评分模型的精确度。



3.2.3 测试记录

①测试人员

②测试时间

③测试环境

④测试工具

⑤测试结果

3.2.4 测试结果评价

四 总结（功能及性能总结，缺点、不足及预期改进方法*（训练集可根据上下牙、切牙、乳牙等分类更加详细）*等）

1. 人工标注存在误差和遗漏，检测模型精确度有偏差；
2. 尚善口腔评分标注包含使用新技术、新工具加分，这一部分在评分模型中无法体现，模型整体评分可能偏低；
3. 数据较少，分类不够详细，还有改进空间。

测试问题：

1. 内存和显存占用情况查看（截图、每一张图像载入后占用的内存2.05G？）
2. 最终ap值测试、最优参数确定，使用49张图像和使用317张图像结果差异较大，317张图像预测比例和精确度太低？

