Readme 补充材料

LLLLC + 第 4 名

算法描述

1 数据

1.1 GtBoxBasedCrop

数据集构造：将kfb格式的文件转化为“.npz”存储，加速读取。

Crop：训练时，在随机抽取的Roi中，随机选择一个gt_box作为定位基准，并以选定的gt_box为参考做一次RandomCrop，裁剪的方式是保证该框完全位于此次范围中，其他框根据裁剪结果计算overlap，确定是否保留。随机crop可以增加对阳性图片的背景的利用。

Patch：Patch大小根据任务修改，pos和Trichomonas的ground true bbox相对比较小，Candida存在较大的ground true bbox (>1000)，所以对前者，我们采用1000, 1200, 1600等尺寸的patch，而对于后者，我们采用2000, 3000, 4000等尺度的patch，确保截取的patch有最后的视野范围。

1.2 ReplaceBackground

为更好抑制假阳，在训练时引入阴性样本数据，具体实现为：训练时，以一定的概率丢弃当前随机抽取的阳性Roi，从阴性样本数据中随机抽取一张阴性Roi作为背景，把当前的阳性样本中的gt_box贴到阴性背景中作为训练样本，用cv2.inpaint对贴合的gt_box边缘进行修复。

1.3 CopyPaste

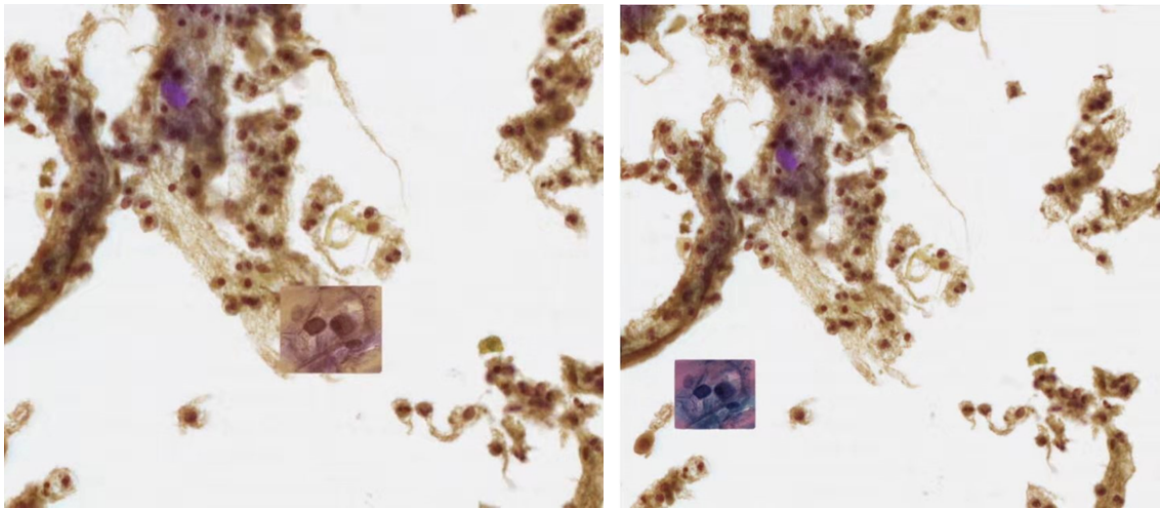
步骤：

1.将所有Roi 的gt_box裁剪后保存下来，建立列表索引。

2.训练时以一定的概率，根据当前随机抽取的阳性Roi中gt_box各个类别的比例，从步骤1中建立的列表索引，随机抽取相同数目的阳性gt_box。

3.用染色剂归一化算法Vahadane将步骤2中随机抽取到的gt_box与步骤一抽取到的Roi进行风格归一化，减少突兀。

4.更新gt_labels



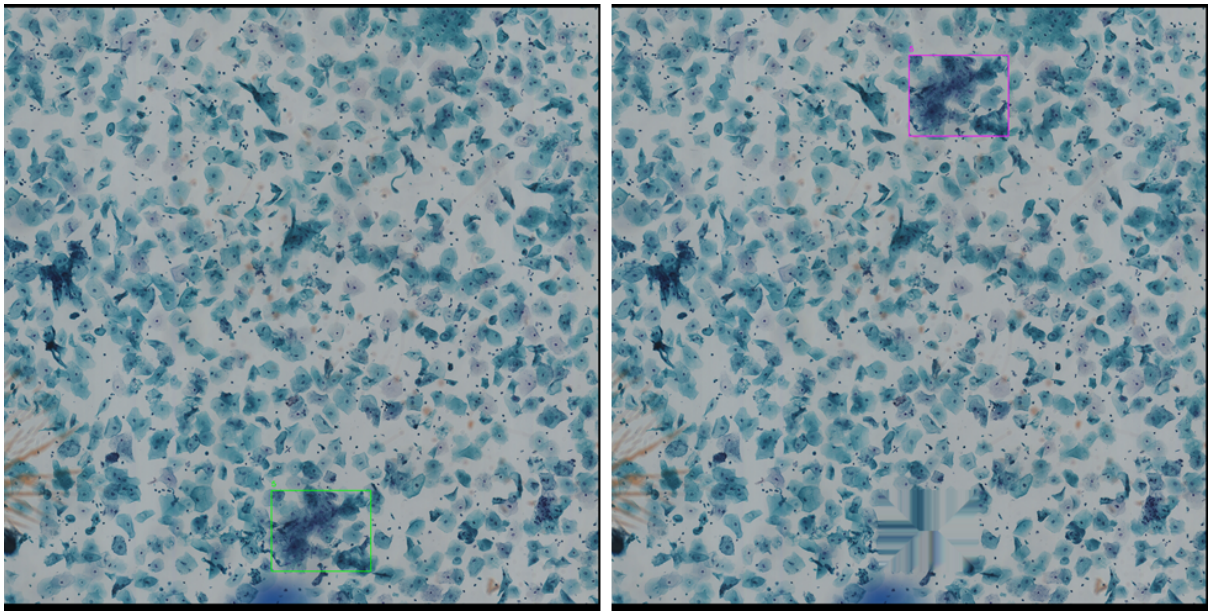
左图为作了Vahadane归一化的CopyPaste效果，右图为不作任何处理直接贴

1.4 Bboxes_Jitter

以一定的概率p对gt_box框作(-0.1,0,1)的缩放

1.5 RandomShiftGtBBox

为了增加gt_box背景多样性，减少模型过拟合风险，以一定的概率p对gt_box做随机平移，用cv2.inpaint对gt_box原始位置及新位置处的外框做修复。



1.6 其它常规的数据增强方法

包括：

1. RandomFlip
2. RandomVerticalFlip
3. Rotate

2 模型和训练

2.1 模型结构

我们主要使用了res50作为backbone的two-stage网络结构作为基础结构，主要使用到了下面的结构：

1. FPN / BiFPN
2. DCN
3. Cascade

2.2 训练过程

训练过程主要有下面的一些改进点：

1. 对类别不平衡的问题的处理。训练集中总共有3670个roi，其中pos和Candida类占据的roi数量相对比较多，但是每个roi中的ground true bbox数量较少，而Trichomonas则相反，其占据比较少的roi，但是每个roi中ground true bbox的数量比较多。我们发现Batch size的大小对训练的效果有影响，当batch size比较大的时候，模型的更新会趋向与Trichomonas，需要使用较小的batch size；
2. 对不同的类别使用针对性的专家模型。因为不同的类别的ground true bbox的尺寸差别大，我们使用不同的模型，对不同的类别进行针对性的优化。这个过程中需要注意利用其它类别的数据，否则会因为数据问题造成过多的假阳性预测。
3. 关于anchor尺寸和梯度贡献的处理。前面提到，不同类别的ground true bbox的尺寸差别大，所以可以使用不同scale的anchor来cover ground true bbox，这个过程中需要注意，当achor的scale发生变化的时候，会导致bbox_head的属于各个类别的proposal数量有所变化，影响梯度更新，并进一步影响模型的倾向性。

3 策略

3.1 后处理策略

医学上，医生宫颈癌细胞看病理切片时，一般先判断有无念珠菌(Candida)，念珠菌一般比较大且特征明显。因为念珠菌和滴虫(Trichomonas)适合生存的PH值不同，二者一般不共存，因此如果存在Candida则可排除滴虫存在。此外，滴虫和念珠菌一般也不与阳性类别“ASC-H”、“ASC-US”、“HSIL”、“LSIL”共存，如果明确存在滴虫和念珠菌，则有较大概率可以排除阳性细胞存在的可能。

根据上述医学上的先验知识，比赛时根据Candida的预测结果对网络的输出作后处理过滤。由于单个模型的置信度可靠性较低，实际后处理时根据模型融合结果，对每个Roi中Candida预测置信度最高的3个gt_box求平均值，如果超过阈值0.85，则对其它所有类别的预测结果进行过滤。

此外我们还发现了下面的现象并使用一些策略进行处理：

- 模型预测的结果有较多的假阳性框，主要表现为：
    1. 同类重叠：同一个区域附近有很多同种类别的预测框；
    2. 异类重叠：同一个区域附近同时预测多个不同类别的预测框；

3.2 融合策略

box融合时考虑分两种场景：

1. 同个模型交叠滑窗预测：box_voting
2. 不同模型之间融合：weight box fusion(WBF) : WBF思想和box voting类似， 区别在投票权重的设置，WBF是根据box confidence score做线性加权，WBF对加权后的方框confidence scores 线性平均，除此外还考虑了方框出现次数与融合模型个数的调整（比如融了6个模型，但对某个Roi中做出预测结果的只有6个中的3个模型，则对confidence 进行线性降权）