kaggle乳腺癌检测竞赛总结

1. 赛题分析

医学图像分类问题。训练集样本量54000多，训练集正样本量1000多一点。图像是DICOM格式且图像大小不一，图像分辨率高（最大的超过4000\*2000），肿瘤目标小。

1. 数据处理
2. DICOM转png。第三方库用dicomsdl和pydicom。线上推理主要用GPU处理，少量无法用GPU处理的用CPU处理，方案参考其他选手，大约耗时两小时多一点。
3. 分割。参考其他选手方案，采用了Otsu's thresholding after Gaussian filtering。但手动查看切割后的图像，效果不是很好。其他选手有的采用了目标检测模型，有的根据DICOM的中心点坐标直接对原图进行了切割。总的来看，都不如用cv2.connectedComponentsWithStats的方案方便。
4. 直接resize到(1440,720)。比例是基本合适的。图像size越大，效果越好，但越吃资源。尝试了使用3090多卡训练(2048,2048)，但效果不好。后使用单卡A100训练(2560,1536)，因时间关系，只跑完1折，5折融合线上效果可能会更好。
5. 尝试了合并csv数据，但没有效果。其他选手大部分也没有使用csv数据。
6. 使用了StratifiedGroupKFold，5折。但根据其他选手，4折效果会更好，因为正样本数量较少，分布很容易不均匀。
7. 没有使用外部数据。但根据其他选手，使用外部数据，效果会更好。
8. 数据增强

训练集使用了垂直翻转和随机旋转，验证集使用了随机翻转。因为做了尝试发现图片稍有变化就会导致分数下降，考虑到医学图像的特性所以没有过多的做数据增强。其他选手有些使用了较多的数据增强，有些较少。

对正样本上采样的同时做了cutout，效果不好。根据其他选手的方案，直接做上采样效果会比较好。

尝试了对正样本做cutmix，效果不好。

1. 模型

使用了预训练的EfficientNet-B2。线下不修改全连接层效果比修改全连接层要好。

尝试了GeM pooling，效果不好。但其他选手有使用，不确定提升效果有多大。

尝试了Focol Loss，效果不好。根据其他选手的分享，情况相同。

根据第一名的方案，Positive label smoothing (0,0.8)提升明显，相比label smoothing(0.05,0.95)可以在线下提升约0.07分。

尝试使用了预训练的EfficientNet-V2-S，但相比B2提升不明显，可能是因为需要调一下参数。大部分top10选手都使用了EfficientNet-V2-S。

1. 其他

训练和推理均采用了混合精度。

有尝试使用TensorRT，不过因为时间关系最终未采用。根据其他选手分享，速度仅提升30%，但线上分数有明显下降。Top10选手最终都没有使用TensorRT。

线上推理时，空间和时间有限，所以采用了批式的方案，按批进行：图像处理-保存图像-推理-删除图像。

尝试了3090多卡训练（6卡），分别实现了DP模式和DDP模式。DDP模式下，实现了BN同步，多卡推理，梯度累加。对于高分辨率的版本，数据量较大但不是特别大的情况下，3090多卡不如单卡A100。

1. 总结

在本次竞赛中，针对医学图像样本极不均衡且正样本量较少、癌症区域目标较小的难点，进行了一些探索。在工程上进行了一些尝试。因为算力和时间有限，未能取得较好的排名，但潜在提升空间很大。总的来看，放大图像size，使用性能更好的预训练模型，上采样，Positive label smoothing，模型融合，这些常规方法都能有效地提升模型效果，且实现起来并不难。