## 防止过拟合及模型调优

### 2.2.1 增加 dropout

在基本的 9 层模型上我们训练 9000 个样本，其中设置 1000 个为验证集，epoch 设置为 20，可以得到准确率图像为：

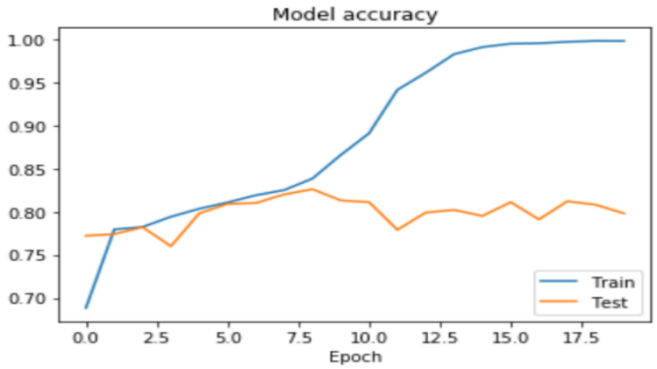


**图 2.2-1 基础 9 层模型准确率**

不难看出在训练集上模型的准确率在第 15 个 epoch 后逼近了 1，而验证集上模型的表现与训练集差的很远，出现了过拟合的情况。通过阅读[8]，我尝试在模型中采用 dropout 的方法去解决这个问题，得到的准确率图像如下图所示：

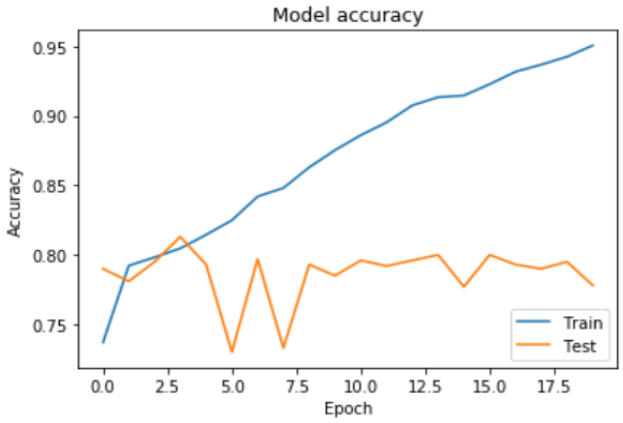


**图 2.2-2 增加 dropout 后的准确率**

我们可以看到，在增加了 dropout 策略之后，经过 20 个 epoch 训练集上的准确率与之前相比稍有下降，训练集和测试集的准确率差值略有降低，过拟合问题稍有改善。

## 2.2.2 正则化方法

根据[8]我又在模型上尝试通过正则化的方式继续改善过拟合的问题，对 L1 正则化分别选取正则化强度为 0.0001，0.001，0.01 的情况，得到的准确率结果如下图所示

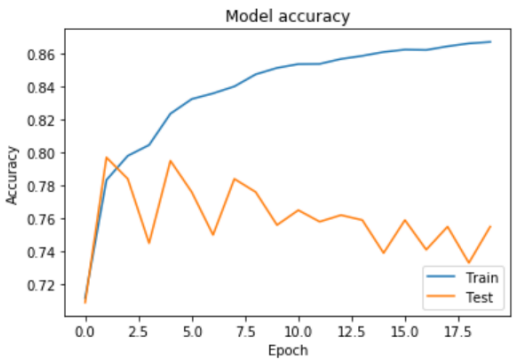


图 2.2-3 α =0.0001

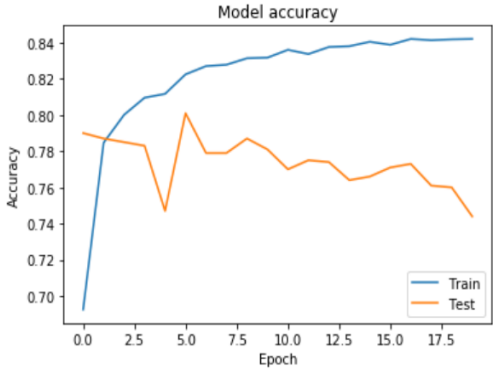


图 2.2-4 α =0.001

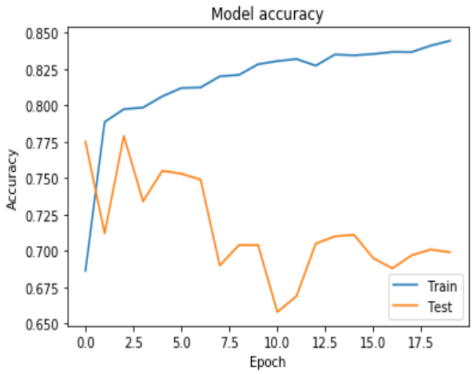


图 2.2-5 α =0.01

随着正则化强度的增加，训练集上的准确率在逐渐地降低，同时测试集上的准确率也在发生变化。当正则化强度增加到 0.01 时，我们可以看到测试集上的准确率开始有明显的下降，出现了欠拟合的情形。相比之下，当 α=0.001 时，训练集的准确率曲线相对比较平稳，与测试集的准确率之差也最小，故选择 0.001 作为正则化强度的参数。

### 2.2.3 调整 batch_size 改善测试集表现

在前两步实验基础之上为了进一步改善测试集的表现，通过阅读[9]，我们尝试改变了模型的 batch_size，将 batch_size 从原来的 10 分别改为 32，64，128，获得的准确率结果如下图所示：

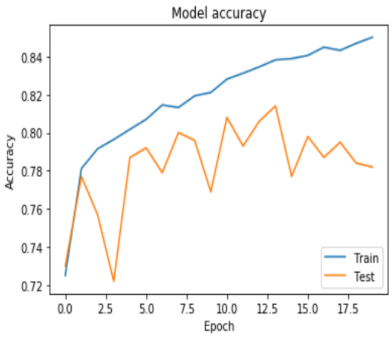


图 2.2-6 batch_size=32

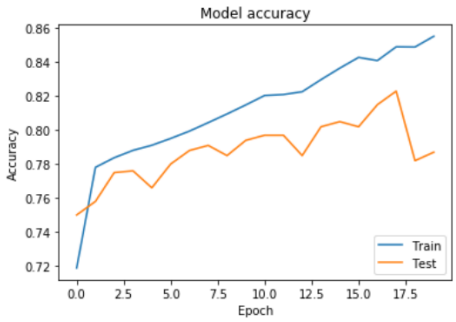


图 2.2-7 batch_size=64

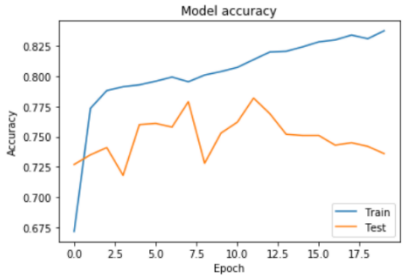


图 2.2-7 batch_size=128

通过对准确率曲线的对比，我们可以发现，当 batch_size 取 64 时训练集与测试集的曲线更为接近，相较 32 与 128 而言是更好的选择。

## 2.2.4 调整权值初始值改善测试集表现

通过以上三个步骤的优化，我们已经得到相对较好的训练模型，在此之上，我们阅读了[10]，尝试通过对神经网络权值的不同初始化方式得到更好的训练结果，除了默认的权值初始化 glorot_normal 方式外，我们还尝试了 lecun_normal, lecun_uniformal 权值初始化方式，得到的准确率图像如下图所示：

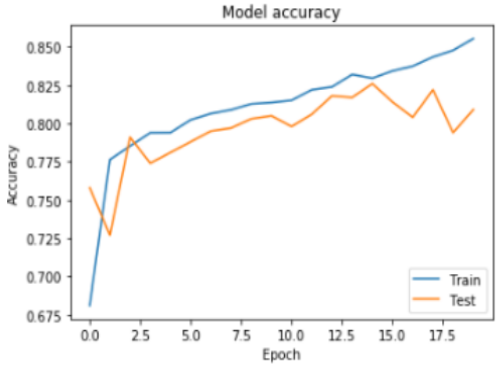


图 2.2-8 lecun_normal

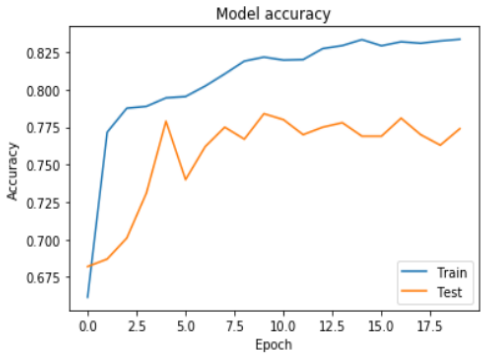


图 2.2-9 lecun_uniformal

从上图我们不难看出，当采用 lecun_normal 的权值初始化方式时，相较默认参数与 lecun_uniformal 的初始化方式，我们可以发现，训练集与测试集的准确率差距相较最低，并且测试集的准确率更高，因此我们选择使用 lecun_normal.

## 2.2.5 图片截取优化模型性能

由于癌症细胞只会出现在图片的中央 32*32 区域，而原图为 96*96，因此我们考虑对图片进行截取，我们分别将图片截取到 96*96，84*84，72*72，60*60，48*48，32*32，进行训练，这里我们使用 ROC 曲线对模型训练的结果进行评价，曲线下面积越大代表模型越好。

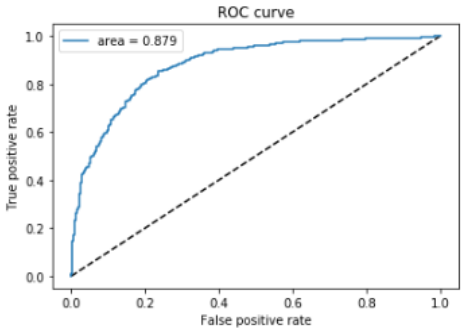


图 2.2-10 96*96

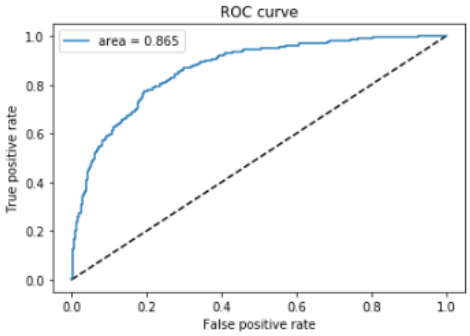


图 2.2-11 84*84

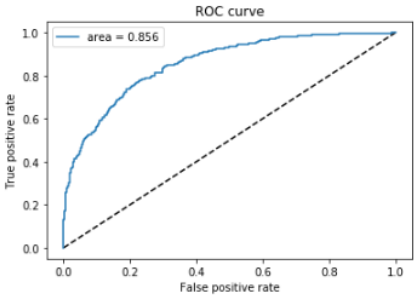


图 2.2-12 72*72

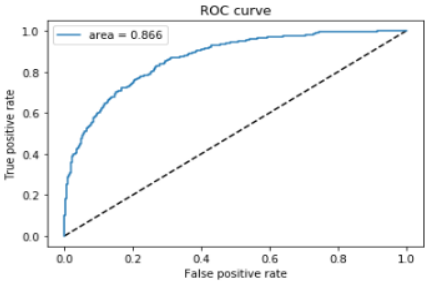


图 2.2-13 60*60

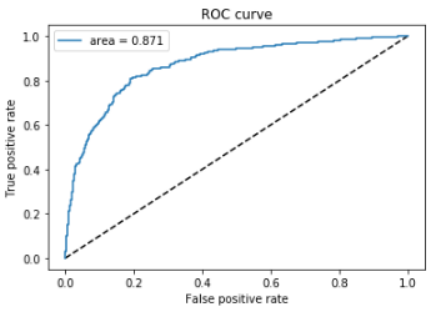


图 2.2-14 48*48

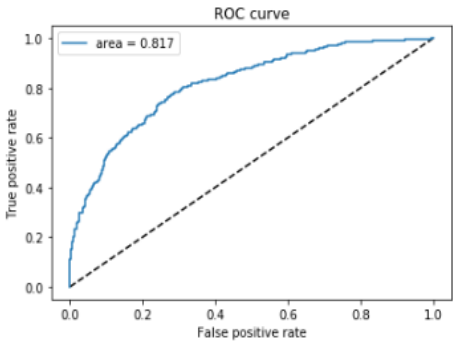


图 2.2-15 32*32

以上数据汇集于下图：

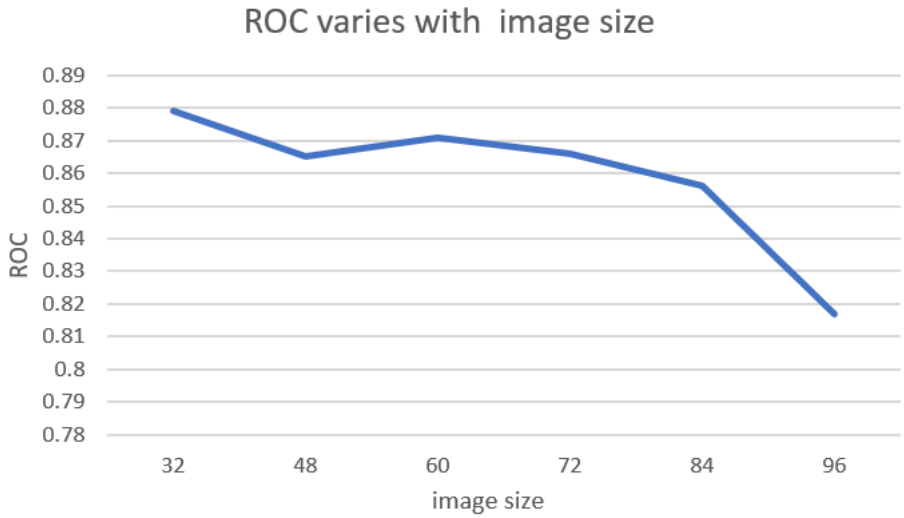


**图 2.2-16 ROC 取值于图片大小关系图**

通过上图，我们可以看出在被训练图片截小之后，ROC 的值也在跟着变小。因此我们可以得到合理的结论：癌症细胞虽然只存在于中央 32*32 区域，但是周边的组织具备很多关于癌症细胞信息，因此在逐渐截取图片后并没有让我们的结果变得更好。因此最终我们选择的图片大小是 96*96。2.3 优化器及学习率调优