# 实验内容

本次实验的主要目标是创建一种算法，以识别从较大的数字病理扫描图片中获取的小图像块中的转移性癌症。实验提供给了我们96\*96像素分辨率的淋巴结组织的显微图像，我们要做的是给出在图像的32\*32的中心区域是否有转移性癌组织的概率，这是一个图像的二元分类任务。

实验的数据是PatchCamelyon（PCam）基准数据集的略微修改版本，（原始PCam数据集由于其概率抽样而包含重复图像，但是，在Kaggle上呈现的版本不包含重复项）。其中大概有220000张图像用于训练，57000张图像用于测试。

# 实验步骤

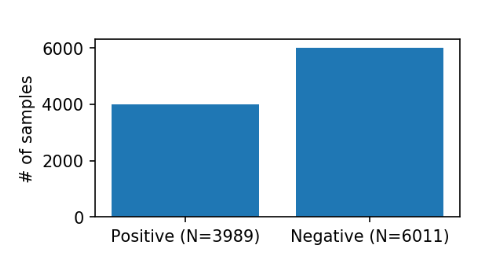
1. 加载数据
2. EDA

EDA（Exploratory Data Analysis）的目的是

* 1. 看一看数据集

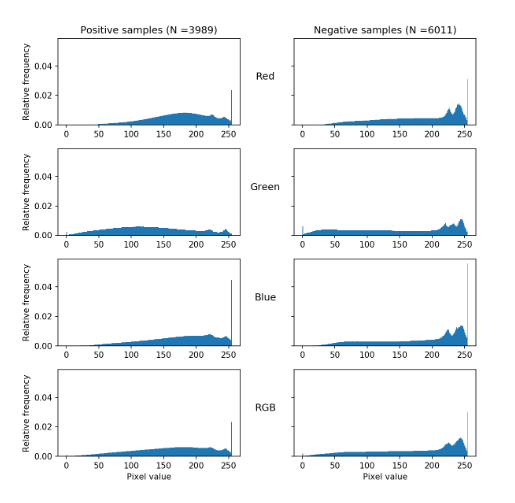
可以让我们先对两种类别有一个大概的了解，这对于后面选择什么特征作为分类依据很重要

* 1. 了解两种类型（有癌症细胞或无癌症细胞）的分布



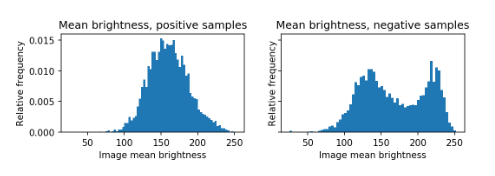
可以看出，在训练集中负样本多于正样本，且正负比例近似为2：3。也就是全部判定为没有癌症细胞也能有60%的准确性，这可能导致分类器偏向于负样本。为了避免分类器中的偏差并改善训练期间的稳定性，我们可以采取过采样和欠采样的方法

* 1. 看一看图片的一些特征，比如RGB通道的分布，平均亮度等
     1. RBG channel



可以看出：

1. 与正样本相比，负样本像素更多在高亮度部分，特别是在绿通道中
2. 正样本中，相比于其他两通道，绿通道中的像素更暗。而负样本不是这样
3. 正负样本中都存在大部分亮白色区域
   * 1. 平均亮度



可以看出：

1. 正样本的平均亮度分布接近于均值为150正态分布
2. 负样本近似双峰分布，峰值为140和225

通过EDA，我们可以得到以下结论：

* 1. 正负样本在像素的分布和平均亮度的分布上有明显的区别，我们的模型可以利用这一点
  2. 一些图片包含了很亮的区域，可能是记录过程中的人为因素导致，我们需要找到一种方法解决他们。他们在正负样本中都有，所以不能简单当作一个特征
  3. 负样本多于正样本很多，可能需要调整

1. 划分训练集与验证集