# 实验内容

本次实验的主要目标是搭建一种图片分类模型，这种模型能够识别图像块是否存在癌症组织。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测试集图像的32\*32的中心区域是否有癌组织的概率，这是一个图像的二元分类任务。

实验的数据是PatchCamelyon（PCam）基准数据集的修改版本，（原始PCam数据集由于其概率抽样而包含重复图像，在Kaggle上呈现的版本不包含重复样本）。图片从较大的数字病理扫描图片中截取。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. 分析问题特征：

淋巴结癌细胞转移有以下特征：

* 1. 关键特征：外来细胞群（经典位置：淋巴结缘窦）
  2. 具有恶性肿瘤细胞学特征的细胞
  3. 核多形性(大小、形状和染色的变化)
  4. 异性核：
     1. 核增大
     2. 不规则核膜
     3. 不规则的染色质图案，尤指不对称
     4. 大的或不规则的核仁
     5. 丰富的有丝分裂图像
  5. 恶性肿瘤中可见的细胞排列结构：高度可变，依赖于肿瘤类型和分化
     1. 腺形成
     2. 单细胞
     3. 集群的细胞

由此可得：不规则的核形状、大小或染色深浅可以提示转移，那么如何将这类数据转换以进行建模？

已知图像的标签只受中心区域（32x32px）的影响，因此只将数据裁剪到该区域对训练模型是有益。但是如果我们将图像块切的过小，一些周围环境中的信息就会被去除。但是调查可知，32x32px效果比48x48px的性能差。

1. 提升数据质量：

可以检查数据是否包含较差的数据（损坏数据或者没有特征的数据），删除他们可以提高数据的整体质量。

预处理和数据扩充：

有几种方法可以避免过度拟合：数据集扩充、增强特征、正则化和选择复杂度较低的模型。每一次增强后相应的数据处理时间都会时间，这被称为测试时间增强(TTA)，如果我们对每幅图像进行多次处理并推断出平均时间，可以改善最终的预测结果。

**可以对图像进行如下处理改善数据集质量：**

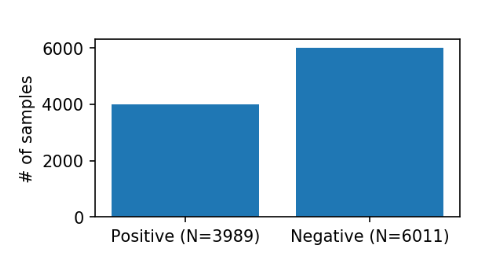
1. 随机旋转
2. 随机选择数据集
3. 随机翻转（水平竖直均可）
4. 随机缩放
5. 随机增亮
6. 高斯模糊
7. 数据集预筛选：
   1. 剔除过亮或者过暗的图像：如果图像过暗或者过亮可能是由于曝光不良或者剪切到空白部分造成的，有部分空白的数据依旧有效，但是曝光不良的图片则损失了特征所以不能使用。
      1. 筛选出一个过暗样本和6个过亮样本
8. EDA

EDA（Exploratory Data Analysis）的目的是

* 1. 对数据集进行粗略观察

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

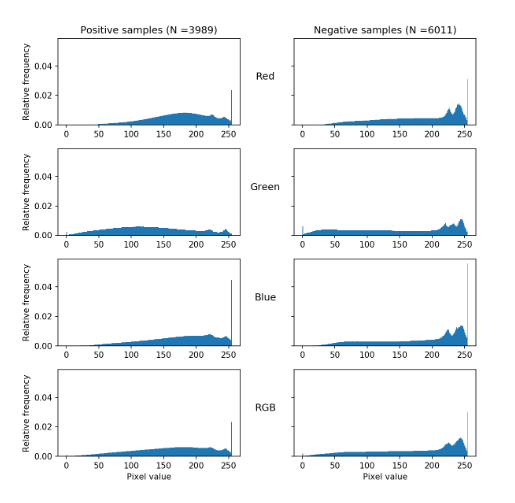
* 1. 了解两种样本类型（有癌症细胞或无癌症细胞）的分布



**图 1 正负样本分布状态**

可以看出，在训练集中负样本多于正样本，且正负样本数比例近似为2：3。也就是全部判定为没有癌症细胞也能有60%的准确率，这种样本分布可能导致分类器偏向于判断结果为负样本。为了避免分类器中的偏差并提升训练期间的稳定性，我们可以采取过采样和欠采样的方法

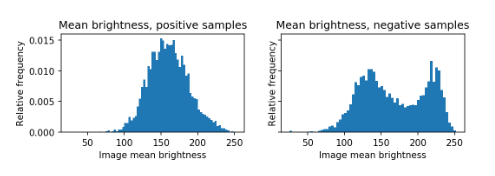
* 1. 观察图片的部分特征，比如RGB通道的分布，平均亮度等
     1. RBG channel



**图2 样本在RGB通道上的分布**

可以看出：

1. 与正样本相比，负样本像素更多在高亮度区域，特别是在绿通道中
2. 正样本的绿通道中的像素相比其他两个通道分布在低亮度区域。
3. 负样本的三通道中的像素在高亮度区域均有峰值
4. 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

① 由于内存的限制，我们通过指定所选图片的索引来进行划分，而不用建立新的存储区。我们将80%的数据用于训练，20%的数据用来验证我们的模型可以推广到新数据。因为两个标签中并没有十分罕见的类别，随机分割不会造成数据分类的减少。

② 为了避免之前数据排列的影响，我们重新对数据进行顺序。

1. 建立神经网络模型

建立一个简单的卷积神经网络模型，它包含三块：卷积层，组正交，池化和丢弃

1. 配置模型训练时的参数
   1. batch\_size是训练一个神经网络时很关键的参数。在训练的时候，我们将训练集分成一个个的batch，然后用batch一个个的去训练我们的网络
   2. 卷积核大小：
2. 训练并且验证模型

每个batch训练三次，每次训练的时候对batch中数据的进行重排