解决方法如下：

1 采样，对小样本加噪声采样，对大样本进行下采样

采样分为**上采样（Oversampling）**和**下采样（Undersampling）**，上采样是把小种类复制多份，下采样是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。

随机采样最大的优点是简单，但缺点也很明显。上采样后的数据集中会反复出现一些样本，训练出来的模型会有一定的过拟合；而下采样的缺点显而易见，那就是最终的训练集丢失了数据，模型只学到了总体模式的一部分。

2 数据生成，利用已知样本生成新的样本

其中最常见的一种方法叫做SMOTE，它利用小众样本在特征空间的相似性来生成新样本。对于小众样本$x_i \in S_{\min}$，从它属于小众类的K近邻中随机选取一个样本点$\hat{x}_i$，生成一个新的小众样本x_{new}：$x_{new} = x_i + (\hat{x} - x_i) \times \delta$，其中$\delta \in [0,1]$是随机数。

SMOTE,即该算法构造的数据是新样本，原数据集中不存在的。该基于距离度量选择小类别下两个或者更多的相似样本，然后选择其中一个样本，并随机选择一定数量的邻居样本对选择的那个样本的一个属性增加噪声，每次处理一个属性。这样就构造了更多的新生数据。（优点是相当于合理地对小样本的分类平面进行的一定程度的外扩；也相当于对小类错分进行加权惩罚

3 进行特殊的加权，如在Adaboost中或者SVM中

4 采用对不平衡数据集不敏感的算法

5 改变评价标准：用AUC/ROC来进行评价

6 采用Bagging/Boosting/ensemble等方法

7 在设计模型的时候考虑数据的先验分布

8 一分类

对于正负样本极不平衡的场景，我们可以换一个完全不同的角度来看待问题：把它看做一分类（One Class Learning）或异常检测（Novelty Detection）问题。这类方法的重点不在于捕捉类间的差别，而是为其中一类进行建模，经典的工作包括One-class SVM等。

如何选择

- 在正负样本都非常之少的情况下，应该采用数据合成的方式；

- 在负样本足够多，正样本非常之少且比例及其悬殊的情况下，应该考虑一分类方法；
- 在正负样本都足够多且比例不是特别悬殊的情况下，应该考虑采样或者加权的方法。
- 采样和加权在数学上是等价的，但实际应用中效果却有差别。尤其是采样了诸如Random Forest等分类方法，训练过程会对训练集进行随机采样。在这种情况下，如果计算资源允许上采样往往要比加权好一些。
- 另外，虽然上采样和下采样都可以使数据集变得平衡，并且在数据足够多的情况下等价，但两者也是有区别的。实际应用中，我的经验是如果计算资源足够且小众类样本足够多的情况下使用上采样，否则使用下采样，因为上采样会增加训练集的大小进而增加训练时间，同时小的训练集非常容易产生过拟合。对于下采样，如果计算资源相对较多且有良好的并行环境，应该选择Ensemble方法。

难分样本问题

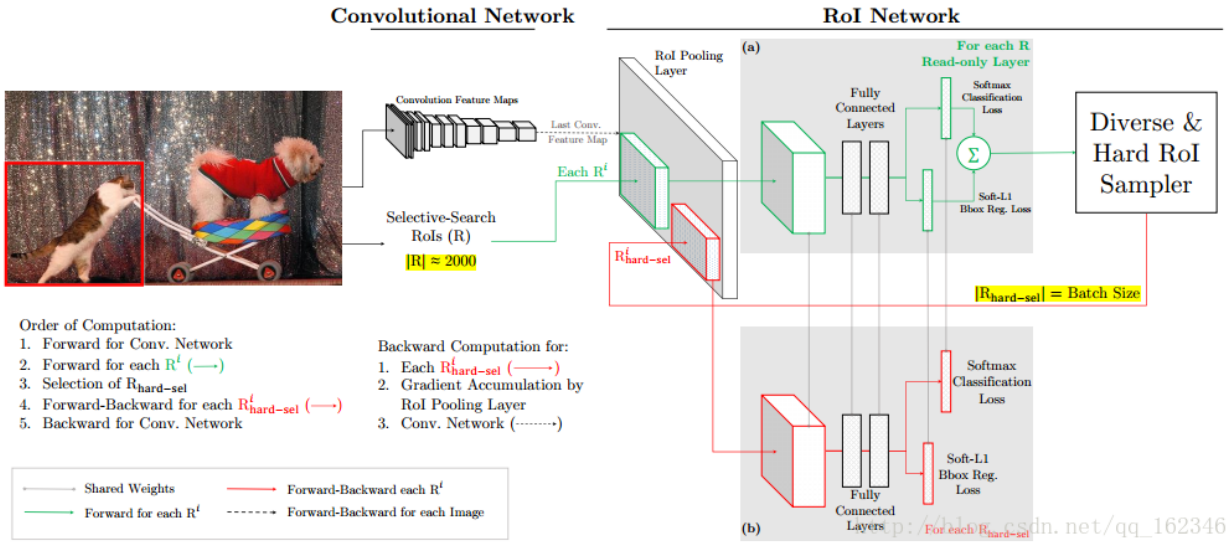
1.focal loss

通过模型预测的概率pt，使用(1-Pt)来代表样本难分程度。可以理解为模型对某个样本预测属于其真实label的概率越高，则说明该样本对此模型比较容易学习，反之则难分。

2、《ScreenerNet: Learning Self-Paced Curriculum for Deep Neural Networks》论文提出一个附加网络来帮助主网络区分样本难易程度。

3、《Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis》论文通过对一张图像进行数据增强生成多张图像，然后使用模型预测每张图像的概率。根据多张相同label的增强图像的概率分布区分其样本难易程度。

4、《OHEM: Training Region-based Object Detectors with Online Hard Example Mining》论文提出先使用模型输出概率，据此选出部分难分样本，然后根据这些样本，更新网络参数。



上图绿色和红色分为两个网络但共享权限，通过将提取的RoI传入绿色的只读网络（只进行forward），计算出每个RoI的loss。根据loss排序（可使用NMS）选出部分样本，再输入红色网络（进行forward和backward）学习并进行梯度传播。文中提出另一种办法，在反向传播时，只对选出的样本的梯度/残差回传，而其他的props的梯度/残差设为0。但容易导致显存显著增加，迭代时间增加。

在深度学习中，有哪些解决样本不平衡的方法?

深度学习同样属于机器学习中的一种典型方法，所以在机器学习中适用的方法在深度学习中同样适用。比如说：扩大数据集、类别均衡采样、人工产生数据样本，添加少类别样本的来loss惩罚项等。

对于数据的方法，这里我们重点介绍：类别均衡采样
把样本按类别分组，每个类别生成一个样本列表，训练过程中先随机选择1个或几个类别，然后从各个类别所对应的样本列表里选择随机样本。这样可以保证每个类别参与训练的机会比较均等。
上述方法需要对于样本类别较多任务首先定义与类别相等数量的列表，对于海量类别任务如ImageNet数据集等此举极其繁琐。海康威视研究院提出类别重组的平衡方法。
类别重组法只需要原始图像列表即可完成同样的均匀采样任务，步骤如下：
1. 首先按照类别顺序对原始样本进行排序，之后计算每个类别的样本数目，并记录样本最多那个类的样本数目。之后，根据这个最多样本数对每类样本产生一个随机排列的列表， 然后用此列表中的随机数对各自类别的样本数取余，得到对应的索引值。接着，根据索引从该类的图像中提取图像，生成该类的图像随机列表。之后将所有类的随机列表连在一起随机打乱次序，即可得到最终的图像列表，可以发现最终列表中每类样本数目均等。根据此列表训练模型，在训练时列表遍历完毕，则重头再做一遍上述操作即可进行第二轮训练，如此往复。 类别重组法的优点在于，只需要原始图像列表，且所有操作均在内存中在线完成，易于实现。

从图像和文本数据的角度出发，我们来看一下有哪些方法？

对于图像数据，解决样本不平衡问题，在深度学习中会用到的方法包括了：类别均衡采样（上已经描述），可以用来解决分类问题。
另外，在笔者看来还值得介绍的方法包括了：OHEM和focal loss。

- OHEM

OHEM（online hard example miniing）算法的核心思想是根据输入样本的损失进行筛选，筛选出hard example，表示对分类和检测影响较大的样本，然后将筛选得到的这些样本应用在随机梯度下降中训练。在实际操作中是将原来的一个ROI Network扩充为两个ROI Network，这两个ROI Network共享参数。其中前面一个ROI Network只有前向操作，主要用于计算损失；后面一个ROI Network包括前向和后向操作，以hard example作为输入，计算损失并回传梯度。作者将该算法应用在Fast RCNN中，网络结构还是采用VGG16和VGG_CNN_M_1024，数据集主要采用VOC2007，VOC2012和COCO数据集。
算法优点：1、对于数据的类别不平衡问题不需要采用设置正负样本比例的方式来解决，这种在线选择方式针对性更强。2、随着数据集的增大，算法的提升更加明显（作者是通过在COCO数据集上做实验和VOC数据集做对比，因为前者的数据集更大，而且提升更明显，所以有这个结论）。

算法的测试结果：在pascal VOC2007上的mAP为78.9%，在pascal VOC2012上的mAP为76.3%。注意，这些结果的得到包含了一些小tricks，比如multi-scale test（测试时候采用多尺度输入），bounding box的不断迭代回归。

代码的github地址：https://github.com/abhi2610/ohem

- Focal Loss

Focal loss主要是为了解决one-stage目标检测中正负样本比例严重失衡的问题。主旨是：ssd按照ohem选出了loss较大的，但忽略了那些loss较小的easy的负样本，虽然这些easy负样本loss很小，但数量多，加起来的loss较大，对最终loss有一定贡献。作者想把这些loss较小的也融入到loss计算中。但如果直接计算所有的loss，loss会被那些easy的负样本主导，因为数量太多，加起来的loss就大了。也就是说，作者是想融入一些easy example，希望他们能有助于训练，但又不希望他们主导loss。这个时候就用了公式进行衰减那些easy example，让他们对loss做贡献，但又不至于主导loss，并且通过balanced crossentropy平衡类别。

OHEM是只取3:1的负样本去计算loss，之外的负样本权重置零，而focal loss取了所有负样本，根据难度给了不同的权重。
focal loss相比OHEM的提升点在于，3:1的比例比较粗暴，那些有些难度的负样本可能游离于3:1之外。之前实验中曾经调整过OHEM这个比例，发现是有好处的，现在可以试试focal loss了。