色情图像识别技术文档 V0.1

作者：@LucasX　　Email: xulu0620@gmail.com

一、 可能的技术方案

目前，对色情图像的识别总结了以下几类解决方案：

方案 1：皮肤区域检测法

与正常图片相比，色情图片的最大特征就是一张图像里包含了大量暴露的皮肤区域，因此可初步通过皮肤区域检测法来快速完成对色情图片的识别。主要方法如下：

1) 将 RGB 通道的图像转为 HSV 或 YCbCr 通道的图像：
2) 分析转换后三通道的图像信息，目前对皮肤区域的定义区间如下：
   a) YCbCr：

$$\begin{cases} 80 \le Cb \le 120 \\ 133 \le Cr \le 173 \end{cases}$$

   b) HSV：

$$\begin{cases} 0 < H < 0.25 \\ 0.15 < S < 0.9 \\ 0.2 < V < 0.95 \end{cases}$$

3) 皮肤区域统计：

$$\mathrm{skin}_{\mathrm{area}} = \frac{\#skin\_color\_pixels}{\#image\_pixels\_total}$$

4) 阈值参数选定：
   选定阈值 t，当$\mathrm{skin}_{\mathrm{area}} > t$时，可认为该图片为色情图片。
5) **总结：**
   该方法通过对图像通道的变化分析，可快速检测出嫌疑色情图像，操作简单，但是对于“性感——如女性泳装照”、“正常——如男性裸露上身、裸体婴儿皮肤”等均会误判为色情图像。

方案 2：机器学习方法

对于一张图像，我们可以广义地定义为 3 类：“正常”、“性感”、“色情”。因此，可以利用机器学习的方法来对训练集图像进行训练，将训练后的模型用于预测。该方法主要思路如下：

1) 分别搜集“正常”、“性感”、“色情”三类大量带有标记的图像样本，保证总数至少为 10 万张，且 3 类样本分布需尽可能均衡，即每类都应为 3 万张左右。
2) 提取图像 LBP、HOG、SIFT 特征，尝试 SVM、Adboost、KNN、Random Forest 等分类器模型进行训练，选择分类性能最佳的模型。

总结：该方法借助机器学习对训练集图像进行训练，因此相比于方案 1，该方法精度会有

明显提升；但是需要大量标注样本（不少于 10 万张）。

方案 3：深度学习方法

近年来，深度学习在计算机视觉领域超越传统机器学习方法取得了突破性的进展，因此可借助深度卷积神经网络对图像进行训练，依据数据量的大小设计不同的模型。将其转换为基于深度学习的图像分类或强监督细粒度图像识别（需要关键部位的 bbox 信息）问题。

Github 上有童鞋整理了一份 NSFW 数据集，数据集下载地址：
https://github.com/alexkimxyz/nsfw_data_scrapper.git

总结：该方法准确率和鲁棒性均可以达到最高，但需要海量标记样本。Fine-tune 或许是个不错的思路。

方案 4：数据挖掘方法

该方法旨在对平台用户积累的历史行为数据进行深入数据挖掘，少数色情图片上传者与大部分正常用户在行为上往往会有很大的不同。该方法对应于数据挖掘中的“离群点分析”。

总结：该方法检测效率较高（对数据库的操作效率远胜于二进制图像数据的操作），但需要平台提供海量用户历史行为数据。

方案 5：MD5 检测

可通过比对上传文件的 MD5 来检测不良图片或视频，该方法效率较高，适合于云盘内容检索等应用场景；但对于直播等 UGC 平台，则该方法不适用。

方案 6：训练音频分类器

视频文件的处理，若按照深度学习与计算机视觉的方法去做视频内容分析，势必会存在极高的计算复杂度。因此，可对视频文件的音轨进行分离，分别提取色情视频与正常视频的音频特征，从而将其转换为机器学习二分类问题。

FFmpeg 分离音频，VGGish 训练音频分类模型。

总结：该方法效率上比视频分析高，但是对无声色情视频无效。

方案 7：迁移学习方法

考虑到昂贵的图像标注成本，因此可借助迁移学习（Transfer Learning）的思想，从 ImageNet Pretrained Model 做 fine-tune。

二、 实际场景应用方案

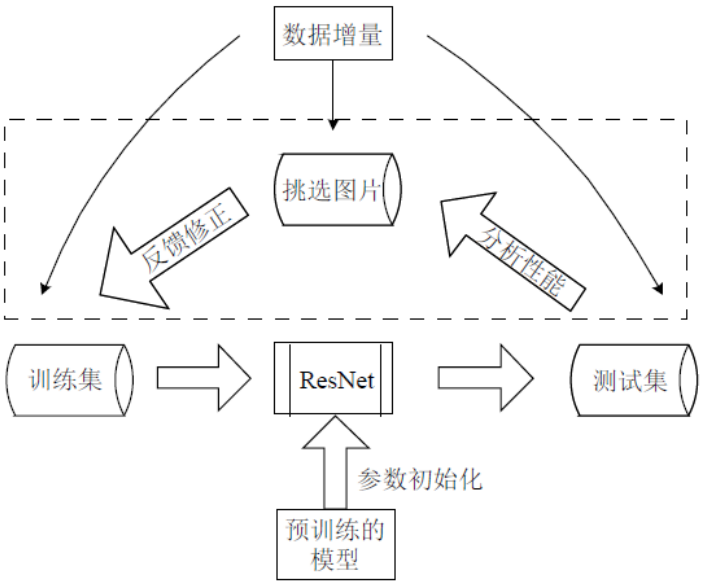
根据初步实验结果显示，将其视为一个简单的 triple classification 任务，远远不能满足实际需求。因为(normal/sexy/porn) semantic meaning 飘得太厉害，尤其是对于 normal 和 porn 这两类，符合要求的图像细粒度类别 intra-class diversity 非常大。模型在学习高层 semantic meaning 时就非常容易混淆。因此实际应用场景下采取了如下方案：

1. 对一张 size 为$W \times H$的图片，因为大多数图片的色情敏感区域只占其中很小一部分，因此我们将其按短边进行等比例缩放到$224 \times L$，然后以$step = 50$的方式进行滑窗，若在滑窗过程中将任何一个子窗口块判断为色情，则结束滑窗，认定该图片为色情。

   注：基础网络可选取 ResNet50/DenseNet101。
2. 对色情/性感/正常这 3 类图片进行进一步的细粒度划分：
   - 类别 0: 良性图片中的正常人物类图片;
   - 类别 1: 良性图片中的正常类图片(无人物);
   - 类别 2: 男性下身类图片，属于不良图片中特征明显的性器官裸露类;
   - 类别 3: 女性上身类图片，属于不良图片中特征明显的性器官裸露类;
   - 类别 4: 女性下身类图片，属于不良图片中特征明显的性器官裸露类;
   - 类别 5: 不良图片中的性行为姿势类图片，一般包含大量的皮肤裸露;
   - 类别 6: 包含儿童色情类的不良图片;
   - 类别 7: 包含色情信息的卡通类图片;
   - 类别 8: 不良图片中的低俗类，包含裙底偷拍、人体内衣敏感部位特写等.
3. 在网络模型输出各类别的 softmax probability 分值后，分类为良性图片和不良图片的方法如下:

   1) 取分值最大的类别n 和分值s;
   2) 如果n=0 或 1，分类为良性图片;
   3) 如果n 是 2 ~ 4，$s = s \times 1.2$;
   4) 如果n 是 5 ~ 8，$s = s \times 0.92$;
   5) 如果$s >= 0.85$，分类为不良图片

   因为其中 2、3、4 这三类是主要的不良图片构成部分,也是特征最明显的类别. 经过试验分析，在统计分值时对这三类硬色情图片的分值乘以系数 1.2，可以增强过滤的准确率. 低俗类的属性特征相对模糊，其边界难以与性感图片区分开. 性行为姿势类、儿童色情类的色情标准同样具有一定的模糊性. 卡通漫画本身的描述方法就具有夸张性，卡通色情类的判定也应比正常色情类弱。将 5、6、7、8 类软色情图片的分值乘以 0.92，进行一定的弱化，可以减少定义模糊的图片类型的干扰. 将最后得到的分值与阈值 0.85 进行比较，当大于等于 0.85 时分类为不良图片。
4. 在每进行 20 个 epoch 之后，从每个类别中随机挑选 100 张图片，分别测试模型的分类准确率. 对准确率小于 0.9 的类别，定向增加其训练样本的容量，包括增加与测试样本具有相似特征的图片以及不同肤色的图片等边缘案例，再继续进行训练. 多次重新设计和构建训练数据集直到模型可以挖掘出更优质的特征。


数据增量
挑选图片
反馈修正
分析性能
训练集
ResNet
测试集
参数初始化
预训练的
模型