基于卷积神经网络的人脸检测

项目报告

郑哲东 2015.11

一．项目目标：

对视频单帧或图像中人脸进行检测，输出人脸框（bounding box）。当前的主流算法可以检测多种形态姿势的人脸，包括一些遮挡和形变，以及平面上和非平面上的旋转。但依旧面临一些漏检错检的问题。本项目将在目前 landmark 的人脸检测数据集 FDDB 上与其他方法做检测评估和比较。

二．项目简介：

人脸检测是传统做人脸对齐、人脸识别的第一步，故对速度和精确度以及召回率都有一定要求。在本项目传统意义中的困难具体表现在以下这几个方面：

1. 人脸内在变化：表情变化、人脸朝向变化（多姿态）
2. 人脸外在变化：光照变化、一定的遮挡

而通过卷积神经网络这些问题都或多或少的克服了。故在实现过程中的困难在于以下几个方面：

1. 负样本数据的人工筛选（指去除混入的人脸图片）
2. 级联的训练多个网络 （需要等前一个网络训完之后）

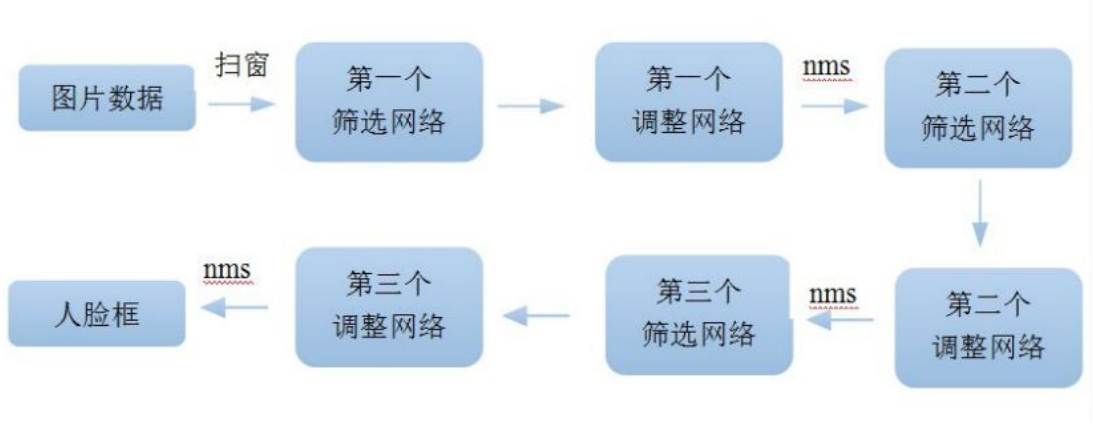
三．实验数据：

AFLW 数据集+CASIA 数据集（通过 dlib 筛选）

四．实验方法：

依据 2015 CVPR Cascade CNNs for Face Detection.(L.Hao,Z.Lin etc.)的方法对其重现及优化部分代码和细节。

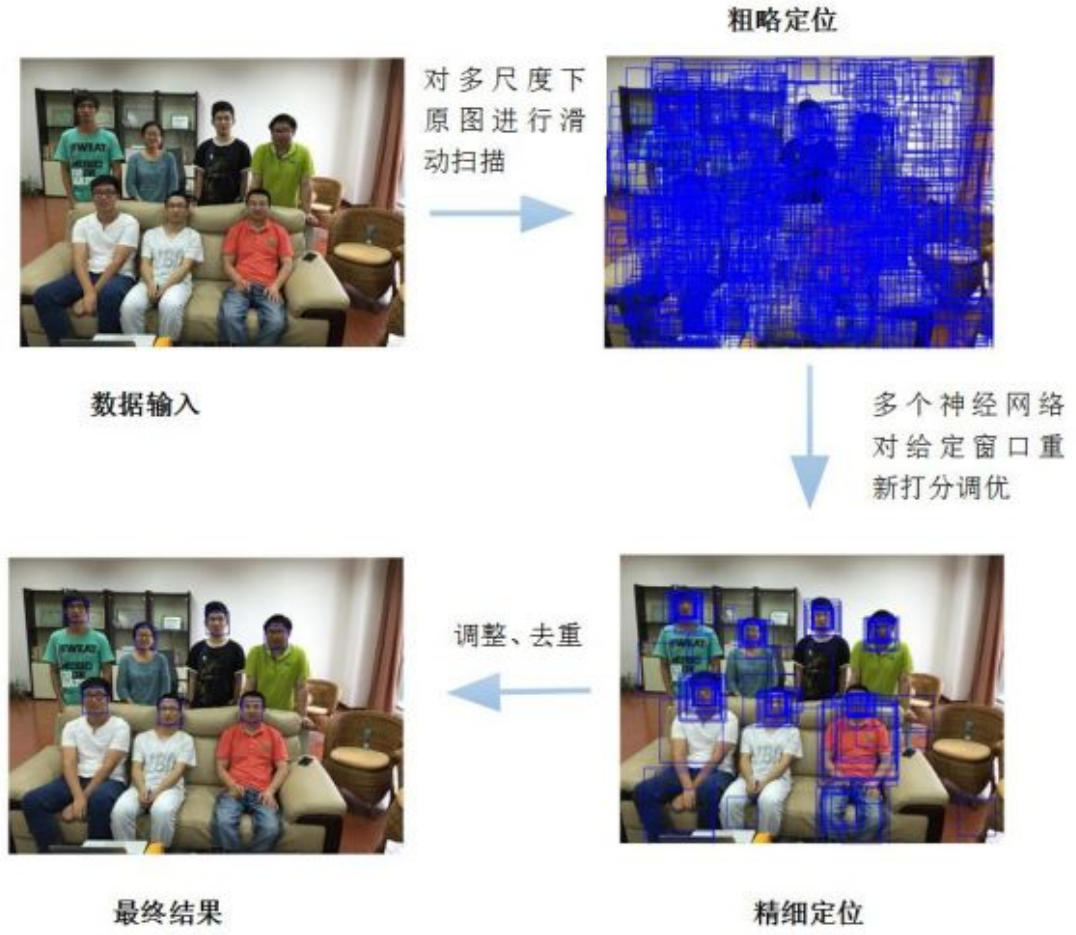
1.对原论文方案的重新描述



a.筛选网络：
通过预先训练好的神经网络得到对于 bounding box 的置信度，输出两维向量。网络结构比较一般，不做赘述。要点在于：利用了小尺度网络全连接层的输出来加强大尺度网络的分辨能力，达到多尺度融合的效果。同时，网络深度都较浅，有助于检测的提速。

b.校正网络：
本论文的另一个创新点在于使用分类而不是回归，通过预先训练好的神经网络得到对于 bounding box 位置的预测，输出 45 维位置的预测向量。同样的，网络深度都较浅，有助于检测的提速。

以下是我实验中的结果



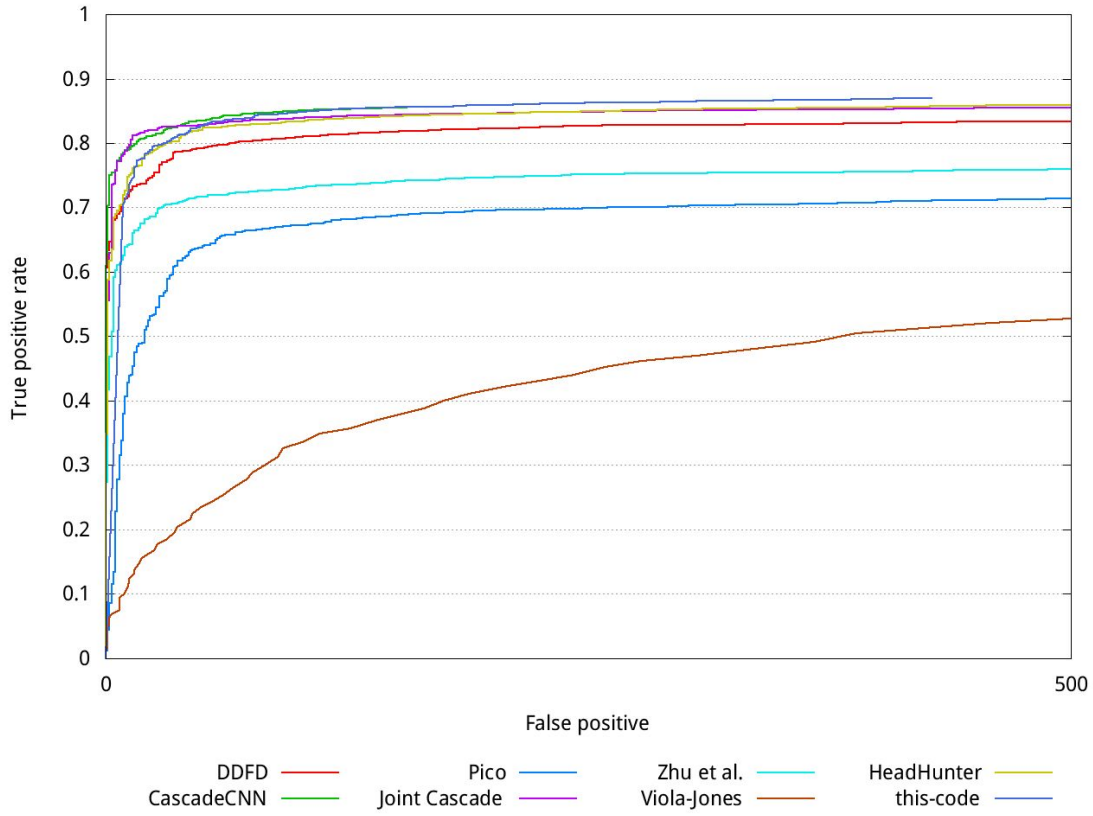
2.优化部分的描述：
a.heatmap(或者叫做全卷积)：
12net 中没有使用原论文中滑动窗口 stride2 的方案。而是利用了全卷积的性质，对于输入任意大小 w*h 的图片输出一个(w-11)*(h-11)的 heatmap。由于减少了冗余信息的输入，所以在速度上也有了一定的提高。

b.对于最后 48net 网络在结构上模仿了 cifar 网络而非原来的结构，为了得到更好的分类效

果。

五．实验结果：

由于 0~2000 的图无法看出细节差距。我这里放的是在 fddb 上 0~500 这个区间内的 roc 曲线。可以看到在某些错误率的范围内，本项目可以做到较好的结果（最上方蓝色的曲线）。但在错误少时准确率仍不足。



六．实验思考及实验细节补充：

1.实验速度分析：

在 fddb 测试中，我使用了 16 种尺度的输入（尺度因子是 1.18）以及对输入 padding 了 20 个像素（为了匹配有一部分在图片之外的脸），为了达到高的精度。而在现实场景下，可以使用 8 种尺度的输入（尺度因子是 1.41），来达到更快的效果。对于大图片来说目前速度还是在 3 到 4 秒一张图，matlab 中的 for 循环可能还是问题的所在。

2.关于 Casia 数据的使用

Casia 数据有 36 万张图片，先用传统方法 dlib 跑一边，得到的正样本，再用我的 48net 测试一下，将置信度较低（<0.93）的图片大约 2 万张，加入正样本。

3.关于负样本的选取

训练 24net 的数据是由 12net 扫描后得到的 postive false，训练 48net 的数据是由 24net 扫描后得到的 postive false。24net 可能得到的 postivefalse 较少，所以丧心病狂的从 imagenet 上搜集了建筑、植物、动物、汽车等等几十万张没有人脸数据集（后来发现还是有混入，再人工删除），以及原来在 AFLW 中产生的负样本混合在一起进行训练（AFLW 中的负样本

是按照与原图交并比小于 0.3 获得的，所以经常有脸的一部分，算是比较 hard 的负样本，对训练比较有效）。

4.在训练时的 trick

加入了 dither（0~0.1 区间内的随机白噪声），训练图片随机旋转一个角度，镜像，模糊等等方法来对人脸做数据增强。从经验主义的角度，旋转角度和镜像会比较实用一点。

5.实现的重点及调节参数

论文中给出了每一次经过网络后的 recall 值，除了最后一个网络我们也需要关心 roc 外，前几个网络的调试过程中，原论文给了我很大的帮助。耐心、细心是学到最多的东西。对于 nms 前两个网络不宜设置过大，用网络阈值来过滤一部分，而最后的结果，采用的是 nms 0.25 以过滤掉很多重叠部分。