人工智能应用实践

开题报告

--------人脸检测算法实践

计算机学院人工智能专业 17 级人工智能班

姓名：丁沐河 学号：201700301062

任课教师：吕琳　　助教：王业超

目录

前言

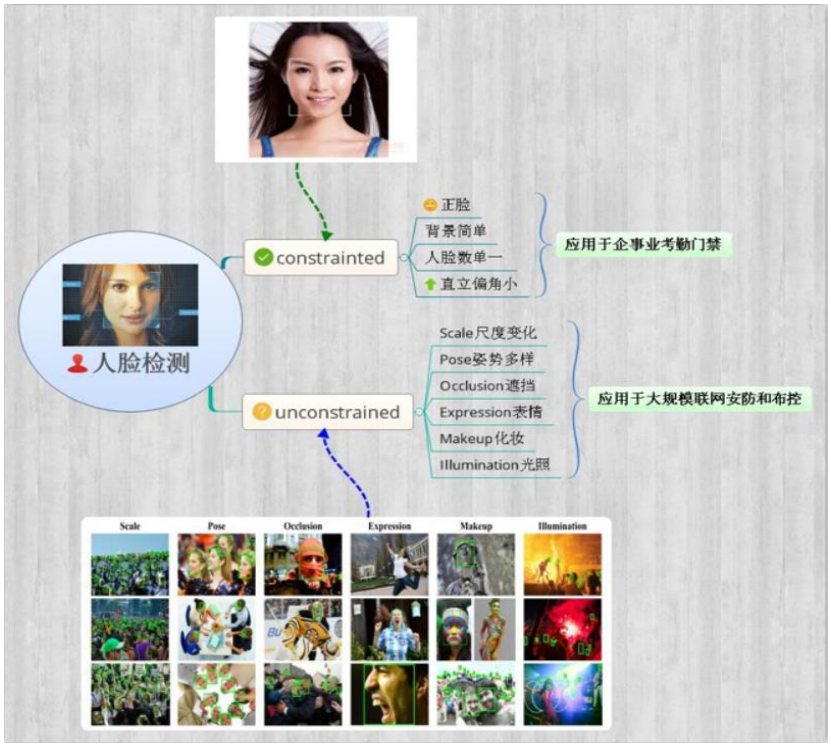
这是人工智能应用实践（关于人脸检测算法实践）的开题报告，报告的大体思路为：首先分析人脸检测的课题背景和人脸检测学术研究的发展现状，发现人脸检测应用市场和前景广阔，技术方面还有较大发展空间。其次介绍了本次课题的目的，重点是学习深度学习卷积神经网络在人脸检测中的应用并进行代码实现。随后将学习的几篇论文和卷积神经网络的理论进行阐述，准备之后对这些理论进行代码实现，训练数据和观察结果。最后，我介绍了人脸检测在表情识别方面的拓展，准备之后进行实现。

一、 课题背景

目前，人脸检测技术受到来自学术界和工业界越来越多的关注，究其原因，至少有三个方面的促进因素：**人机交互方式的演变**、**生物特征识别的发展**、**物体检测的研究**。首先，人脸检测技术的提出是人机交互研究发展的需要。其次，在生物特征识别技术中，作为人脸自动识别系统的先决条件，人脸检测技术有着十分重要的作用。最后，在理论上讲，人脸检测对于人脸识别，表情识别，目标检测和识别有着重要的研究意义。

而且，人脸检测技术是所有人脸影像分析衍生应用的基础，这些扩展应用细分有人脸识别、人脸验证、人脸跟踪、人脸属性识别，人脸行为分析、个人相册管理、机器人人机交互、社交平台的应用等。

从应用领域上可以分为：①以企事业单位管理及商业保密为主的商用人脸检测；②大规模联网布控的多角度多背景的安防人脸检测；③反恐安全、调查取证、刑事侦查为主的低分辨率尺度多样的军用/警用人脸检测；④当然还有基于互联网社交娱乐应用等的一般人脸检测。在学术研究中分为**约束环境人脸检测**和**非约束环境人脸检测**，如下图。



二、 发展现状

2.1 基于机器学习的算法

所谓人脸检测，就是给定任意一张图片，找到其中是否存在一个或多个人脸，并返回图片中每个人脸的位置和范围。人脸检测的研究在过去二十年例取得了巨大进步，特别是 Viola and Jones 提出了开创性算法，（**Viola-Jones 人脸检测器**）他们通过 Haar-Like 特征和 AdaBoost 去训练级联分类器获得实时效果很好的人脸检测器，然而研究指出当人脸在非约束环境下，该算法检测效果极差。这里说的非约束环境是对比于约束情况下人脸数单一、背景简单、直立正脸等相对理想的条件而言的，随着人脸识别、人脸跟踪等的大规模应用，人脸检测面临的要求越来越高（如上图）：人脸尺度多变、数量冗大、姿势多样包括俯拍人脸、戴帽子口罩等的遮挡、表情夸张、化妆伪装、光照条件恶劣、分辨率低甚至连肉眼都较难区分等。用经典 VJ 人脸检测器（2010 年更新）在非约束评测集 FDDB 中验证显示：当限定误检数为 10 个时，准确率不超过 10%；为 500 个时，检测率

仅仅为52.8%。所以丞待更好的算法以应用于大规模安防布控等非约束人脸检测场景。

2.2 基于深度学习的算法

14年底微软美国研究院首席研究员张正友等在CVIU上发表了非约束人脸检测专题综述，文中指出过去十年里，当限定误检数为0或不超过10个时，人脸检测算法的查准率也就是准确率（true positive rate）提高了65%之多（最新基于CNN的算法和传统Vj-boosting算法的对比结果）。

文中总结了现今出现的优异算法主要得益于以下四点：

①越来越多的鲁棒特征提取方法：LBP、SIFT、HOG、SURF、DAISY等；

②开放的数据库和评测平台：LFW、FDDB（报告中性能对比主要用的一个，更新于2016.4.15）、WIDER（汤晓欧团队发布的，更新于2016.4.17，不完整）；

③机器学习方法的发展和应用：boosting、SVM、深度学习等；

④高质量的开源视觉代码库的良好发展与维护：OpenCV、DPM、深度学习框架-caffe等。

人脸检测算法以往被分为**基于知识的、基于特征的、基于模板匹配的、基于外观**的四类方法。随着近些年DPM算法（可变部件模型）和深度学习CNN（卷积神经网络）的广泛运用，人脸检测所有算法可以总分为两类：①Based on rigid templates：代表有boosting+features和CNN ②Based on parts model：主要是DPM。

基于深度学习的人脸检测方法可以作为第一类方法的代表，同时也是检测某一种深度学习架构或新方法是否有效的评测标准。往往一个简单的卷积神经网络在人脸检测就能获得很好效果，同时有文献验证了深度卷积神经网络的第一层特征和SIFT类型特征极其相似。

DPM算法由Felzenszwalb于2008年提出的一种基于部件的检测方法，对目

标的形变具有很强的鲁棒性，目前已成为分类、分割、动态估计等算法的核心组成部分。应用 DPM 的算法采用了改进后的 Hog 特征、SVM 分类器和滑动窗口检测思想，在非约束人脸检测中取得极好效果。而其缺点主要是计算复杂度过高。

随着 DNN 的发展，基于深度学习的方法获得了 state of art 的效果，可见未来人脸检测算法主要的发展将围绕 DPM 和 DCNN 展开。同时将 DPM 和 DCNN 结合的方法也将是研究趋势。

三、课题目的

一、了解和学习基本的人脸检测算法，基于知识的、基于特征的、基于模板匹配的、基于外观的四类方法，学习计算机视觉里常用的进行人脸检测的几个特征，并且基于传统的人脸检测算法实现人脸检测的功能

二、了解最新前沿的人脸检测算法，实现在约束环境人脸检测和非约束环境人脸检测，并能实现人脸对齐，标记出人的眼睛、鼻子、嘴巴等特征点。

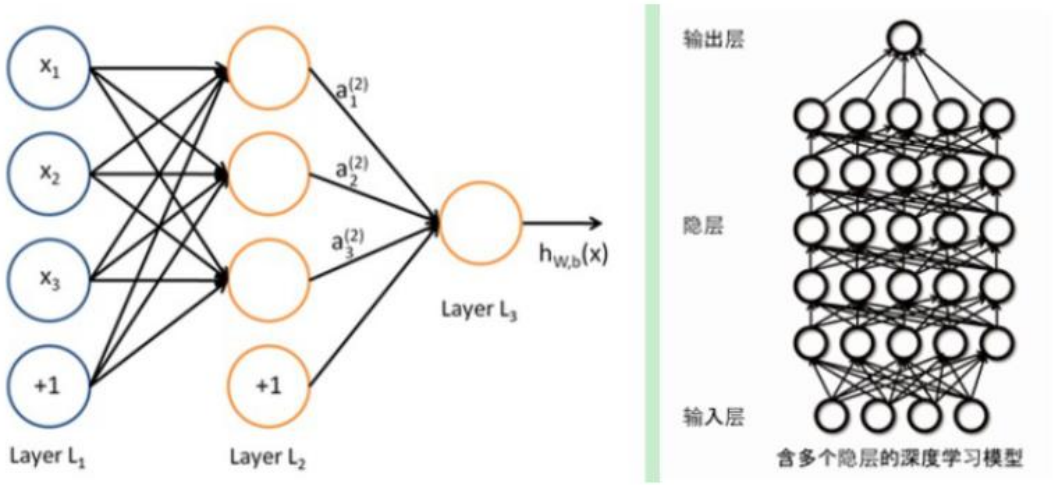
三、学习和了解卷积神经网络（CNN）和深度学习的知识，将深度学习，卷积神经网络（CNN）应用到人脸检测中，提高人脸检测的精度

四、学习三篇关于使用深度学习和卷积神经网络解决人脸检测的论文 ’Improved Selective Refinement Network for Face Detection’，’Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks’，’A unified embedding for face recognition and clustering’，实现代码来进行人脸检测，并对比评价人脸检测的精度

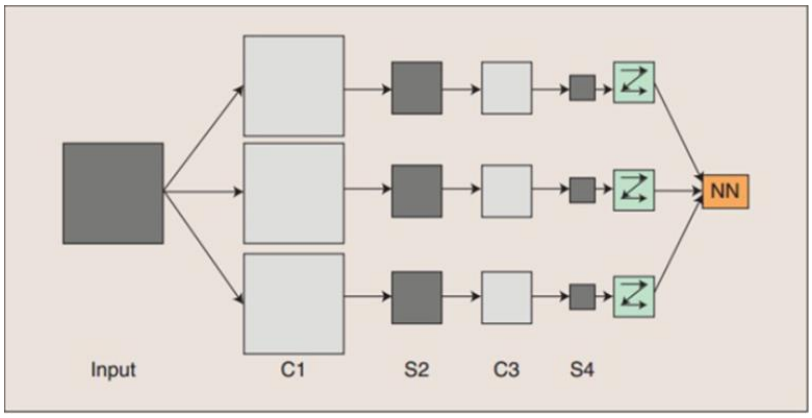
四、基于深度学习的算法理论

4.1 深度学习理论

深度学习采用了与传统神经网络相似的分层结构，系统由包括输入层、隐层（多层）、输出层组成的多层网络，只有相邻层节点之间有连接，同一层以及跨层节点之间相互无连接，每一层可以看作是一个 logistic regression 模型；这种分层结构，比较接近人类大脑结构。



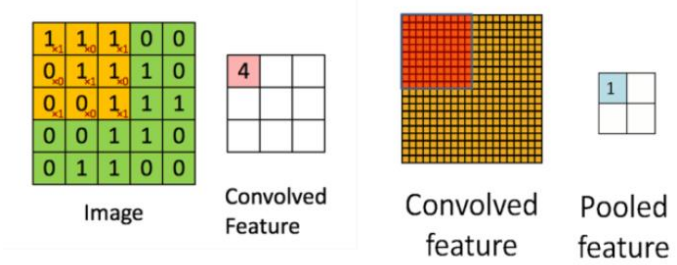
一个深度卷积神经网络通常包含输入层、多个卷积层（convolutional layer）、对应的降采样层（pooling layer）和归一化层。最后通过全连接层（fully connected layer）将二维的 feature maps 连接成一个向量输入到最后的分类器，得到概率（二分类时 0/1）输出，如下图。



卷积神经网络的概念示范：输入图像通过和三个可训练的滤波器和可加偏置进行卷积，滤波过程如上图，卷积后在 C1 层产生三个特征映射图，然后特征映射图中每组的四个像素再进行降采样求和（池化 pooling），加权值、偏置，通过一个 Sigmoid 函数得到三个 S2 层的特征映射图。这些映射图再进过滤波得到 C3 层。这个层级结构再和 S2 一样产生 S4。最终，这些像素值被光栅化，并连接成一个向量输入到传统的神经网络全连接层，分类得到输出。

卷积层（Convolution）：通过若干个滤波器与输入的二维特征平面（第一个卷积层是原始图片）进行卷积提取得到数据的显著性特征（如下左图），滤波器的大小（3X3,一般不大于 5X5）决定了提取到的特征对应的感知区域的大小，每一个特征都对应输入空间的一个小的感知区域，提取到的特征通过降采样等操作提高特征对输入样本微小畸变的鲁棒性。在卷积层中，卷积核是在整个输入平面上进行平移的（步长 stride=1），不同卷积层的卷积核提取不同尺度的特征，所以卷积神经网络提取的特征具有很高的平移不变性和尺度不变性，下图中，5X5 的输入 image 与 3X3 卷积核卷积得到 3X3 的 Convolved feature。

池化层（Pooling）：图像具有一种“静态”的属性，也就是意味着在一个区域内有用的特征极有可能在另一个区域适用。为了描述大的图像时，又可以做到降维，我们很自然的就对不同位置的特征进行聚合统计，旨在提高网络对输入样本微小形变的鲁棒性，从而增强网络的泛化能力，有以下三种：平均池化 Average pooling、最大池化 Max pooling、重叠池化 Overlapping pooling，一般 size=stride，如上右图，上一步得到的卷积特征图的左上角经过特定的运算，可以得到右边池化特征图的值，注意池化作用于图像不重叠区域，有别卷积操作。



4.2 数据集

- LFW

全名是 Labeled Faces in the Wild.这个数据集是人脸评估一定会用到的一个数据集，包含了来自 1680 的 13000 张人脸图，数据是从网上搜索来的。基本都是正脸。这个数据集也是最简单的，基本主流算法都能跑到 99%以上，貌似有 6 对 label 错了，所以最高正确率应该是 99.9%左右。这个都跑不到 99%的话别的数据集表现效果会更差。一般来说这个数据集是用来做人脸识别验证的。

- CelebFaces

总共包含 10177 个人的 202599 张图片，也是从搜索引擎上爬过来的,噪声不算多，适合作为训练集。同时这个数据对人脸有一些二元标签，比如是否微笑，是否戴帽子等。如果需要特定属性的人脸，也可以从中获取。

- CASIA-WebFace

该数据集是从 IMBb 网站上搜集来的，含 10K 个人的 500K 张图片。同时做了相似度聚类来去掉一部分噪声。CAISA-WebFace 的数据集源和 IMDb-Face 是一样的，不过因为数据清洗的原因，会比 IMDb-Face 少一些图片。噪声不算特别多，适合作为训练数据。

- VGG-Face

来自 2622 个人的 2 百万张图片。每个人大概要 2000+图片，跟 MS-Celeb-1M 有很多重叠的地方（因为都是从搜索引擎来的），这个数据集经常作为训练模型的数据，噪声比较小，相对来说能训练出比较好的结果。

4.3 深度学习框架（MTCCN）

Multi-task Cascaded Convolutional Networks（MTCCN）

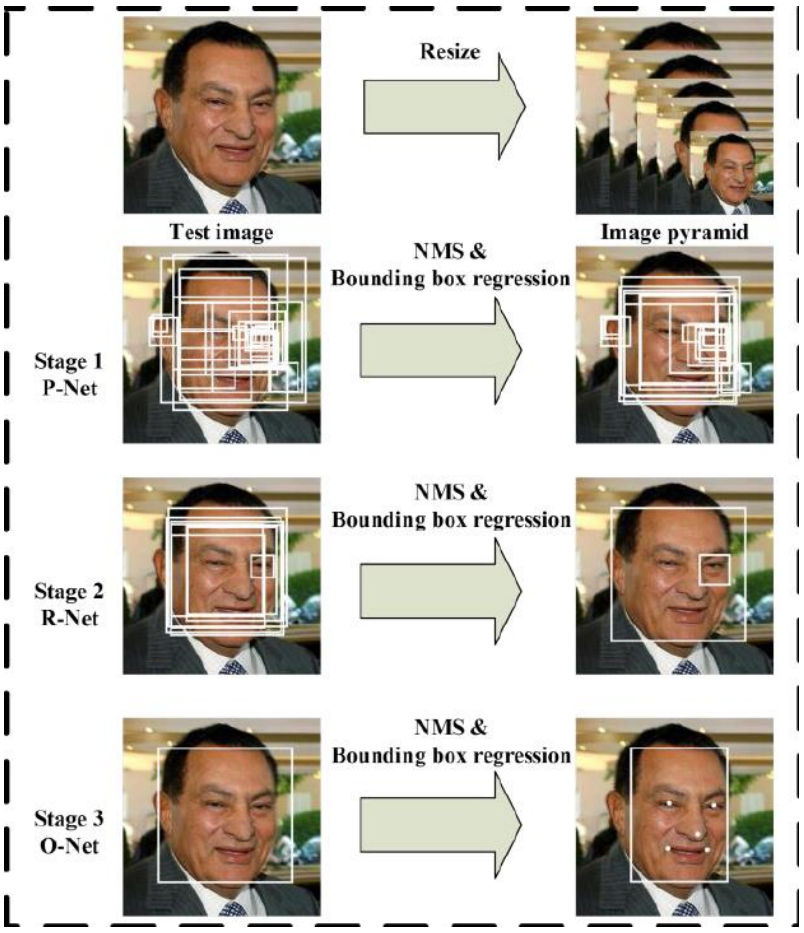
MTCNN 顾名思义是多任务的一个方法，它将人脸区域检测和人脸关键点检测放在了一起，同 Cascade CNN 一样也是基于 cascade 的框架，但是整体思路更加巧妙合理，MTCNN 总体来说分为三个部分：PNet、RNet 和 ONet，具体步骤为：

Stage 1 ： 我们利用全卷积网络，建立建议网络(P-Net)，获得候选窗口及其边界框回归向量。然后利用估计的边界盒回归向量对候选对象进行校正。然后，我们使用非最大抑制(NMS)来合并高度重叠的候选项

Stage2： 所有的候选人都被反馈给另一个 CNN，称为 Refine Network (R-Net)，它进一步拒绝大量的假候选项，使用包围盒回归进行校准，并合并 NMS 候选人

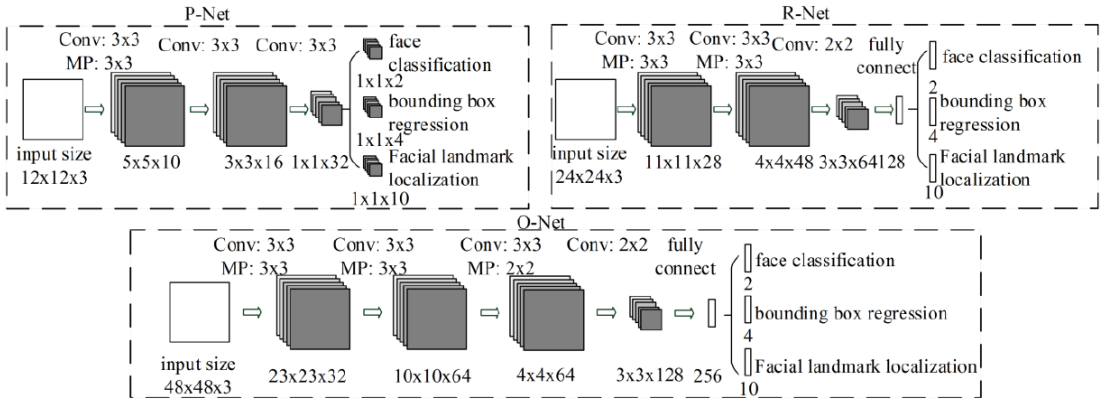
Stage3: 这一阶段与第二阶段相似，但在这一阶段，我们的目标是更详细地描述脸部。建立，输出网络(O-Net)产生最终的边界框和面标记位置地标的位置。该网络将输出 5 个面部地标的位置。

如下图所示：



MTCNN 在测试第一阶段的 PNet 是全卷积网络（FCN），全卷积网络的优点在于可以输入任意尺寸的图像，同时使用卷积运算代替了滑动窗口运算，大幅提高了效率。

除了增加人脸 5 个关键点的回归任务，另外在 calibration 阶段采用了直接回归真实位置坐标的偏移量的思路替代了 Cascade CNN 中的固定模式分类方式，整个思路更为合理。



MTCNN 的整体设计思路很好，将人脸检测和人脸对齐集成到了一个框架中实现，另外整体的复杂度得到了很好的控制，可以在中端手机上跑 20~30FPS。该方法目前在很多工业级场景中得到了应用。

4.4 FaceNet

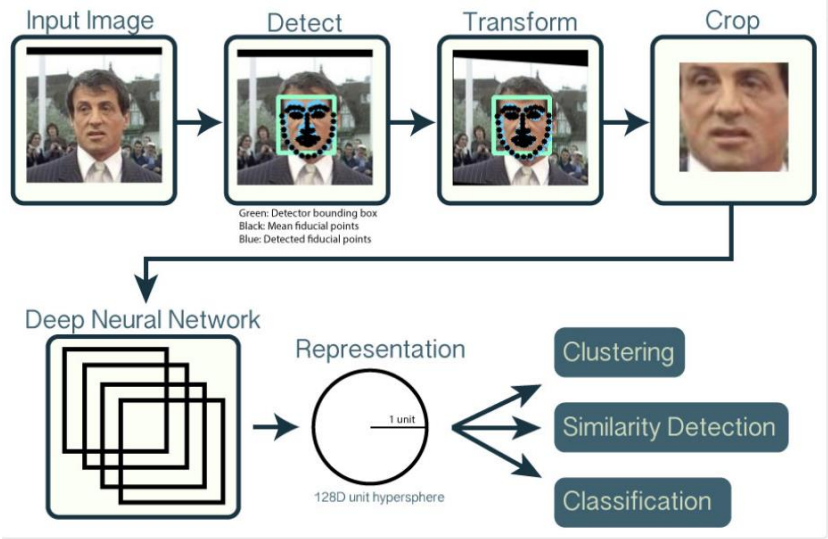
FaceNet: A Unified Embedding for Face Recognition and Clustering

基于论文 FaceNet: A Unified Embedding for Face Recognition and Clustering 建立了有名的 Openface 系统，这是一个基于深度神经网络的开源人脸识别系统。Openface 是卡内基梅隆大学的 Brandon Amos 主导的。

具体步骤：

1. 使用来自 dlib 或 OpenCV 的预训练模型检测人脸 。

2. 为神经网络变换人脸。该存储库使用 dlib 的 实时姿势估计 以及 OpenCV 的仿射变换 来尝试使眼睛和下唇出现在每个图像的相同位置。
3. 使用深度神经网络在 128 维单位超球面上表示（或嵌入）面部。嵌入是任何人脸部的通用表示。与其他人脸表示不同，此嵌入具有很好的特性，即两个人脸嵌入之间的距离较大意味着这些人可能不是同一个人。此属性使聚类，相似性检测和分类任务比其他人脸识别技术更容易，在其他人脸识别技术中，特征之间的欧几里得距离没有意义。

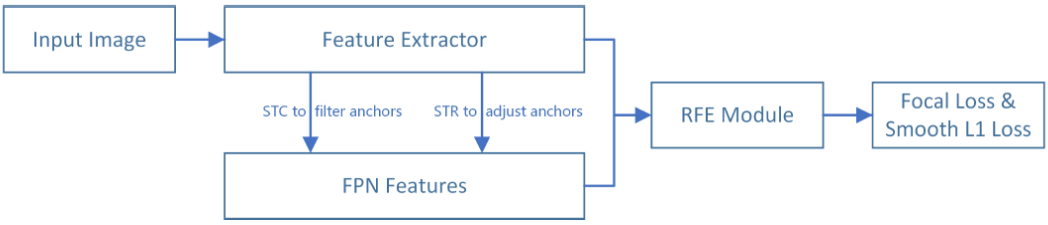


4.5 SRN

Improved Selective Refinement Network for Face Detection

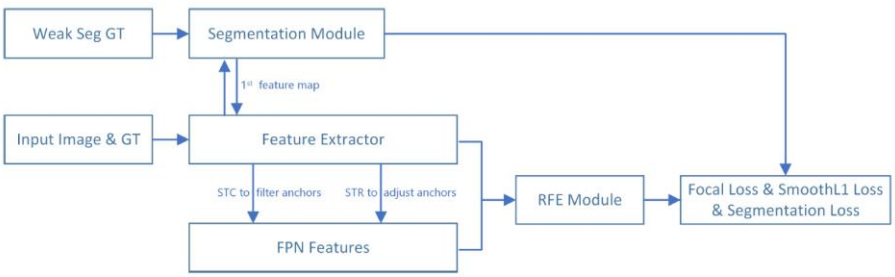
本篇文章，提出了一种选择性细化网络（SRN）人脸检测器，选择性细化网络（SRN）人脸检测器选择性地将两步分类和回归操作引入到基于 anchor 的人脸检测器中，以同时减少误报并提高定位精度。此外，它设计了一个感受野增强模块，以提供更多样化的感受野。还有一些为了进一步提高 SRN 的性能的一些现有技术，包括新的数据增强策略，改进的骨干网络，MS COCO 预训练，解耦分类模块，分段分支和 Squeeze-Excitation 块。其中一些技术可以带来性能改进，

而其中很少一些技术不能很好地适应我们的基线。通过将这些有用的技术结合在一起，提出了一种改进的 SRN 人脸检测器，并在广泛使用的人脸检测基准 WIDER FACE 数据集上获得了最佳性能。



它包括选择性两步分类(STC)和选择性两步分类回归(STR)和感受野增强(RFE)。

这篇文章主要解决图像中存在很多小的人脸（many tiny faces）情况。文章提出的 Selective Refinement Network (SRN)是一个 single-shot face detector，提出的结构主要减少 false positive。整体网络由两个部分组成：the Selective Two-step Classification (STC) module and the Selective Two-step Regression(STR) module. STC 旨在从低级检测层中过滤掉大多数简单的负 Anchor，以减少后续分类器的搜索空间，而 STR 则设计为从高级检测层粗略调整 Anchor 的位置和大小，以便为随后的回归量。此外，我们设计了一个感受野增强（RFE）模块，以提供更多样化的感受野，这有助于更好地捕捉某些极端姿势的面部。



显著解决两个问题：1. 减少 False positive 2. 捕获某些极端姿势的面部

•我们提出了一个 STC（选择性两步分类）模块，用于过滤掉来自低层的大多数简单负样本，以减少分类搜索空间。

•我们设计了一个 STR（选择性两步分类回归）模块，可以从高级层粗略调整锚点的位置和大小，为后续的回归量提供更好的初始化。

•我们引入 RFE（感受野增强）模块，为检测极端姿势面提供更多样化的感受野增强。

五、人脸检测扩展——表情识别

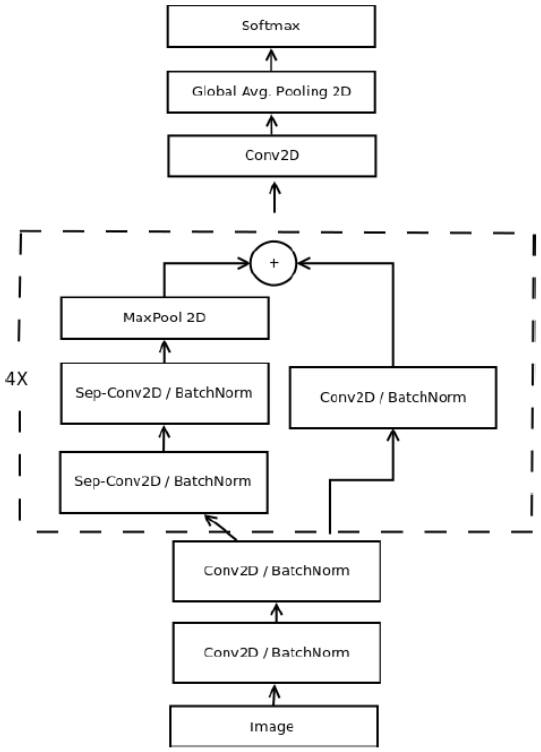
我们在实现人脸检测的算法之后，便可以开始应用扩展了，其中一个特别吸引人的应用扩展就是——表情识别。我们阅读论文，决定选用 Real-time Convolutional Neural Networks for Emotion and Gender Classification 的代码来进行实现应用

5.1 训练集：

训练数据集使用的是 2013 年 Kaggle 比赛的数据——FER2013，该数据集含 28709 张训练样本，3859 张验证数据集和 3859 张测试样本，共 35887 张包含生气、厌恶、恐惧、高兴、悲伤、惊讶和正常七种类别的图像，图像分辨率为 48×48。

5.2 实现方法

阅读 Real-time Convolutional Neural Networks for Emotion and Gender Classification 论文，使用论文中提出的这里用到的是 CNN 的主流框架之 mini_XCEPTION。XCEPTION 是 Google 继 Inception 后提出的对 Inception v3 的另一种改进，主要是采用深度可分离的卷积（depthwise separable convolution）来替换原来 Inception v3 中的卷积操作。具体方式如下：



附. 参考文献

[1] Zhang S , Zhu R , Wang X , et al. Improved Selective Refinement Network for Face Detection[J]. AAAI2019.

[2] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016.

[3] Schroff F , Kalenichenko D , Philbin J . FaceNet: A unified embedding for face recognition and clustering[J]. CVPR, 2015

[4] Octavio Arriaga, Matias Valdenegrotoro, Paul G Ploger Real-time Convolutional Neural Networks for Emotion and Gender Classification. arXiv: Computer Vision and Pattern Recognition .2017