学校代码: 10246
学　　号: 15210180026

復旦大學

硕 士 学 位 论 文

Faster-RCNN网络对人脸年龄及性别的多标签预测

Multi-label prediction of face age and gender basing on Faster-RCNN network

院　　系: 数学科学学院
专　　业: 应用数学
姓　　名: 梁欣
指 导 老 师: 卢文联　教授
完 成 日 期: 2018年3月30日

目录

摘 要

年龄和性别是人的两个重要特征。性别可以视为一个二分类问题，但对于年龄这个特征来说，随着人年龄增长，脸部的皮肤纹理等特征会不断变化。而脸部的这些特征参数也可以反映出当事人的年龄。传统的分类器先基于人工编写的特征对人脸进行检测，从而预测年龄，性别等问题。但随着计算机硬件性能的提升，以及深度学习的快速发展，神经网络在计算机视觉(CV)，自然语言处理(NLP)等领域有了非常突出的科研成果。本文主要讲述了深度学习在人脸实际年龄和性别预测问题中的应用。利用Faster-RCNN的检测模型，对图片中的人脸进行检测，同时对人脸的实际年龄段和性别做出分类预测。其中检测模型使用VGG16作为图片特征提取的网络。该网络已在ImageNet图片集上完成预训练，使得网络能够更好地抓取图片较为粗糙的特征。另选取了从IMDB-WIKI数据库抽取出的43000余张图片作为模型的训练集，这是目前已知的最大的同时拥有年龄和性别标签的数据集。

关键字: 深度学习，人脸检测，性别识别，年龄预测，多标签学习.

Abstract

Age and gender are two basic and significant characteristics of people. Gender recognition can be regarded as a two-category problem. However, as the person ages, face features such as skin texture will constantly change. And these characteristic parameters of the face also indicate the age of the people. Traditional classifiers extract facial features based on hand-made filters to predict age and gender. With the improvement of hardware computing performance and the development of deep learning, neural networks have made outstanding scientific research achievements in areas such as computer vision and natural language processing. This paper mainly discusses the application of deep learning in age estimation and gender recognition. Basing the Faster-RCNN detection architecture, our model detects the face in the picture, at the same time, estimates the age and recognize the gender of the people. Aiming to capture features better, we uses VGG16, which has been pre-trained on the ImageNet dataset, as the network for extracting features. In addition, we selecte more than 43,000 images from the IMDB-WIKI database as the training set of the model.

Keywords: Deep Learning, Facial Detection, Gender Recognition, Age Estimation, Multi-labels.

第一章　相关研究

本章将对常用的图像特征描述子以及人脸检测中的两个经典网络，包括Viola-Jones物体检测框架和DPM模型，做相关介绍。

第 1 节　图像特征描述子

一般地，在计算机视觉和图像处理中，检测模型有几个常用的图像特征，包括harr特征，方向梯度直方图、SIFT特征以及它们相关的一些变体。

1.1. harr-like 特征

harr-like 是由Papageorgiou C. 基于人脸的几个共性而提出的一种脸部特征的描述子。这些特征包括:

1.眼睛附近区域的颜色要比额头区域深

2.鼻子附近区域的颜色要比眼睛区域浅

其中所定义的颜色是指灰度值的高低。根据这些人脸特性，提出了由图中的四种矩形框定义的特征，即最初的harr特征，也称为矩形特征。其中黑色区域对应人脸部颜色较深的部位，白色区域对应人脸部颜色较浅的部位。通过计算两个部分覆盖的灰度差，得到相应的矩形特征。具体运算如下：

$$f = \Sigma_{(i,j)\in \mathrm{B}} p\,(i,j) - \Sigma_{(i,j)\in \mathrm{W}} p(i,j)$$

其中，$p(i,j)$代表图像中点(i,j)的像素值，B和W分别代表矩形框的黑色区域和白色区域。除了Papageorgiou C.最开始提出的四种harr特征，后来还出现了一些变体，称它们为类harr特征。根据矩形中黑色区域和白色区域的数目，对应称之为两矩形特性，三矩形特征等。如图1.1。

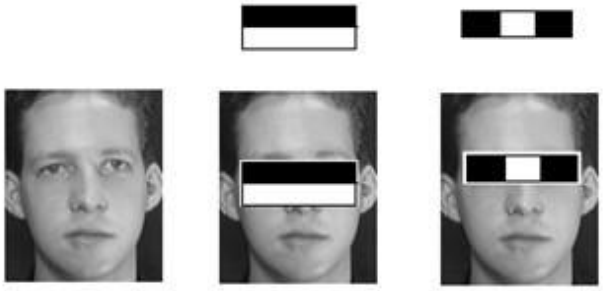


图 1.1: harr-like 特征

1.2. 方向梯度直方图

图像中局部物体的表象和形状可以由局部梯度分布或边缘方向密度描述，因此在做目标检测任务时，常常选用方向梯度直方图(HOG)作为检测系统所用的图像特征。

方向梯度直方图的具体步骤如下:

1.梯度计算

对像素点的梯度计算，最常用的方法是分别在水平和垂直两个方向上，对相邻像素点的像素值做差分。即:

$$
\begin{aligned}
G_i(i,j) &= I(i+1,j) - I(i-1,j) \\
G_j(i,j) &= I(i,j+1) - I(i,j-1) \\
G(i,j) &= \sqrt{G_i(i,j)^2 + G_j(i,j)^2} \\
\alpha(i,j) &= tan^{-1}\left(\frac{G_j(i,j)}{G_i(i,j)}\right)
\end{aligned}
$$

其中$I(i,j)$是点(i,j)的像素值。$G_i(i,j)$和$G_j(i,j)$分别表示图像在像素点(i,j)水平方向和垂直方向的梯度，$G(i,j)$和$\alpha(i,j)$为像素点在(i,j)的梯度幅值和梯度方向。

2.定向直方图

根据第一步的计算，得到细胞单元内的每一个点的梯度值大小$G(i,j)$和梯度方向$\alpha(i,j)$。为计算出细胞单元内的定向直方图，单元内的像素点需要对某个方向的直方图通道进行加权投票。权重由像素点梯度值的函数计算。直方图通道平均分布在$0° \sim 180°$(无向)或是$0° \sim 360°$(有向)范围内。通常选用无向的，均分为9个的梯度直

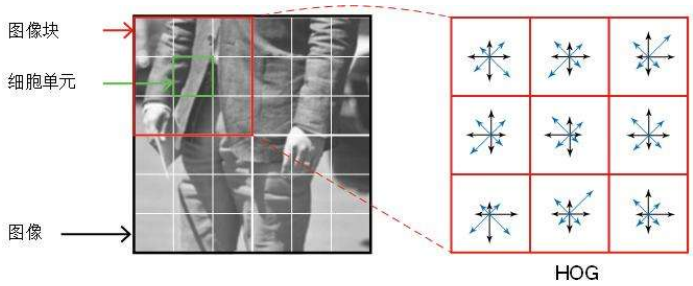


图 1.2: HOG 特征

方图通道，即0°，20°，40°,,,180°这9个方向分布。

3.区块直方图及归一化

考虑到图像中的每点光照或对比度的影响，需要对局部的梯度强度归一化。合并邻近的细胞单元，组成空间上较大的连接块，计算这些较大区域的梯度图。常用的区块形状的选择有矩形，圆形和中心环绕三种。若取区块为矩形形状，且该区块合并了4个细胞单元。级联这4个单元内的9向直方图，可得到该区块的4×9=36维的特征，再对区块级的方向梯度做归一化。由于在将细胞单元连接成区块的过程中，某些细胞单元同时出现在不同区块的计算中，这也就意味着，这些细胞单元的输出多次作用在最终的描述子上。

4.HOG特征

级联图像中所有区块的直方图。如图像划分为100个矩形区块(区块间可能有细胞单元的重叠)，每个区块包含上述的4个细胞单元，级联所有区块的特征，可得到$100\times4\times9$=3600维特征。

图1.2为定向梯度直方图计算过程。

从方向梯度的计算过程可知，由于图像几何形变及光学的形变只会出现在较大的空间尺度上，而方向梯度直方图是在局部的细胞单元内计算的，所以它对这两种形变都能保持很好的不变性。

1.3. SIFT

David Lowe在1999年提出了尺度不变特征变换(Scale-invariant feature transform, SIFT)的方法来处理图像识别的问题。顾名思义，该算法最突出的特性就是尺度不变

性，并且不会受到图像旋转的影响。此外，该方法对于光线变化、加噪声、仿射变化等处理后的图片依旧表现出良好的鲁棒性。由于这些良好的性质，自提出后便得到广泛使用。其大致算法构造如下：

1、极值检测

我们需要引入高斯核函数对原图像进行光滑处理。由于处理的都是图片信息，这里，我们只需要用到两维高斯核函数：

$$G_\sigma(x,y)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

通过卷积运算$L(x,y,\sigma)=G_\sigma(x,y)*f(x,y)$得到对于特定的尺度$\sigma$进行磨光后的图像。

接下来，需要建立图片的尺度空间。为此，通过对图片进行降采样结合不同尺度高斯滤波的方式，得到不同尺度图片构成的图片金字塔。在检测图片极值点之前，还需要对金字塔中的图片使用高斯差分(DoG) 方法进行处理。这里DoG方法即是指将不同尺度的滤波后图像做差，由此得到了高斯差分图像，同样也是一个金字塔的形式：

$$D(x,y,k\sigma)=L(x,y,k\sigma)-L(x,y,\sigma).$$

而我们需要的极值点则是通过将像素点的值和相邻像素点(这里相邻既包括同DoG图层，也包括相邻DoG图层)做比较，选取其中的极大极小值点作为备选特征点，这个过程可以排除掉图像中绝大部分的点。

2、关键点定位

由于采用降采样的方法改变了图片尺寸，并且尺度σ的变化也并非连续。所以为了获取精确的位置信息，在得到候选特征点后需要进行空间曲面拟合来校正得到的位置和尺度信息。在这个过程中，也会对低对比度和部分边缘点进行过滤。

3、确定特征点方向

SIFT方法一个重要性质就是旋转不变，为达到这个目的，需要为特征点分配一个或多个方向。为此，对于每个特征点，我们需要在其所处的尺度σ内计算该点对应的梯度，其模值和方向计算如下：

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2},$$

$$\theta(x,y)=tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))).$$

将360度平均划分为36个方向范围，统计特征点附邻近区域(区域大小的选取依赖于尺度σ)内每个像素点对应的梯度。将其分配到对应的方向范围，并统计邻域内各个

方向范围出现次数，绘成直方图，将出现次数最多的方向最为该特征点的主方向。此外，为了增加算法的鲁棒性，我们也保留直方图中超过80% 的方向作为辅助方向。

4、生成特征描述子

最后，依据前面得到的特征点方向旋转坐标轴，将特征点附近16×16的邻域划分成4×4的子区域，统计子区域内像素点的梯度方向信息。通过高斯加权处理生成一个8个方向的梯度直方图，由此统计得到的$4 \times 4 \times 8 = 128$ 维向量经过归一化处理后即为该特征的描述子。

第 2 节 人脸检测

一般地，人脸识别任务可以分为以下四个内容。(一)人脸检测识别图像中人脸的位置;(二)人脸校准，人脸的特征点定位，这些特征点包含但不限于眼角，眉毛，瞳孔，嘴巴，鼻子等等人脸共用的特征属性;(三)人脸校验，验证不同图片中出现的是否为同一人的脸;(四)人脸鉴别，根据图片中的人脸，找出数据库中对应的信息。

下述的两个经典的检测网络，主要是建立在第一个类别——人脸检测问题的基础上，利用相关的图像特征构造并训练探测器。在测试阶段，模型通过计算图片的特征，分类器根据这些特征判断图中是否有人脸。

2.1. Viola-Jones目标检测框架

人脸检测在计算机视觉领域是一个比较成熟的课题。Viola P和Jones M.于2001年，在[1]一文中提出的人脸检测模型，是一个较经典的算法。该篇文章主要提出以下三个创新性的研究成果，很大程度地提高了人脸检测效率。

1.用积分图进行图像表征

文章提出一种新的图像表征方式——积分图，通过计算图像的积分图，探测器能很快地对特征进行评估。和之前很多探测器采用的特征一样，Viola-Jones网络采用harr-like特征。根据前面的描述，harr-like特征是为符合人脸的某些区域比另一些区域或深或浅的特性而设计。但在拍摄图片时，人站立的位置，或是焦距的选择，呈现出的人脸占据整个照片的比例也会不同。为不漏掉人脸特征，需选取不同尺寸的矩形特征，这增加了很大的计算量。虽然harr-like特征运算简单，但是其庞大的数目，给检测网络带来了一定的效率限制。为了提高多尺度下harr-like特征的计算效率，文章提出采用一个中间表示方法来计算图像的矩形特征，即积分图像的概念。

事实上，可以将积分图像视为一个双变量函数，两个变量为图像的位置点x, y。积分图像$I(x, y)$表示，包含点x, y上方及其左侧的所有像素点的像素和，即：

$$I(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

式子中$I(x, y)$是积分图像，$i(x, y)$是原始图像在点(x, y)的像素值，由积分图的定义可知，：

$$I(x, y) = i(x, y) + I(x, y-1) + I(x-1, y) - I(x-1, y-1)$$

当积分图像被计算出来后，计算任何矩形区域的强度总和只需要四个数值，而和该区域的面积无关。举例来说，$A = (x_0, y_0), B = (x_1, y_0), C = (x_0, y_1), D = (x_1, y_1)$,由这四个点构成的矩形区域的像素和$i(x, y)$为：

$$\sum_{\substack{x_0 < x \leq x_1 \\ y_0 < y \leq y_1}} i(x, y) = I(D) + I(A) - I(B) - I(C)$$

这样仅需要较少的计算量，积分图像就可以通过已知原始图像而一步求得。当计算得到了整幅图的积分图像后，可以很快速的计算出任意的类harr特征，加快了模型的检测速率。

2.增强学习算法

积分图加快了特征的计算速度，但是由于这些矩形特征庞大的数量，在测试图像时若对它们全部进行评估，计算代价较昂贵。文章提出了一个构造分类器的方法，通过Adaboost的变体挑选出少量但最佳的分类特征。在任意图像的子窗口里，类harr特征的数目非常多，远高于图像像素的数目。为了让分类器能快速的判断类别，大部分类harr特征需要被丢弃掉。分类器只需关注那些数目较少的关键特征即可。特征筛选的机制通过对Adaboost程式的简单修改完成：限制比较弱的分类器，使它们仅和其中一个特征相关。这样每次提升的过程，即每次选出一个弱分类器的过程，可以看成一个特征筛选的过程。

3.级联形式的分类器

当图像被送进检测器时，会通过滑窗搜索的得到很多子图像。原检测问题被分解为数目众多的子图像的人脸检测，使得检测任务效率下降。为提高效率，文章提出用级联形式不断地合并分类器，使得图像中的背景部分可以快速地被分类器筛选出。则计算资源被分配到了疑似检测目标的区域上。

尽管Viola-Jones目标检测框架取得了当时较高的检测率，但从其原理可知，该探测器只能对在正常环境拍摄、且为正面的人脸图像才能表现良好。因为级联训练中简单的harr特征不足以捕捉到更复杂的人脸参数。当图像的拍摄环境不是正常状态（比如拍摄时的光照，被拍摄者的表情，人脸被环境所遮挡），以及人脸非正面朝向镜头时，探测器对人脸的检测就差强人意了。

2.2. DPM

2011年，Felzenszwalb 等人在[20]中提出的基于混合多尺度可变形部件模型的目标检测系统，目前已经成为很多分类问题，图像语义分割，行为分类和人体姿态等问题的重要部分。

文中提出了星型模型的混合检测系统，由一个基本上覆盖了全部目标的比较粗糙的根滤波器，和几个覆盖目标中较小部分的部件滤波器构成。其中，部件滤波器使用的特征分辨率是根滤波器的两倍，这样可以捕捉到相对于根滤波器更为精确定位的特征。为了获得不同位置与不同尺度下的特征，文章引入了特征金字塔的概念。特征金字塔由几个不同尺度构成的特征映射组成。首先通过不断的平滑和子采样计算一个标准的图像金字塔，然后计算金字塔中每层图像的特征。

DPM使用的图像特征可以看做是方向梯度直方图的扩展，大体思路与之一致。在得到直方图特征后，用隐藏变量支持向量机(LSVM)训练得到物体的梯度模型。得到模板后，就可以将模型和目标进行匹配了。

由于匹配过程中，会得到目标实例的多个重叠的检测结果。文章采用一个贪心的非极大值抑制(Non-maximum Suppression)程序来消除重复检测。在对图像作用算法后，会得到图像对应某个类别的一个检测结果集合S。S中每个检测结果对应一个预测框和一个得分。按得分对S中的检测结果排序，贪心地选择具有最高分数的结果，并跳过与之前所选检测结果的预测框覆盖面积超过50%的检测结果。

这种方法结合了训练时对齐和聚类的潜在变量估计，使用多个组件和可变形部分来处理类别间的差异。且由于其稳健的工程设计，使得模型更具鲁棒性也足够快。因此用监督部分位置训练的树型DPM模型，已经成功地应用于人脸检测和基准点估计问题上。

第二章　目标检测算法

第 1 节　卷积神经网络

2012前，目标检测，例如前面所述的Viola-Jones目标检测框架及DPM模型，采用的特征大多是人工选定的。测试阶段，通过计算这些图像特征，最后将特征向量送入训练好的支持向量机(SVM)或多层感知器(MLP)进行分类。但随着人们对数据的重视，越来越多的标注数据集被建立和维护。这些大型的标注数据给神经网络的发展提供巨大的发展动力。Alex 在2012年[4]中，用神经网络的模型AlexNet夺得了ILSVRC 2012的冠军。在百万量级的ImageNet数据集上，效果大幅度超过以往传统的方法，top-1和top-5的错误率分别降到了37.5%和17.0%。Alex使用的卷积神经网络，在计算机视觉领域作为特征提取强有力的工具，推动了深度学习在包括计算机视觉的许多的领域的研究热潮。

卷积网络最初因为生物学的启发而被提出。19世纪60年代，科学家通过对猫的视觉皮层细胞研究发现，每一个视觉神经元只会处理一小块区域的视觉图像，这块区域我们称之为感受野。当刺激在感受野的范围内出现时，视觉皮层的神经元才会对它作出反应。由于不同神经元的感受野会有部分的重叠，所有神经元的感受野会覆盖整个视野，这就实现了对物体从局部到整体的认知。

在深度学习领域，卷积神经网络(CNNs)指的是，包含多个卷积层的一类前馈式人工神经网络。和具有类似大小网络层、别的标准前馈神经网络相比较，卷积神经网络因为共享权重，神经元之间有更少的连接，这使得它们更容易训练。对于传统的图像分类算法，为了更好的获取图像特征，一般说来，会有很多预处理的过程，例如图像色彩和Gamma值归一化等等。而卷积神经网络则使用了较少的预处理步骤。传统算法中，特征提取网络所学习的滤波器是通过人工设计的，这往往需要很多关于图像的知识储备。较之这些传统分类器神经网络不需要先验知识，也不需要人工来设计这些特征滤波器，成为其最大的优势。

下面的内容将借助AlexNet的网络框架，介绍CNN的一些基本结构，及相关网络的发展。

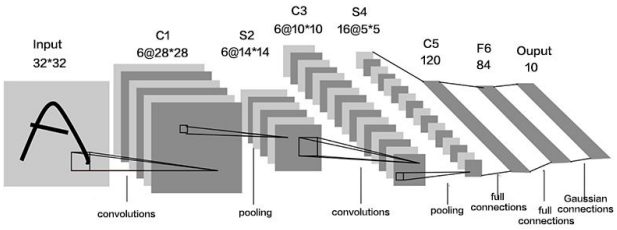


图 2.1: AlexNet网络结构图

如图2.1所示的就是AlexNet的一个网络结构，该网络结构由以下几个部分组成：卷积层，池化层，激活层，全连接层，softmax分类层。接下来将简单介绍一下各网络层的构成以及它们的一些变体和发展。

1.卷积层

卷积核是卷积神经网络(CNNs)的重要部分，同时这些卷积核也是卷积层所需要学习的参数，我们观察一幅图像时，每一个视觉神经元只能感受到一小块区域的视觉图像(感受野)，由于神经元的感受野会有重叠，所有视觉神经元的感受野就覆盖了全部视野，我们也就获取了整张图像的信息。受感受野的启发，我们利用卷积核来实现感受野的概念，通过对整幅图像的卷积操作，模拟视觉神经元对图像的认知过程。

卷积操作有以下几个概念：卷积核，步长，填充。

卷积核，即特征滤波器，不同于传统图像分类网络，它不是人工设计，而是整个网络训练所学习的参数。假设输入一张图像，将其看成一个矩阵I，定义一个3×3的卷积核K，那么卷积核作用在图像I上面可得到一个新的矩阵，称之为特征图。卷积过程如图2.2。

若卷积核每次间隔一个单位进行下一次卷积操作，则称此卷积步长为1，以此类推，可以根据需求取步长为2，为3，甚至更大。卷积层还有一个填充的概念，为了获取特定大小的特征图，会对原图像或中间特征图进行边缘的填充，即将边缘用0值补齐到某个尺寸。

卷积核的大小即代表了感受野的大小。网络越靠近输入层，卷积核尺寸越大，即感受野越大，网络学习到较为低级的、更为普适的图像特征。越靠近输出层，卷积核尺寸越小，感受野越小，学习到的就是比较精细的特征，也就越有利于对图像的分类。然而卷积核越大，所带来的计算量也就越大。尽管随着各种硬件设备的发展，计算资源限制的问题得到很大的改善，但如何将网络高效的利用这些资源，让卷积神经网络

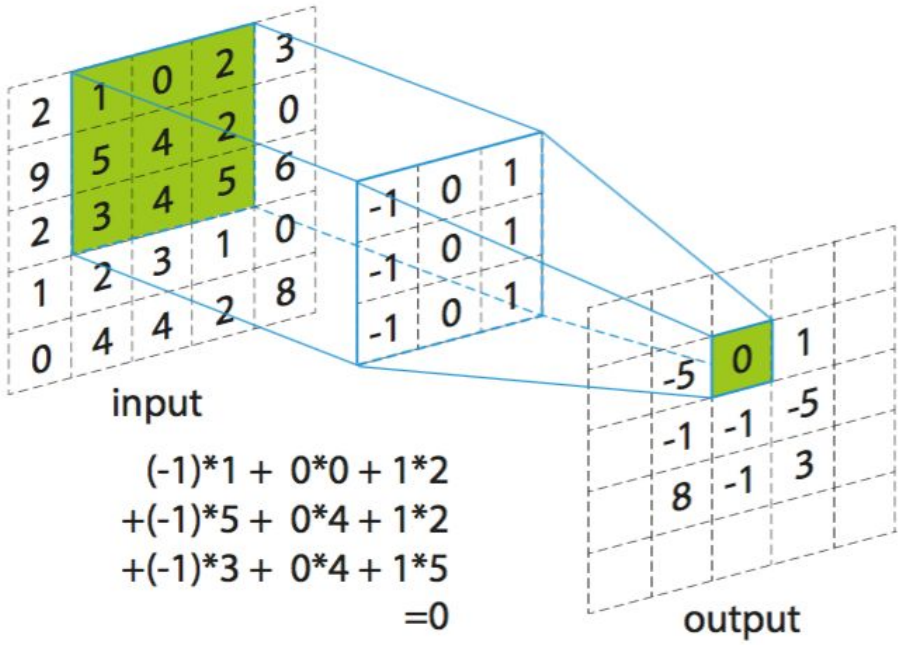


图 2.2: 卷积过程

能够更深，成为了后来学界和业界一直努力的方向。

为了提高神经网络的表达能力，2013年的[5]提出了微型网络的结构，这个结构换个角度来看，就是引入了1×1的卷积结构。受该文章的启发，Google于第二年在[6]中提出了Inception的微型网络的结构，并通过搭建这些Inception，得到了一个深度为22层的神经网络GooLeNet。在这个Inception的微型网络中，作者丢弃了较大尺寸的卷积核，仅仅使用了5×5,3×3,以及最关键的1×1。这个1×1的卷积核最主要的目的是，通过控制通道的个数降低维度。降维解决了计算资源紧张的问题，这也是限制整个网络大小最大的障碍。随着资源的释放，就可以增加网络的深度，同时不会影响网络的宽度。

2.池化层

池化是存储技术中一种常用手段，借助池化，网络存储可以有效提升存储的利用率。在卷积神经网络中，池化是一种非线性实现降采样目的的方法。该网络层的提出源于这样一个直觉，即一个特征的确切位置并不比其他特征的粗略位置重要。常见的池化有两种，一种是最大化池化，一种是平均池化。卷积神经网络中池化层的输入是卷积层输出的特征图，特征图经过池化层，将会被划分为一组互不重叠的矩形框，这些矩形框对应的子区域就是一组组邻近神经元组，输出是这些神经元组的最大值(最大化池化)或是均值(平均池化)。

在AlexNet之前，特征图上被池化的邻近神经元组两两之间没有重叠覆盖。准确说来，一个池化层可以视为由间隔s个像素的池化单元网格组成，每个聚合以池化单元的位置为中心的大小为$z\times z$的领域。若令$s=z$，则这就是没有覆盖的池化。若令$s<z$，就是AlexNet中提出的有覆盖的池化。本质上，池化过程可以看成卷积操作，对应的s为步长，$z\times z$为卷积核大小。根据比较有覆盖和没有覆盖的池化层的试验结

果，发现有覆盖的池化的模型会更不容易过拟合。

在卷积神经网络结构中，池化层可以逐渐的较少特征表征的大小，减少参数的数目和整个网络的计算，也因此可以适当的抑制过拟合。

3.激活函数层

神经元有两种状态：兴奋状态和抑制状态。一般情况下，大多数的神经元处于抑制状态，一旦神经元受到刺激，导致它的电位超过一个阈值，那么该神经元就会被激活，转变成兴奋状态，再接着向别的神经元继续传递。神经网络添加激活函数层来模拟神经元状态的改变。传统神经网络中最常用的两个激活函数，同属Sigmoid系的Tanh-Sigmoid函数

$$f(x) = tanh(x)$$

以及Logistic-Sigmoid函数

$$f(x) = (1 + \mathrm{e}^{-x})^{-1}$$

2001年，神经学科学家从生物学角度，模拟出了脑神经元接受信号更精确的激活模型。而这个模型恰好与2001年提出的ReLU函数很是相似，该函数表达式为：

$$f(x) = max(0, x)$$

在真实数据上实验发现，在梯度下降的训练过程中，传统的激活函数和ReLU相比，后者会下降的更快。

4.全连接层

图像经过卷积层，池化层，以及激活函数层的作用后，捕获到图像的特征。然后神经网络最高层的推理将通过全连接层完成。全连接层的每一个神经元都将和前面一层的所有神经元连接。所以全连接层的输入需为固定尺寸的特征向量。这种全连接的方式就会产生大量的参数，也会占据很大的计算资源。AlexNet网络结构由5个卷积层和3个全连接层组成，其中3个全连接层带来的参数和计算占了整个网络的几乎全部。由于全连接层带来的大量参数计算，有些网络，如GoogLeNet，为了节省计算资源，选择丢弃了全连接层，用平均池化的层加上一个线性层来替代。

5.softmax层

特征在经过全连接层作用后，会被送入用于分类的softmax层。该网络层通过softmax函数：

$$\sigma : \mathbb{R}^K \to [0,1]^K$$

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for j=1,..., K}$$

输出K个取值范围在$[0,1]$的概率，每个值对应划分为该类的概率。

6.损失函数层

损失函数层一般是网络的最后一层，该层的函数选择决定了该如何处理预测值与真实值的偏差。对于不同任务(分类或回归)，需选择不同的损失函数。比如Softmax损失函数用于N个互相排斥类别的分类问题。Sigmoid交叉熵用于预测$[0,1]$中的K个独立概率值。Euclidean损失用于实值回归问题。后面的检测网络在预测框的回归问题上用了SmoothL1Loss函数。

上面说过，损失函数层一般是在网络的最后一层。但在神经网络搭建的越来越深深的过程中，梯度消失的问题使得网络可能训练不佳。为了减轻网络深度带来的这些问题，一些网络结构选择将中间层的损失输出，并参与网络的反向传播。比如，GoogeLet除了在网络的最后用softmax分层得到损失函数外，它还在前面的稍浅的网络层添加了softmax层。并且从该层所在的位置开始参与反向传播，更新权重，以期改善梯度消失的问题。

第 2 节　检测网络

不同于图像分类问题，目标检测网络需给出一张图像里的物体的定位。卷积神经网络在抓取图像特征上强大的优势，使得该结构在各种图像分类赛事里成为了主流网络。在这之前，检测网络最常用的图像特征是SIFT和HOG，但是基于这些特征的模型，无论是通过集成的方式或是对算法更优的改良，结果提升并不明显，这反映出了SIFT和HOG的局限性。自然的，人们开始利用将卷积神经网络能很好抓取图像特征的这个特点，将其用于目标检测网络上。

2.1. R-CNN

其实在R-CNN提出前，已经有提出建立在类似于CNN网络结构的，使用滑窗方法的检测器OverFeat。但通过在ILSVRC2013检测的表现上，R-CNN在两百个种类上的检测表现，都远优于OverFeat。

如同OverFeat一样，以往的检测网络通过滑窗操作，获取一系列候选框，再将这些候选框送进分类网络。滑窗搜索是一种穷举搜索：选择一个窗口大小，滑动几个像素点扫描整张图像。由于图像中物体的大小不一，所以还需要改变窗口的大小和长宽比例，继续重复扫描整幅图。可以看到，在这“撒网”似的捕捉下，利用穷举的框将物体框住。但通过滑窗生成的候选框由于数目过于庞大，生成很多候选区域，还需

要将这些区域都送入分类网络，使得整个检测网络效率非常低下。为了这个检测网络更加的高效，R-CNN选择了[8]中提出的区域生成(region proposal)，采用了选择性搜索(selective search)的方式，得到可能还有检测目标的图像候选区域。

图片中包含了很多种不同层次的语义，拥有非常多的信息，包括纹理，形状，颜色等等。显然，所属同一个物体的颜色和纹理都应该是高度相近的。将算法中相似性的定义取为上述的图像特征的相近，建立合并规则，选择对图像优先合并以下四种情形：

1.颜色（颜色直方图）相近的

2.纹理（梯度直方图）相近的

3.合并后总面积小的

4.合并后，总面积在其BBOX所占比例大的，保证和并后形状规则

选择性搜索的具体算法如下：

Algorithm 1: 选择性搜索

Data: 彩色图像
Result: 所有合并过的区域r_t的集合
完全割裂整个图像，获取初始分割区域$R = r_1, r_2, , , r_n$;
初始化相似度集合$S = \emptyset$;
for 相邻的区域对(r_i, r_j) **do**
    计算(r_i, r_j)之间的相似度$s\left(r_i, r_j\right)$;
    $S = S \cup s\left(r_i, r_j\right)$;
end
while $S \neq \emptyset$ **do**
    得到最高的相似度值:$s\left(r_i, r_j\right) = max\left(S\right)$;
    合并相似区域: $r_t = r_i \cup r_j$ 从S里面移除所有关于区域r_i的相似度:
    $S = S \setminus s\left(r_i, r_*\right)$;
    从S里面移除所有关于区域r_j的相似度: $S = S \setminus s\left(r_j, r_*\right)$;
    计算r_t与它相邻区域的相似度得到相似度集S_t;
    更新相似度集:$S = S \cup S_t$;
    更新区域集:$R = R \cup r_t$;
end

经过选择性搜索，将会得到若干个候选区域。接下来就是对这些候选区域进行特征提取。

但在将这些候选区域送入CNN前，还需要对这些不同大小的候选区域做裁剪和填充，使得它们具有和网络兼容的输入尺寸，然后再被送进网络训练。

特征提取部分的卷积神经网络由五个卷积层以及两个全连接层组成。网络预先在ImageNet完成训练，由于ImageNet一共有1000种物体分类，当在小数据集微调时，需要把最后的输出改成N+1，其中N为小数据集的类别数，增加的一类为背景类。将训练集通过选择性搜索的方式，得到每张图像的若干候选区域。在将这些候选区域送进卷积神经网络之前微调之前，还需要对它们进行正反例的标注。把与原标注框的交并比(IoU)为0.5及以上标注为正例，反之交并比低于0.5的标注为反例。其中交并比的定义为：

$$IoU = \frac{area(A) \cap area(B)}{area(A) \cup area(B)}$$

其中A和B表示任意两个区域，area表示面积。用这些标注过的训练集微调神经网络，训练特征提取网络。

经过特征提取网络后，每个候选区域对应得到还需要将所得的图像特征输入训练好的支持向量机（SVM）分类器，对图像进行分类。每一类物体将会对应训练一个SVM分类器，其训练集依然由设置的IoU阈值选出。将IoU低于0.3的视为反例，IoU在0.3以上的视为正例，依次训练出对应N个类别的N个SVM分类器。

为了提高对物体定位的表现，还将对候选框进行框的回归。当经过SVM分类器后，候选区域对应每一类别会得到相应的打分。但在最终比对标注的框前，还需要对候选区域针对每个类别进行框的回归，也就意味着需要训练N个回归的式子。

回归算法的输入是N个训练对$\{(P^i, G^i)\}_{i=1,\ldots,N}$，其中$P^i = \left(P^i_x, P^i_y, P^i_w, P^i_h\right)$，分别代表候选框的中心点的横纵坐标及候选框的宽度和高度。每个标注的框G也用同样的方式表示，$G^i = \left(G^i_x, G^i_y, G^i_w, G^i_h\right)$。该回归学习的目的就是学习到从候选框$P$到标注框$G$的映射变换。

选取四个函数$d_x(P)$,$d_y(P)$,$d_w(P)$和$d_h(P)$来表示转换的参数。前两个表示候选框的中心点的尺度不变的平移，后两个表示候选框的宽度和高度的对数空间平移。当学习到这几个函数后，就可以通过下面的式子，将输入的候选框P转化成模型给出的最终

预测的框$\hat{G}$:

$$\hat{G}_x = P_w d_x(P) + P_x$$
$$\hat{G}_y = P_h d_y(P) + P_y$$
$$\hat{G}_w = P_w \exp(d_w(P))$$
$$\hat{G}_h = P_h \exp(d_h(P))$$

可以看到，相比滑窗似的“撒网”搜索，选择性搜索生成的候选区域可信度高，且数目大大的较少。且R-CNN提出对框的回归，校正预测值的偏移，提高了检测的准确率。

2.2. SPPnet

R-CNN中使用的选择性搜索。会得到大小不一、长宽比例也不同的候选区域。所以在将候选区域送进卷积神经网络前，还需要对它们进行一定的裁剪填充和放缩。然而这一步很有可能仍然包含不了整个检测的物体，或者使得原本的检测物体发生形变，影响分类的判断。

卷积神经网络中的全连接层，其输入是一个固定尺寸的特征图，这也是需要将候选区域处理成固定大小的原因。由前面的介绍可知，卷积层和池化层都是滑窗的操作，它们可以作用于任意尺寸的输入，同时它们的输出大小是可变的。但全连接层或者分类层是一个线性操作，是将输入的特征和权重矩阵做乘法，故而需要固定输入图像的尺寸。

既要保证图片的完整性和真实性，又要保留神经网络中的全连接层，用于完成最高层的推理。SPPnet便提出在原来的卷积网络中加入空间金字塔池化，Spatial Pyramid Pooling(SPP)。任意尺寸的特征图输入SPP层后，输出特征图都是固定大小。将它放在最后一层卷积层后，全连接层前，相当于在网络层级较深的地方做了信息的叠加，以省掉在输入神经网络前对候选框尺寸的预处理，却不影响卷积网络的结构。

空间金字塔池化是词袋模型的一种延伸，是计算机视觉领域最成功的方法之一。它将图像划分成从精细到粗糙不同的程度，然后把它们的局部特征相叠加。具体说来，设得到的特征图为f，其大小假设为$W \times H$。在f上划分4个层级的空间金字塔，即将其划分为1×1,2×2,4×4,6×6四种规格，那么每个图像块的大小就是$\frac{W}{k} \times \frac{H}{k}$,其中$k = 1, 2, 4, 6$。每张图无论其尺寸大小，皆可得到50个图像块。对每个小块做最大池化，每个图像块将输出一个1维的向量，50个图像块则会得到一个50维的向量。由于原网络最后的卷积层输出的通道为256,则经过SPP层后，每张图像会得到$50 \times 256 = 12800$维的表示，然后再将这12800维的向量输入全连接层。

相比传统的池化层，SPP具有以下三个特点：

(1)无论输入的图像的尺寸如何，SPP都能生成一个固定长度的输出，而滑窗式的池化层的输出大小和输入的大小是紧密相关的；

(2)SPP使用多层级的空间信息，然而滑窗式的池化每次只是用了窗口大小的图像信息。相较而言，多层级池化对物体的形变会更具鲁棒性；

(3)由于输入特征图尺寸的灵活性，SPP能够池化各种尺度下的特征。

SPPnet和R-CNN一样，通过选择性搜索得到候选框。但是R-CNN对所得的候选区域全部输入神经网络分别进行特征提取，SPPnet仅仅需要将原始的图像送入神经网络，共享最后一层卷积层输出的特征图。对比R-CNN的历遍两千次CNN的计算，SPPnet仅计算一次，所以相比R-CNN的检测速度，SSPnet的检测速度明显显著提高。

2.3. Fast-RCNN

尽管较于R-CNN，SPPnet在训练和测试阶段的速度都有很大提升。但依然存有和R-CNN一样的缺点，即整个物体检测是一个多步骤的程序，包括特征抓取，用对数损失微调网络，训练SVMs，最后对候选框回归。这些分任务给检测带来了速度上的影响。此外，抓取的特征还需要存储在硬盘上，也是对存储资源的限制。而且SPPnet微调不像R-CNN，它仅有SPP层后面的全连接层会更新权重，因为卷积特征是线下计算的，从而无法在微调阶段反向传播误差。显然，这限制了比较深的网络的表现。

针对上述的问题，提出了更为迅速高效的网络Fast-RCNN，无论是速度还是准确率上，相较于R-CNN和SPPnet，都有了很大的提升。

和R-CNN和SPPnet一样，Fast-RCNN用选择性搜索的到近2000个候选区域，然后将整张图像送入卷积网络，通过卷积层和最大池化层后得到一个特征图。对于每一个候选区域，RoI(Region of Interest)池化层将从特征图中抓取一个固定长度的特征向量，将该特征送入网络的全连接层，最终接入两个输出层：一层生成N+1维的softmax概率，一层生成对应N个类别的预测候选框的四个实值坐标。

其中RoI池化层是SPP层的一种特殊情形，它只将特征图划分为$h \times w$区域，例如取$h = w = 7$，同样对每个图像块最大池化，每个通道的特征图对应得到7×7=49维向量。

对于在ImageNet上完成预训练的卷积神经网络，当被作为Fast-RCNN的特征抓取滤波器导入后，还需要做一些变化。首先将最后一个最大池化层替换为上述的RoI池化

层;其次网络的最后一个全连接层和softmax层替换为两个输出层，一为全连接层加一个N+1类的softmax分类层，另一个候选框的回归；最后，整个网络的输入为两部分，包括所有的图片以及这些图片对应的候选区域。

网络的训练通过反向传播误差来更新权重。当每个训练样本(候选区域)来自不同的图像时，通过SPP层的反向传播会非常低效。原因是每个候选区域可能具有非常大的感受野，通常包括整个输入图像。由于前向传导必须处理整个感受野，训练输入就非常大。

Fast-RCNN提出了更有效的训练方法，即训练时的特征共享。训练过程中，随机梯度下降算通过分级采样获取小批量训练数据，首先随机取样N张图片，然后每张图片取样R个候选区域。文章选取每个小批量构成方式为，取训练集中的两张图片，每张图片取64个候选区域，一共得到128个候选。取和标注框交并比大于0.5，且为25%的候选框，将其标注为对应的类别，剩余的交并比属于[0,0.5)的候选区域将被打上背景的标签，达到相同图像的候选区域在前向和反向传导中共享计算和内存的目的。

除了分级采样外，Fast-RCNN使用一个精简的训练过程，一次微调过程将联合优化softmax分类器和候选框的回归。而不再是训练softmax分类器，SVMs和对候选框回归三个分开的过程。

Fast-RCNN有两个输出层，设softmax的输出为$p = (p_0, p_1, ...p_N)$,N为所有类别的总数，回归层输出为$t_n = (t_x^n, t_y^n, t_w^n, t_h^n)$，其中$n$是某一个分类，$t^n$和R-CNN中候选框回归一样，代表候选框相对于标注框尺度不变的平移和宽度高度对数空间的变化。每个候选框标注了类别c和一个标注的框的坐标v。对每一个候选区域使用多任务损失函数去联合训练分类和回归。

第三章　数据集介绍

第 1 节　数据来源

互联网时代也是信息共享时代，信息通过网络平台传递和分享，各种网络数据爆炸式增长。而学术界因为实际数据的缺失，很多算法没有大量真实样本的检验，其发展和应用都具有很大的限制性。有一些学者和机构开始意识到，可以利用这些网络资源构建和维护大批量的标注数据集。这其中就包括手写数字数据库MNIST，目标检测小数据库PASCAL VOC,用于图像识别/分割/释义的数据库COCO，以及本文预训练所用数据集，用于图像识别和目标检测的ImageNet数据库。

ImageNet是由斯坦福的李飞飞团队创建，是一个巨大的用于视觉物体识别的图片库。根据图中的物体信息，每张图片被人工标注对应的种类标签。此外，还有一百多万张照片可用于目标检测，也被标记了标注框。截至笔者写文章的此时，该数据集已经有14,197,122张图像，涵盖21,841种类别。ImageNet在2009年计算机视觉和模式识别会议上首次被公开。2010年开始，每年ImageNet举办大规模视觉识别挑战赛(ILSVRC)，利用标注图像对1,000类图片进行分类任务测试。2011年，ILSVRC分类top-5错误率为25%。2012年，深度卷积网络Alexnet错误率降到16%。到2015年，已有算法在ILSVRC分类任务已经超过了人眼识别率。在历年的ILSVRC中，优秀的算法不断涌现，其中深层神经网络的不断发展，使分类任务的准确率不断地取得突破，最终在小任务上的表现超过了人眼。

对于人脸检测任务来说，公开的标记人脸数据集样本数量较少，而且一般只有人脸的标注而没有年龄的信息。为了同时拥有年龄信息和性别信息。我们选用了IMDB-WIKI数据集。其中IMDb全称为互联网电影资料库(Internet Movie Database)，是一个关于电影演员、电影、电影制作、电视明星和电视剧的在线数据库。截止到2012年2月24日，IMDb共收录了4,530,159名人物的资料。从这些人物的资料中，可以获取出生年月，性别等相关信息。再从人物对应的图片库中获取图片，这些图片包括了剧照，活动照片等等。根据图片拍摄的时间戳，减去人物的出生时间，便可得到图片拍摄时，图像中人物的真实年龄。

经过筛选，一共选取了IMDb榜单上全球最知名的100,000演员，爬取了他们的出

生日期，姓名，性别和该人物相关的所有图片。此外维基百科也同时包含了人物的生日，性别和照片信息，和IMDB数据搜集方式一样，爬取整理了WIKI数据库。移除了没有时间戳的照片后(没有拍摄时间信息的图片)，不失一般性，我们假定只有一个人像的照片就是人物本身，并且假定人物的出身日期以及拍摄年份皆为正确的，这样就可以获取到有人物年龄信息的人像数据。除了时间戳可能有误以外，有很多照片是来自电视或电影的剧照，取了该影视剧的出品时间为拍摄时间。由于演员的角色设定，妆容一定程度上也会对数据的正确性有所影响。最后，IMDB-WIKI数据集一共收集了来自IMDB的20284个名人的460,723张人脸相片以及来自维基百科的62,328张人脸照片，总计 523,051张图片。

有了人脸的性别信息以及年龄信息，下一步就是对图像中的人脸进行矩形框的标注。使用[14]中提出的vanilla DPM人脸检测方法，对数据集中的数据进行打框标注。

此外，本文还用了ChaLearn Looking at People(LAP) 2015年龄预测赛的公开数据。这些人像数据被标注了年龄信息，但没有性别信息。为了将这些人脸的年龄信息利用上，我们随机标注了他们性别，作为噪音数据传入网络参与训练。另，LAP 2015由于是人像年龄的分类比赛，数据中并未标注相应的矩形框。我们利用训练好的caffe网络对图片的人脸进行了矩形框标注，作为检测网络的训练集。

第 2 节　数据预处理

由于IMBD数据库中收集了很多剧照和活动照片等，包括了不止一人，无法对应唯一的人物信息。加上计算资源的限制，我们从50多万张的数据集中，挑选了维基百科的6万多张照片，作为本次训练的所有数据。在计算图像的人物年龄时，由于图像的拍摄时间只有年份而没有月份，我们假定拍摄时间为当年的年中，也即6月30日。根据数据集提供的出生日期，计算出图像拍摄时人物的年龄。由于数据仍然有些错误和缺失，为了尽力保证数据的准确性，我们剔除了性别信息缺失，有多张人脸，计算出来的年龄为负数或是超过100岁，以及没有能标注出人脸位置等情况的数据。经过筛选，剩余5万余张人脸图像。最后，由于标注的框是通过别的传统模型标注的，普遍偏小。为了让检测网络更好的捕捉到人脸位置，我们将框的尺寸增大为数据原标注框的120%。

把数据按照每间隔十岁划分为十类，即0-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99这十个分类。由数据的来源可知，数据主要集中在20代、30代和40代，0-9岁年龄段和10-19年龄段的名人由于较少，所以数据较20代和30代的来说样本比较少。同样，70代和80代以上的数据也偏少。我们发现所得种类的的样本数目分布不均衡，为了尽力让每类样本的分布能够平均，我们会通过重采样的方式，对原本类别中数据

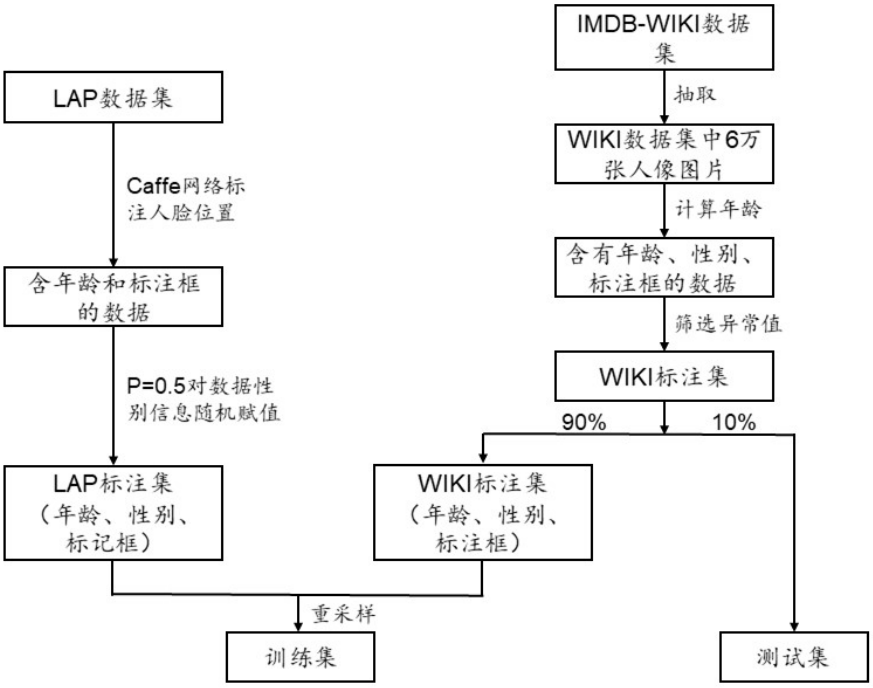


图 3.1: 数据处理流程图

较少的类别，重复采样三次，送进网络训练。

此外，经实验对比，发现模型对性别的识别率很高。所以为了更好的让模型学习到年龄段的分类，将LAP公开的人脸数据集也做为训练集送入网络。该训练集的人脸数据多为单人且拍摄条件都较好的样本，但是只有年龄的标签标注。为了更好的利用这些数据，我们p=0.5随机对样本的性别取了值，将该数据集作为训练集的一部分送入网络训练。当然，由于性别标注的不真实性，并未取该数据集中的数据作为测试集。

经过预处理的筛选，将剩余所有数据划分为训练集和测试集，其中训练集有4万多张图像，对每一年龄段分类，取测试集样本数为训练集样本数的10%,去除重采样和补充的数据的部分，抽取了测试集为4千多张图像。具体过程如3.1。

第四章　网络框架以及模型结果

一直以来，人脸检测都是计算机视觉领域的一个经典课题。该课题不仅具有重要的理论意义，也具有广泛的商业应用场景。例如，数码相机的智能人像对焦，社交网络中图片的人像圈定，智能手机的人脸解锁等等商用场景都是生活中常见的应用。

从传统的人脸检测网络到后来卷积神经网络，深度学习的发展克服了传统方法特征设计的限制，在图像分类任务上不断取得突破。为了将图像分类任务的长处和特点运用到目标检测网络上，我们需要解决以下两个问题：

一是如何利用深度网络来对检测物体进行定位;

二是如何用数量不多的标记检测数据训练出高性能的检测模型。

卷积神经网络在连年的ILSVRC 等图像分类赛事中表现突出。从LeNet，到AlexNet，再到之后的VGG和GoogLeNet，以及后来层数更深的Resnet，这些卷积神经网络结构在物体分类问题上，都是很重要的突破。为了充分利用卷积神经网络在图像分类上的突出表现，选择把检测网络视为数个候选区域的分类问题，也就解决了第一个问题。

对于第二个问题，神经网络的训练需要足够充分的标注数据。虽然图像分类的数据集已经很大的量级，但是针对检测问题的标注数据还是匮乏。因为检测问题的标注数据不仅需要给出物体的分类，还需要给出物体的定位，所以仅是标注这些数据就已经是一项比较耗费人工的工作。在少量的标注检测数据集上训练一个高效的模型，就需要借助神经网络浅层学习低级特征，而深层网络学习高级特征的特点。在样本量较大的标注数据集上预先训练图像分类模型，把其作为预训练模型载入，再用该预训练模型在小型数据集上进行特定领域的微调。也就是用迁移学习的方法来克服标注数据少的问题。

下面我们将结合上述两点，给出我们本次年龄和性别多标签预测任务的网络框架和一些具体参数。

第 1 节　Faster-RCNN

第二章内容中提到，Fast-RCNN将特征提取，物体分类以及边框回归整合到一个卷积神经网络中，大大提高了物体检测的效率。但和RCNN以及SPPnet一样，候选区域的生成采用了选择式搜索方法，独立于卷积神经网络。在Fast-RCNN的基础上，提出区域生成网络，Region Proposal Network(RPN)，通过共享卷积神经网络输出的特征图，生成原始图像的候选框。从而提出检测速度更快的网络结构——Faster-RCNN。

Faster-RCNN由两部分组成。第一部分是深层的全卷积网络RPN，用于生成候选区域;第二部分是Fast-RCNN网络，对RPN生成的候选区域进行目标检测。整个网络融合了区域生成网络和目标检测网络，是一个统一的检测结构。下面将主要介绍RPN网络的原理。

为了共享卷积层网络的特征计算，在Faster-RCNN中，RPN和Fast-RCNN共享了卷积层。RPN网络相当于在共享的网络层后再加了两个卷积层，一层将特征图的每一点编码成一个低维的向量，另一层用于输出分类概率和候选框的坐标。即特征图各点对应感受野为前景或者背景的概率，以及不同尺寸不同长宽比的k个候选框的边界。因此，本质上RPN网络是一个全卷积网络，所以输入的图像可以是任意尺寸。

由于图像中的物体有不同的大小和形状，因此标注矩形框也需具有不同的尺寸和长宽比。为了生成的候选框能尽量贴合物体的形状，需要k个不同大小和比例的矩形框。对于不同的矩形框，分类层和回归层输出对应的比例和坐标。分类层将判断是否是前景还是背景，输出2k个预测的概率。回归层将输出k个框的坐标，也即4k个输出。

而这k个不同的矩形框，就是RPN生成候选框的核心——anchor boxes。一个anchor box的中心，我们称之为锚点。通常选取3个尺寸和3个不同宽高比，那么特征图的点对应得到3×3=9个anchor boxes。对于尺寸为$W\times H$的卷积特征图，一共得到9WH个anchor boxes。

无论是锚点还是相对于锚点计算出候选区域的函数，都具有平移不变性。当图像中的物体进行了平移，相应候选区域也该有对应的平移，同样的函数也可以预测出候选区域的位置。anchor boxes重要的平移不变性同时也减少了模型的大小。

类似于第一章所述的DPM模型，RPN网络生成的候选框彼此间会有重叠。故得到的预选框需经过极大值抑制，筛选和真实标注框的交并比高于0.7的anchor boxes，再挑选交并比值排名前300或是前600的框送入RoI池化层。

设$x_i \in \mathbb{R}$是送进到RoI 池化层的第i个输入，其中RoI池化层将特征图划分为$H \times W$个区域。令y_{hw}表示区域$R(h,w)$对应的池化输出。则$y_{hw} = x_{i^*(h,w)}$，其中$i^*(h,w) =$

$\text{argmax}_{i' \in R(h,w)} x_{i'}$。由于一个$x_i$可能和几个不同的输出$y_{hw}$相关联，故反向传播时计算的损失函数对$x_i$的偏导数为：

$$\frac{\partial L}{\partial x_i} = \sum_h \sum_w \left[i = i^* (h, w) \right] \frac{\partial L}{\partial y_{hw}}$$

简单说来，对每个候选区域r和每个池化层的输出神经元y_{hw}，若i是被对应于y_{hw}的最大池化所选出的指标，且i只和一个池化区域相关，那么$\frac{\partial L}{\partial y_{hw}} = 1$。若$i$和不同的$y_{hw}$关联，则累加所得的偏导。

第 2 节　人脸检测网络

2.1.　多标签学习

年龄和性别是人脸的两个重要特征。性别是一个固定的二值特征，年龄是一个连续的特征。但由于将年龄看成实数难以计算，且样本的数目较少，本文将年龄预测视为分类问题来处理。

之前的工作对于年龄预测问题的处理，常常是将其拆成两个子问题处理。如[16]中，先利用传统的人脸检测网络定位出人脸的位置，再将裁剪后的人脸图像送进神经网络做分类训练。这种分开训练的网络，检测器和分类器各自有不同的目标函数，可能与整个系统整体的性能指标存在偏差。而且，前面检测网络的性能高低会影响到后面的分类器的表现，分类器性能的上限由前面所训练的检测器确立。

一般地，年龄和人物的性别以及种族等因素都有很大的关系，它们并非是独立的特征。相同年龄的男性和女性，常常在脸部的纹理上表现出不同程度的衰老。同年龄段的男性会比女性看起来年纪更大。同样，种族的不同会导致肤色等的不同，不同种族的样本也会在年龄预测问题上有所影响。为了尽力消除这些原因带来的影响，除了对人脸的年龄预测，我们还引入了二值的性别标签，将问题变成多标签的预测。

为了让训练的网络在整体上达到最优的性能，本文把人脸检测和年龄性别预测视为一个统一的问题，我们将整个模型构建成一个端到端的网络。以Faster-RCNN为模型的基础，针对人脸识别的特性做出相应的更改，使模型对此问题更具有适应性。网络的框架结构具体如图4.1。

整个网络的目标函数由两个分问题的目标函数组成:

$$L = L_{\text{rpn}} + L_{\text{rcnn}} \tag{4.1}$$

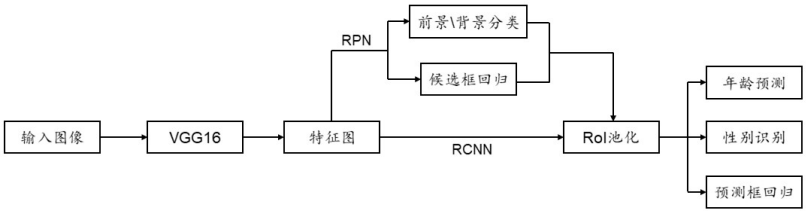


图 4.1: 网络结构

其中L_{rpn}为RPN网络层的损失函数，包含预选框的前景背景分类损失，以及候选框的回归损失。L_{rcnn}表示被检测人脸的年龄和性别分类损失，以及预选框的回归损失。如下:

$$L_{\text{rcnn}} = L_{\text{age}} + \lambda L_{\text{gender}} + \beta L_{\text{box}}$$

上式子中超参数λ，β控制着三个损失函数之间的比重，本文取$\lambda = \beta = 1$。L_{age}和L_{gender}是年龄和性别的交叉熵损失，如下:

$$\begin{aligned}\text{loss}(x, c) &= -\log\left(\frac{\exp(x_{\text{c}})}{\sum_j \exp\left(x_j\right)}\right) \\ &= -x_{\text{c}} + \log\left(\sum_j \exp(x_j)\right)\end{aligned}$$

其中x对每个分类的概率打分，其大小为$[M, C]$。M表示每个批量数据集的样本数，C表示所有的类别数。式中的c是大小为M的一维张量，代表每个样本的真实分类标签。

而第二个损失函数L_{box}，是对真实分类c对应的标注框$v = (v_x, v_y, v_w, v_h)$，和对类别c预测的框$t^c = (t_x^c, t_y^c, t_w^c, t_h^c)$定义的:

$$L_{\text{box}}(t^c, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^c - v_i)$$

其中L_1光滑函数定义为

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < \ 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

从上式可以看出，L_1损失比R-CNN和SPPnet中使用的L_2损失更具有鲁棒性。一旦$|x|$超过了1，L_2损失会呈现指数级的增长。因而更需要调节学习率去避免梯度爆炸的情形。

2.2. 预测框回归

Faster-RCNN在对不同种类的物体检测时，由于种类之间的外形差异，标注的矩形框也大小比例不一。在对预测的框做回归的时候，需要对每一类的框分别做回归。但是在人脸检测中，虽然要对年龄做十个类别的分类预测，但是候选框框住的内容只是人脸。所以我们更改了回归网络层的输出，将对十类框的回归改为了一类框的回归，提升网络的训练速度。

RPN网络在训练时，需给每一个anchor box标注一个二值标签，即是否含有待检测的目标。标记规则具体如下：

(1) 标记和一个真实的标注框交并比最高的anchor box为正例；

(2) 标记和任意的标注框交并比高于0.7的anchor box为正例；

(3) 标记和所有标注框的交并比小于0.3的anchor box为反例。

定义好正反例后，最小化4.1中RPN的目标函数L_{rpn}：

$$L_{\text{rpn}} = \sum_i L\left(\{p_i\}, \{t_i\}\right)$$

其中$L\left(\{p_i\}, \{t_i\}\right)$为：

$$L\left(\{p_i\}, \{t_i\}\right) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}\left(p_i, p_i^*\right) + \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}\left(t_i, t_i^*\right) \tag{4.2}$$

上式中，i表示每个小批量数据集中第i个anchor box，p_i是表示第i个anchor box是检测目标的概率。若第i个anchor box是正例，则p_i^*为1。若该anchor box是反例，则p_i^*为0。t_i表示预测的候选框的四个坐标，t_i^*表示标记为正例的anchor box所对应标注框的坐标。分类损失函数L_{cls}是两个是两个类别(前景和背景)的对数损失。

对于式4.2在中的回归损失函数$L_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*)$，R是如前定义的$L_1$的光滑函数。由于RPN的框的回归是在候选框和真实标注框之间的回归，两个框的差值可能偏大。为了使损失函数在差值稍大的情况下能减轻梯度爆炸的情况，修改了其中自变量的取值范围。更改过候选框的损失函数如下：

$$L_{\text{reg}}(t_i, t_i^*) = \begin{cases} (t_i - t_i^*)^2 \cdot \dfrac{9}{2} & \text{if } |t_i - t_i^*| < \dfrac{1}{9} \\ |t_i - t_i^*| - \dfrac{1}{18} & \text{otherwise} \end{cases}$$

由于框的回归只对含有待检测目标的anchor boxes有作用，所以式4.2中的$p_i^* L_{\text{reg}}$代表只有当anchor box为正例时，回归函数的损失才被激活。

由整个检测过程可知，我们将得到三种不同的矩形框。第一种是图像本身的标注框，第二种是RPN网络的anchor box，第三种是候选框。在RPN网络中，从anchor box到候选框还需要做一次框的校正回归。

对于anchor box和候选框的回归,不妨设函数$\mathrm{d}i(P)$(i代表x,y,w,h中的一个)是候选区域的最后一个池化层输出特征的线性函数$\phi(P)$。因此可以将$\mathrm{d}i$表示为$\mathrm{d}i(P)=w_i^{\mathrm{T}}\phi(P)$，其中$w_i$即是回归模型所需要学习的参数，通过岭回归来学习得到：

$$w_i=\arg\min_{\hat{w}_i}\Sigma_k^N(t_i^k-\hat{w}_i^{\mathrm{T}}\phi(P^k))^2+\lambda\|\hat{w}_i\|^2$$

其中$\lambda\|\hat{w}_i\|^2$为权重的正则项，t_i^k表示第k个候选框和anchor box的偏差，t_i如下定义：

$$\begin{aligned}t_x&=(x-x_a)/\omega_a\\t_y&=(y-y_a)/h_a\\t_w&=\log(\omega/\omega_a)\\t_h&=\log(h/h_a)\end{aligned}$$

则回归校正后所得候选P的中心点坐标及宽高为：

$$\begin{aligned}p_x&=\omega\mathrm{d}x+x-0.5\omega\mathrm{e}^{\mathrm{d}\omega}\\p_y&=h\mathrm{d}y+y-0.5h\mathrm{e}^{\mathrm{d}h}\\p_w&=\omega\mathrm{d}x+x+0.5\omega\mathrm{e}^{\mathrm{d}\omega}\\p_h&=h\mathrm{d}y+y+0.5h\mathrm{e}^{\mathrm{d}h}\end{aligned}$$

本文采取近似联合训练，让模型根据数据本身，对两个网络自动调节，提升模型整体的契合度，达到整体上的最优性能。

第 3 节　预训练模型

本文年龄和性别预测任务，选取的特征提取网络为VGG16。VGG网络由牛津大学计算机视觉组提出，通过搭建更深层的卷积神经网络，提高神经网络对特征的提取能力。VGG16在2014年ImageNet定位和分类挑战赛上分别获得了第一和第二的成绩。

VGG网络舍弃了Alexnet中使用的类似11×11，5×5较大的卷积核，而是通过全部使用了3×3的小卷积核，释放了计算资源，将神经网络的层数搭建到16-19层。本文所用的VGG16包含了13个卷积层和3个全连接层，选择其前13层作为RPN网络和Fast-RCNN的共享卷积层。

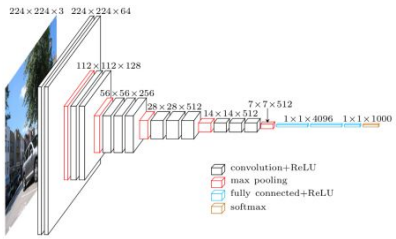


图 4.2: VGG16网络结构

神经网络的训练需要大量数据样本的支持。越来越多标注数据集的出现，使得神经网络在图像分类和目标检测等问题上不断取得突破。但是这些数据库并不能覆盖所有的领域。有很多实际的问题因为数据过少，缺乏标注的样本数据。而训练神经网络的基本，就是要有足够大的训练集。期望利用在已有的大量的标注数据集上训练好的网络，帮助模型在新的少量数据集上取得不错的结果，这也是迁移学习所关心的内容。

一般地，对应不同物种间的分类问题，网络可以通过较为粗糙的轮廓或形状特征对物体进行分类。比如数据集PASACL VOC 2007，该数据库包含了近1万张图片。这些照片被分成20个目录，包括人，动物(猫、牛、马、狗、羊等)，交通工具(飞机、自行车、火车、摩托车等)，室内用品(椅子、餐桌、沙发、电视等)。种类和种类之间差别越大，就可以用较浅层的特征将它们区分。Faster-RCNN在PASCAL VOC 2007数据集上的mAP可以达到70%。

然而人脸年龄分类任务属于同个物种间的区分，浅层网络提取的特征抓取的轮廓形状等信息，不足以用来作出种类间的分类预测。还需要捕捉更为深层的图像特征，才能对不同人脸作出年龄分类。

本文选用在ImageNet上完成训练的VGG16，作为网络的预训练模型。VGG16是一个比较稳定也比较经典的模型，所以很多迁移学习的任务都选择用它作为特征提取网络。VGG16一共有16层网络，除去最后3层全连接层，一共含有13层卷积层。将这13层卷积层作为检测网络的特征提取器。最后一层卷积层输出的特征图被输入进RPN网络层生成候选框，然后生成的候选框会在特征图上做RoI池化，所得特征将用于年龄段分类、性别分类以及预选框的回归。

通过第三章对卷积神经网络的介绍可以知道，神经网络浅层学习到的是较为粗糙，较为低级的图像特征。而较深的网络学习到的是较为精细，较为高级的特征。为了利用网络在ImageNet上学习的较为低级图像特征，我们固定了网络的前五层卷积层，这

前五层卷积层将参与网络的前向传播过程，负责特征的提取，但是不会参与损失的反向传播，即对应的权重不会更新。

第 4 节　抑制过拟合

通常情况下，在小数据集上训练的网络，可能会存有在验证集和测试集上不泛化的问题。因此模型常常会出现过拟合的现象。直观上，增加训练集的样本数，就可以使得模型训练时一定程度上减少过拟合。然而，数据来源通常是不易获取的。在条件限制的情况下，利用有限的数据，抑制过拟合现象的发生，成为一个神经网络非常关注的问题。本文采用了dropout和数据增强两种方法。

4.1. dropout

一般地，多种模型的结合会吸收不同模型的优势。特别是不同的模型在不同的数据上训练得到，且具有不同的结构的时候。但同样也会带来一些问题：

1.训练多个网络需要调一大堆参数;

2.训练大的网络需要很大的训练集，若数据集不够大，那么不同的网络就不能用不同的数据集;

3.就算上述两点所述都能完成，最后得到一个较大的网络，在测试集上测试也会较慢。

为了减少过拟合，Hinton在[17]中提出了dropout机制，即训练时，把某些神经元（中间层或是可见层都可以）暂时地移除。此外，和这些神经元的连接也暂时的移除。也就是说，在训练过程中停掉某些神经元的作用随机地移除网络层中的神经元，停止该神经元和别的神经元的所有联系，从而达到通过训练一个网络，得到2^n网络集成的效果。模型加上一个dropout相当于训练了指数级别的多个网络，而且还是不同结构的。

如图如见，左边是原来的网络，有图是随机地移除了某些神经元，这些神经元用打叉的节点标记。那么可以从图中看出，移除了这些节点后，得到比原网络“更瘦”（网络层的宽度变小）的网络。

一般说来，对神经元的选择采取随机移除的机制：每个神经元相互独立，被dropout的概率是p。本文中训练模型时，以$p = 0.5$随机移除神经元，那么若网络中一共包括n个神经元，则会产生2^n种组合的网络。而权重其实是共享的，也就最多原来n个节点全连

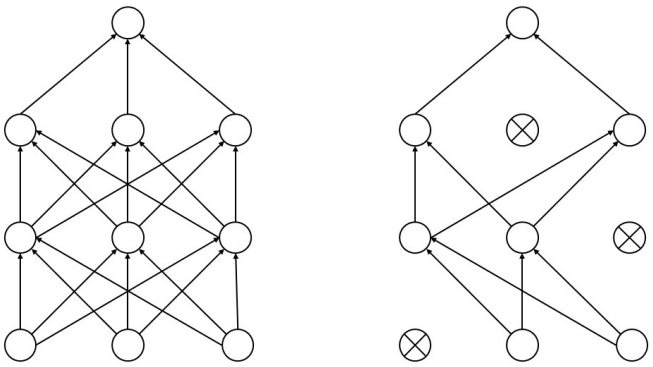


图 4.3: 左为含两个隐藏层的标准神经网络；右为对左边标准模型使用dropout机制

接的参数个数$\mathcal{O}_(n^2)$。这样就完成了训练大量的网络，而参数个数又没有增加，所以达到了防止过拟合的作用。

4.2. 数据增强

除了上述的dropout的使用，数据增强也是一个抑制过拟合的方法。即利用这些图片数据，生成新的数据，以增加数据的样本量。在图片分类问题中，由于图片的翻转和旋转并不会影响图像自身的分类。所以在保证标签不变的情况下，对图像进行平移，缩放和旋转训练集，就可以增加样本的数目。

本文使用的样本集数目较少，训练时很容易出现过拟合现象。由于是检测任务，对数据增强增强时，还需考虑到标注框的位置，不能随意的平移。最后选择了一共五种数据增强的方式，分别为镜像翻转，随机裁剪，亮度增强，对比度增强，添加高斯噪音点。

以上五种数据增强的方法，除了镜像翻转和随机裁剪以外，其他三种产生的图像由于物体位置并未移动，所以标注框的位置不变。数据增强算法可见算法2。

除了算法中所述的四种数据增强方法外，我们还对输入图像做了随机的剪裁。检测任务不像图像分类任务，图中标注框也需要被考虑进去。设原图的宽度和高度分别为W和H，我们取裁后的照片宽度为w，高度为h。随机取$w \in [0.3 \cdot W, W]$，$h \in [0.3 \cdot H, H]$。再随机取新图像的左上角坐标为$x_1 \in [0, W - w]$，$y_1 \in [0, H - h]$，故新图像在原图像的坐标表示为$[x_1, y_1, x_1 + w, y_1 + h]$。接下来考虑新图像与标注框的交并比，我们选择$\text{None}, 0.3, 0.5, 0.75$种模式，每次随机裁剪开始前先随机选择一个模式。

Algorithm 2: 数据增强

Data: 带标注的图像数据
Result: 增强后的图像数据
数据集$P = \{P_1, P_2, , , P_N\}$;
初始化新数据集合$S = \emptyset$;
for $k \leftarrow 1$ **to** N **do**
    计算P_k的宽度M和高度H，记录标签$label_k$ =[年龄段，性别];
    及P_k的标注框$bbox_k = [x_1, y_1, x_2, y_2]$;
    和; 对r随机赋值0或1;
    if $r == 1$ **then**　　// 镜像翻转
        $P_k(i, j) = P_k(M - i, j)$;
        $bbox_k = [M - x_1, y_1, M - x_2, y_2]$，$label_k = label_k$;
    end
    对r随机赋值0或1;
    if $r == 1$ **then**　　// 亮度增强
        在区间$[-255, 255]$内对δ随机赋值;
        $P_k(i, j) = P_k(i, j) + \delta$;
        $bbox_k = bbox_k$，$label_k = label_k$;
    end
    对r随机赋值0或1;
    if $r == 1$ **then**　　// 对比增强
        随机赋值$\alpha \in [0.5, 1.5]$;
        $P_k(i, j) = \alpha P_k(i, j)$;
        $bbox_k = bbox_k$，$label_k = label_k$;
    end
    对r随机赋值0或1;
    if $r == 1$ **then**　　// 增加高斯噪音
        在原图的每个通道每个点加上高斯项;
        $bbox_k = bbox_k$，$label_k = label_k$;
    end
    $S = S \cup P_k$;
end

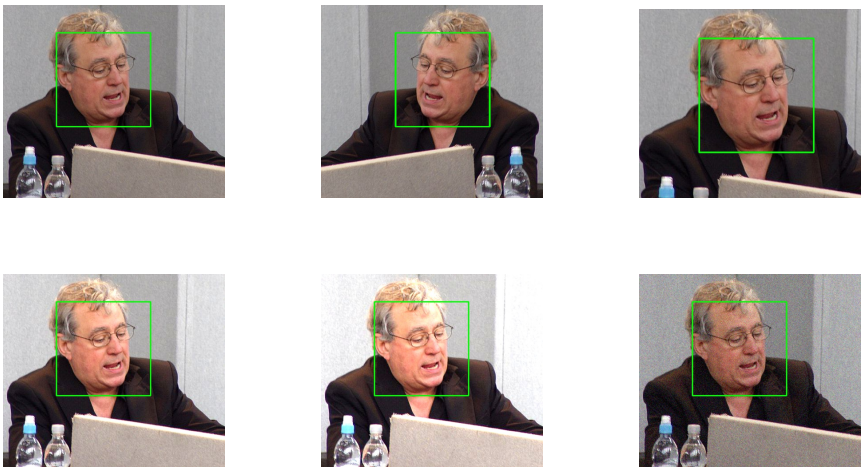


图 4.4: 图从上到下,从左到右依次为：原始图片，镜像翻转图片，随机裁剪图片；增强光照图片，对比增强，添加高斯噪音图片

None表示不裁剪，其他四种表示裁剪后的图像和标注框的交并比必须大于该比例，否则将再重复随机裁剪过程，直至输出。

数据增强算法可见算法2，增强的图片效果可见图4.4。

第 5 节　批量数据集的选取和处理

在将图片送入网络训练时，硬件资源不足以计算全部的样本，所以每次将随机抽取一部分样本集参与网络的训练。这部分样本称之小批量样本。换言之，每次网络反向传播的损失，只是这些小批量样本产生的损失，继而更新权重。下一次训练又将抽取另一批小批量样本，继续更新网络层的权重参数。如此循环往复，直至取完全部的样本。

通过对样本的观察发现，不同图片的长宽比，往往反应了图像中人被拍摄的比例。例如，宽高比较大的图片一般只摄取了人的上半身，宽高比较小的图片一般拍摄了人腰部以上，甚至全身照片。所以不同宽高比的图像一定程度上反映了图像中的人脸的位置或是比例特征。

由于每次网络的更新都是依赖于这些小批量样本的学习，为了提高网络的学习能力，我们选择把宽比高相近的图片放进同一个小批量样本中。特别地，我们将小批量样本中的数据全部处理成同样的宽高比例。这样，网络每次训练都是基于这些相似比例的图片，就会更集中地学习到它们的共性。同时，每次取得宽高比例都是随机的。


图 4.5: 图从上到下依次为：宽高比较大的照片；宽高比较小的照片

所以不存在对某些特征的图片过分学习的情况。

具体的做法如下。在读入训练集数据后，计算每张图片的宽高比例。根据图片的顺序，得到一个对应每张图片宽高比的列表。按从小到大的顺序对比例表进行排序。列表将会按照小批量样本的数目被均分(不若足重复采样补齐)成多组，这些比例组对应的图片，就是训练时被一起送进网络训练的小批量样本。

由于小批量样本中，图片的宽高比例按从小到大的顺序排列，在处理它们为同样宽高比时，需要考虑比例最小和最大的图片。设整个小批量样本集最小宽高比为r_1，最大宽高比为r_2。若r_2小于1,则取整个小批量样本集的宽高比为r_1。若r_1大于1，则取整个小批量集的宽高比为r_2。其余情况，将小批量集的宽高比选取为1。

对于比例和尺寸不在常规范围内的图片，我们会对其进行相应的裁剪或者调整图片的大小。例如，对于宽高比大于2或者小于0.5的图片，我们对其进行裁剪，使裁剪后图片的宽高比例在$[0.5, 2]$范围内。若图片的宽度或者高度超过设定的600像素，将通过双线性插值法调整图片的大小，使其尺寸在设定的阈值范围内。

第 6 节　环境及参数配置

本次年龄性别多标签预测的任务环境为：操作系统ubuntu16.04 64位操作系统，CUDA的版本为8.0.61，选用深度学习框架pytorch搭建程序。网络在6块型号为TITAN Xp的GPU上完成训练。学习率初始化取为1e2，每10个epoch衰减一次，衰减为上一次

学习率的0.1，模型优化方法选用了随机梯度下降算法。

第 7 节　模型结果

我们将训练好的模型在之前划分的6k张测试集上进行测试。模型在包含4331张图片的测试集上测试，用时为1314秒，平均每张图耗时0.3秒。相比将问题分为几个子问题来处理，我们的模型可以在较短时间内同时给出年龄段的分类和性别的识别。

其中年龄的预测结果如下表格：

表 4.1: 年龄段分类

年龄段	gen_1	gen_2	gen_3	gen_4	gen_5	gen_6	gen_7	gen_8	gen_9	gen_10	mAP
AP	0.237	0.401	0.667	0.468	0.456	0.419	0.428	0.399	0.401	0.305	0.418

可以看出，模型在20-30岁间的人脸图像上表现较好，mAP的值达到了0.667,这和我们之前统计的样本数目分布基本一致。20-30岁年龄段的图像数目较多，所以模型对该类别学习的比较好。而由于0-9岁和90-99岁的样本数目较少，尽管已经对这几个样本数较少的类别进行过重采样，模型对这两类的学习仍然不够充分，识别的效果比较差。

而模型在性别的二分类问题上表现较好，如下表：

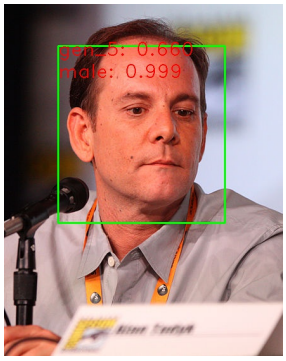
表 4.2: 性别识别

性别分类	female	male	mAP
AP	0.906	0.895	0.901

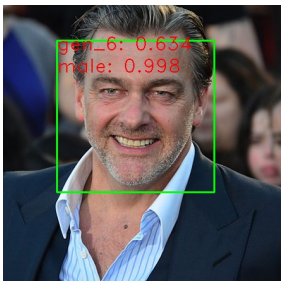
在前面介绍训练集的时候说过，为了让模型能对年龄段分类学习的更好，我们选择牺牲部分性别的准确率，将LAP数据集中的图片数据随机附上了性别的二值标签。

如图4.6展示了部分模型输出的结果：

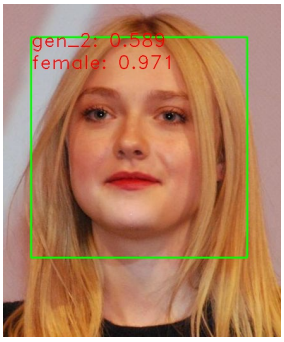
以上的结果都是给出分类并分类正确的图片。尽管给出了正确的分类，但是年龄段的分类概率都不是很高，说明模型对种类之间的特征区分度不是很高。而对于性别识别任务，可以看出除(5)外，其余给出的性别分类概率都在0.97以上。而图(5)中的人像由于年龄偏大，人眼观察也比较难以给出性别的判断。这也反应出，对年龄偏小和偏大的人做性别识别，存在一定的困难。



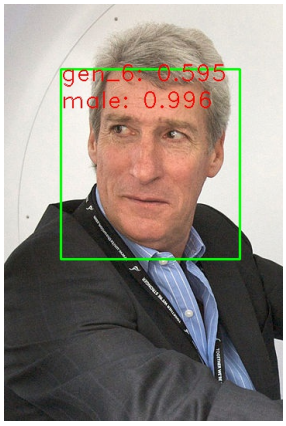
(a) gen_5; male



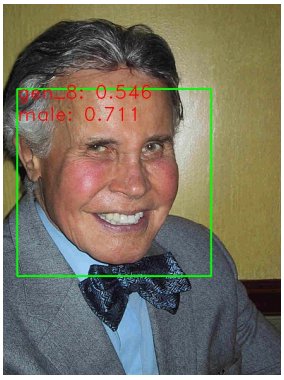
(b) gen_6; male



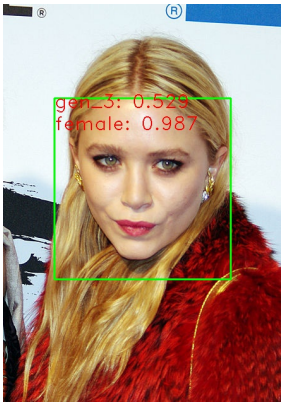
(c) gen_2; female



(d) gen_6; male



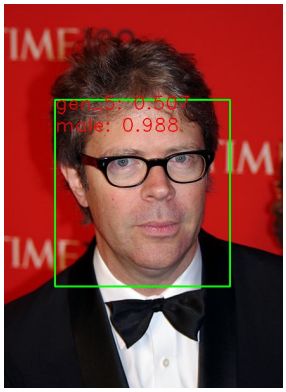
(e) gen_8; male



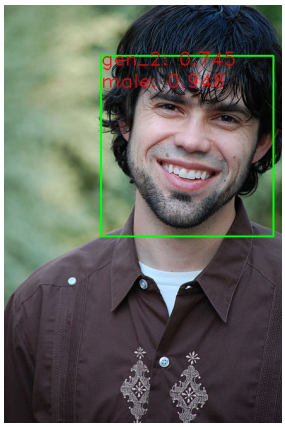
(f) gen_3; female

图 4.6: 分类正确的图片

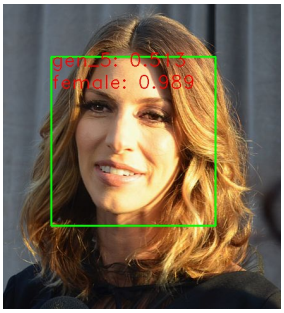
除了上图正确的样本分类外，图4.7展示了部分无法分类或是分类错误的人像：



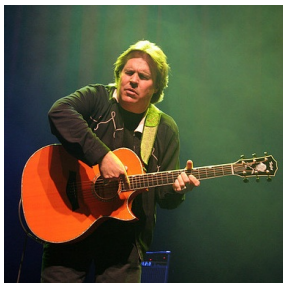
(a) gen_6; male



(b) gen_4; male



(c) gen_4; female



(d) gen_6; male

(e) gen_7; female

(f) gen_4; female

图 4.7: 分类有误的图片

可以看出，因为邻近年龄段界限比较模糊，有些人像被分到了前一个年龄段或者后一个年龄段，如4.7a和4.7c。而有的图片因为拍摄环境过于不清晰，脸部的特征不明显，导致图片无法给出具体的年龄分类，如4.7d。还有一些图片，如4.7f因为脸部特征被遮挡，无法得到较精细的特征进行分类。此外，模型到年纪较大的人像学习效果不佳，导致对年龄较大的人像识别不出来，如4.7e。

第五章　总结

人脸的年龄估计和性别分类有着广泛的应用场景，也是图像识别领域的经典课题。相对于二分类的性别识别问题，人的年龄因为种族，皮肤纹理，个人护理等因素，往往较难给出预测。

前面很多的工作主要是将该问题分解成几个子问题来处理。先通过人脸检测获得人脸在图像中的位置，根据位置结果裁取图片获得人脸图像。然后训练特征滤波器获得相应的人脸特征向量，最后通过训练不同的分类器分别估计年龄和识别性别。在深度学习兴起之前，传统的分类器通常选为支持向量机。随着卷积神经网络在图像识别任务上不断取得突破，传统人工设计的特征在其强大的特征抓取能力前显示出了局限性。所以，也有工作将特征滤波器和分类器的训练用一个卷积网络替代，通过CNN网络抓取特征并实现分类任务。

本文通过利用Faster-RCNN的目标检测框架，提出预测人像年龄和性别的多标签学习任务。我们将年龄预测视为分类问题，取间隔为十岁，将0 ~ 99岁的年龄划分为十个年龄段。将性别视为二分类问题，其中年龄段预测和性别分类预测共享了RoI池化层的特征向量。此外，因为所预测的目标皆为人脸，我们更改了RPN网络中的回归类别数，统一只做一类框的回归，提高训练的效率。在模型训练阶段，为了提高模型的泛化性，我们对输入的人脸数据做了随机的镜像翻转、裁剪、光照增强、对比增强、增加光照噪音五种数据增加的方法，抑制模型过拟合，提高其泛化性。且由于人脸图像不同的宽高比例，往往反映出人脸的位置以及比例等信息。为了让网络更好地集中学习到人脸的特征，我们按照图片的宽高比例排序，每次取比例邻近的图片为小批量数据集送入模型训练。

致谢

参考文献

[1] Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

[2] Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[3] Lowe, D.G., 1999. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on (Vol. 2, pp. 1150-1157). Ieee.

[4] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[5] Lin, M., Chen, Q. and Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

[6] Szegedy, Christian, et al. Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2015:1-9.

[7] Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. International journal of computer vision, 104(2), pp.154-171.

[8] Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. International journal of computer vision, 104(2), pp.154-171.

[9] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[10] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[11] He, K., Zhang, X., Ren, S. and Sun, J., 2014, September. Spatial pyramid pooling in deep convolutional networks for visual recognition. In european conference on computer vision (pp. 346-361). Springer, Cham.

[12] Girshick, R., 2015. Fast r-cnn. arXiv preprint arXiv:1504.08083.

[13] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[14] Mathias, M., Benenson, R., Pedersoli, M. and Van Gool, L., 2014, September. Face detection without bells and whistles. In European Conference on Computer Vision (pp. 720-735). Springer, Cham.

[15] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.

[16] Rothe, R., Timofte, R. and Van Gool, L., 2015. Dex: Deep expectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 10-15).

[17] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp.1929-1958.

[18] Chen, B.C., Chen, C.S. and Hsu, W.H., 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Transactions on Multimedia, 17(6), pp.804-815.

[19] Ren, S., Cao, X., Wei, Y. and Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1685-1692).

[20] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9), pp.1627-1645.

[21] Denton, E.L., Zaremba, W., Bruna, J., LeCun, Y. and Fergus, R., 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In Advances in neural information processing systems (pp. 1269-1277).

[22] Hosang, J., Benenson, R., Doll á r, P. and Schiele, B., 2016. What makes for effective detection proposals?. IEEE transactions on pattern analysis and machine intelligence, 38(4), pp.814-830.

[23] Lazebnik, S., Schmid, C. and Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer vision and pattern recognition, 2006 IEEE computer society conference on (Vol. 2, pp. 2169-2178). IEEE.

[24] Chen, B.C., Chen, C.S. and Hsu, W.H., 2014, September. Cross-age reference coding for age-invariant face recognition and retrieval. In European Conference on Computer Vision (pp. 768-783). Springer, Cham.

[25] Eidinger, E., Enbar, R. and Hassner, T., 2014. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12), pp.2170-2179.

[26] Han, H., Otto, C. and Jain, A.K., 2013, June. Age estimation from face images: Human vs. machine performance. In Biometrics (ICB), 2013 International Conference on (pp. 1-8). IEEE.

[27] Escalera, S., Fabian, J., Pardo, P., Bar ó , X., Gonzalez, J., Escalante, H.J., Misevic, D., Steiner, U. and Guyon, I., 2015. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1-9).