

本科生毕业论文

|  |  |
| --- | --- |
| **题目**： | **面向拥挤场景的人物检测和计数方法的设计与实现**  **Design and implementation on people detection and counting in crowded scenes** |
|  |  |

姓 名： 尤安升

学 号： 1300013023

院 系： 信息科学技术学院

本科专业： 计算机科学与技术

指导导师： 童云海

二○一七 年 五 月

北京大学本科毕业论文导师评阅表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 | 尤安升 | 学生学号 | 1300013023 | 论文成绩 |  |
| 学院（系） |  | | | 学生所在专业 |  |
| 导师姓名 |  | 导师单位/  所在研究所 |  | 导师职称 |  |
| 论文题目  （中、英文） | |  | | | |
| 导师评语  （包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价）  导师签名：  年 月 日 | | | | | |

**版权声明**

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

随着城市人口增加，影响社会公共安全的事件层出不穷，智能视频监控系统受到越来越多的重视。目标检测和人群计数是智能视频监控系统比较重要的课题，通过计算机视觉减少对人力资本的消耗，同时目标检测对泛身份识别领域有着关键性的研究价值和重要意义，而人群计数则能够预警人群安全事故，具有重要的现实意义。随着深度学习的发展，目标检测和人群计数得到了较为快速的发展，近几年提出了一系列基于深度学习的解决方案。本文在广泛调研国内外目标检测和人群计数方法的基础上，通过考虑视频帧的连续帧信息，即时序信息，设计并实现了面向拥挤场景视频的基于连续帧信息的目标检测和人群计数方法。

本文的研究内容和主要成果包括：

（1）独自标注了拥挤场景视频帧2350帧，同时探索了模拟光照、模糊处理、噪声处理、颜色抖动等一系列数据增强方法，效果较为逼真，在一定程度上解决了数据量少的问题，最后还设计实现了目标检测和人群计数系统；

（2）设计并实现了基于连续帧信息的目标检测和计数方法的一系列模型。本文提出了两种有效连接时序信息的方式，一种是沿时序卷积，通过1x1的卷积核将连续帧的特征进行线性组合；另一种是使用LSTM对当前帧的特征表示进行预测，最后将预测得到的结果与当前帧的特征进行线性组合；

（3）在标注数据集上进行训练测试，实验证明了连续帧信息对当前帧的目标检测和计数方法效果提升有一定作用。除此之外，与Faster R-CNN的比较证明了使用LSTM代替NMS的有效性，使用ResNet代替GoogLeNet证明了好的特征提取网络能提升目标检测和人群计数效果。

**关键词：数据增强、时序信息、目标检测、人群计数**

# Abstract

With the increase of urban population, there are more and more incidents that affect social public security. Meanwhile, the intelligent video surveillance system is paid more and more attention. Object detection and crowd counting are the most important topics of intelligent video surveillance system reducing the consumption of human resources through computer vision. Furthermore, object detection is of great significance in the field of universal identity recognition, and crowd counting can predict the accident of the crowd, which has important practical significance. With the development of deep learning, object detection and crowd counting have been developed rapidly and a series of solutions based on deep learning have been proposed in recent years. In this paper, we, based on the extensive research on object detection and crowd counting methods at home and aboard, using deep learning technology, present and implement a deep neural networks for crowded video scences considering the continuous frame information of video frames, namely, timing information.

The research contents and main achievements of this paper include:

(1) Annotated 2350 crowded scene video frames, and explored a series of data augmentation methods such as simulating illumination, fuzzy processing, adding noise and color jittering, which solve the problem of lacking of data to some degree, on account of the realistic effects. Finally, we have designed and implemented an object detection and crowd counting system.

(2) Designed and implemented a series of models of object detection and crowd counting methods based on continuous frame information. In this paper, we proposed two ways to effectively concatenate the timing information. One is to convolve along the time, using 1x1 convolution to linearly combine continuous frames. The other is to use LSTM to predict the feature map of the current frame. Finally, we linearly combine the predicted result and the current frame.

(3) The experiments were carried out on the annotation dataset. Luckily, the experiments proved that the continuous frame information had a certain effect on the effect of the object detection and crowd counting. In addition, the comparison with Faster R-CNN demonstrates the effectiveness of using LSTM to replace NMS. Moreover, Using ResNet to replace GoogLeNet demonstrates that a good feature extraction network can improve object detection and crowd counting.

**Key Words：data augmentation, timing information, object detection, crowd counting**

**全文目录**

[摘要 4](#_Toc482520081)

[Abstract 5](#_Toc482520082)

[第一章 绪论 8](#_Toc482520083)

[1. 研究背景和意义 8](#_Toc482520084)

[1.1 研究背景 8](#_Toc482520085)

[1.2 研究意义 9](#_Toc482520086)

[2. 主要研究内容 10](#_Toc482520087)

[第二章 相关研究现状 11](#_Toc482520088)

[1. 传统目标检测算法 11](#_Toc482520089)

[1.1 算法概述 11](#_Toc482520090)

[2. 基于Region Proposal的深度学习目标检测算法 11](#_Toc482520091)

[2.1 算法概述 11](#_Toc482520093)

[2.2 代表性算法之R-CNN 12](#_Toc482520094)

[2.3 代表性算法之Fast R-CNN 12](#_Toc482520095)

[2.4 代表性算法之Faster R-CNN 13](#_Toc482520096)

[3. 基于回归方法的深度学习目标检测算法 14](#_Toc482520097)

[3.1 算法概述 14](#_Toc482520099)

[3.2 代表性算法之YOLO 14](#_Toc482520100)

[3.3 代表性算法之SSD 16](#_Toc482520101)

[4. 本章小结 17](#_Toc482520102)

[第三章 具体实验 18](#_Toc482520103)

[1. 实验理论基础 18](#_Toc482520104)

[1.1 GoogleNet 18](#_Toc482520105)

[1.2 RNN和LSTM 20](#_Toc482520106)

[1.2.1 循环神经网络（RNN） 20](#_Toc482520107)

[1.2.2 长短期记忆网络（LSTM） 21](#_Toc482520108)

[1.3 实验理论基础 22](#_Toc482520109)

[2. 前期数据准备 23](#_Toc482520110)

[2.1 数据采集和标注 23](#_Toc482520112)

[2.2 数据增强 24](#_Toc482520113)

[3. 基于当前帧的目标检测和计数方法的实现 26](#_Toc482520114)

[3.1 算法概述 26](#_Toc482520116)

[3.2 损失函数 28](#_Toc482520117)

[4. 基于连续帧的目标检测和计数方法的实现 29](#_Toc482520118)

[4.1 算法概述 29](#_Toc482520120)

[4.2 通过1x1卷积核来连接连续帧信息 29](#_Toc482520121)

[4.2.1 1x1卷积核 29](#_Toc482520122)

[4.2.2 设计思路和网络结构 31](#_Toc482520123)

[4.3 通过LSTM来连接连续帧信息 32](#_Toc482520124)

[4.3.1 设计思路和网络结构 32](#_Toc482520125)

[5. 本章小结 34](#_Toc482520126)

[第四章 实验结果与分析 35](#_Toc482520127)

[1. 数据集和评测标准 35](#_Toc482520128)

[1.1 数据集 35](#_Toc482520129)

[1.2 评测标准 35](#_Toc482520130)

[2. 实验设置 36](#_Toc482520131)

[3. 实验结果 36](#_Toc482520132)

[4. 实验结果分析 37](#_Toc482520133)

[5. Demo展示 38](#_Toc482520134)

[6. 本章小结 39](#_Toc482520135)

[第五章 总结和展望 40](#_Toc482520136)

[1. 本文工作总结 40](#_Toc482520137)

[2. 改进意见和展望 40](#_Toc482520138)

[参考文献 42](#_Toc482520139)

[致谢 43](#_Toc482520140)

# 第一章 绪论

## 研究背景和意义

### 研究背景

随着城市化进程的不断推进，人群聚集的现象越来越多，无论是在室内还是在室外，随处可见大群的人聚集在一起，与此同时也带来了诸多的安全隐患和社会压力。2004年2月5日，北京密云元宵灯展发生踩踏事故，37人死亡；2014年12月31日晚23时35分许，上海外滩陈毅广场发生群众拥挤踩踏事故，致35人死亡，至少48人受伤。不只是中国，几乎在世界各地，尤其是人口规模较大的一线城市及旅游城市，都发生过此类由于人群密集引发的安全事故。比如， 2014年6月29日，几内亚首都科纳克里附近一处海滩举行的一场音乐会上发生踩踏事件，造成至少24人死亡。2014年10月10日，巴基斯坦反对党正义运动党领导人伊姆兰·汗，当晚在木尔坦一个运动场举行政党集会。活动结束后，人们从运动场离开，因出口开放数量有限，民众相拥而出，发生踩踏事故，造成40余人伤亡。这些事故都是在人群密度过大、过度拥挤的条件下发生的，当出现人群聚集却缺乏有效的管理和控制时，一旦个体发生意外则可能引起群体的盲目恐慌，演变成群体事故。如果能够对人群及其运动特征进行统计分析，及时对存在安全隐患的密集人流进行合理的梳导，对可能出现的安全事故做到提前预警，将有效减少甚至避免伤亡和社会损失。

由于计算机硬件和软件技术的快速发展，视频监控系统作为治理城市安全的有效方式之一，具有很大的价值和意义。监控行业诞生于英、美等发达国家，促成其广泛应用的直接原因是 1993年和 1994年的两起恐怖袭击。监控摄像头能够对犯罪分子起到威慑作用；便于城市安全和管理人员事后取证；对重点区域实时监控，还可以有效地预防危害公共安全的事件发生。为保障城市安全，公安部启动了“科技强警”战略，以科技的手段提高执法监督效率，推进“和谐社会”的构建。城市社会治安视频监控系统是“科技强警”建设的重要组成部分，它可以对一些治安重点监控区域，如居民小区、城区路面、商业中心、娱乐场所、车站广场、重点单位、卡口等场所实施远程实时监控，及时了解现场的车流、人流及异常情况，并进行远程录像备份。城市社会治安视频监控系统作为公安集中指挥系统中不可缺少的子系统，可以实现派出所、分局、市局的多级监管，信息共享，起到社会治安综合治理的效果。据悉，不少省市的公安系统已与国内人工智能公司展开了业务上的合作，在目标检测、泛身份识别领域开展相应的项目，可能在不久的将来将会全面部署智能监控系统，保障城市安全，对危险人物和危险场所能做到提前监控和预警。

从 2006 年以来，在 Hilton、Bengio、LeChun 等人的引领下，大量深度神经网络的论文被发表，尤其是2012年，Hinton课题组首次参加ImageNet图像识别比赛，其通过构建的CNN网络AlexNet[1]一举夺得冠军，从此神经网络开始受到广泛的关注。深度学习利用多层计算模型来学习抽象的数据表示，能够发现大数据中的复杂结构，目前，这项技术已成功地应用在包括计算机视觉领域在内的多种模式分类问题上。计算机视觉对于目标运动的分析可以大致分为三个层次：运动分割，目标检测；目标跟踪；动作识别，行为描述[2]。其中，目标检测既是计算机视觉领域要解决的基础任务之一，同时它也是视频监控技术的基本任务。由于视频中的目标（或行人）具有不同姿态且经常出现遮挡、其运动具有不规则性，同时考虑到监控视频的景深、分辨率、天气、光照等条件和场景的多样性，故即使在技术发展的今天，目标检测这一基本任务仍然是非常具有挑战性的课题，存在很大的提升潜力和空间。而目标检测算法的结果将直接影响后续的跟踪、动作识别和行为描述的效果，也将直接影响计数方法的准确性

### 研究意义

目标检测是计算机视觉和数字图像处理的一个热门方向，广泛应用于机器人导航、智能视频监控、工业检测、航空航天等诸多领域。因此，目标检测也就成为了近年来理论和应用的研究热点，它是图像处理和计算机视觉学科的重要分支，也是智能监控系统的核心部分，同时目标检测也是泛身份识别领域的一个基础性的算法，对后续的人脸识别、步态识别、人群计数等任务起着至关重要的作用。它的目的就是如何快速、准确地检测出监控视频中的目标物体或人物，即从序列图像中将运动目标提取出来。

人群计数作为智能视频监控系统的重要课题之一，具有重要的社会意义。对高密度人群的智能检测和计数，能够减少对人力资本的消耗，预防人群安全事故，有助于保障重点区域（比如：广场、体育场、人行道、机场等）人群的人身安全。如：可以通过统计等候电梯的人数来优化调度电梯，以此提高电梯的利用率，减少用户的等待时间。可以通过统计经过十字路口、丁字路口人群流动繁忙的交通场合的人数，可以合理安排交通警察或保安人员的工作时间和工作额度。此外，大多数的人群状态模型以及其他更高层次的群体行为研究也是建立在人群计数的基础之上，因此，人群计数也有着不可忽视的研究价值。

## 主要研究内容

本文在广泛调研国内外目标检测和人群计数方法的基础上，通过考虑视频帧的连续帧信息，即时序信息，设计并实现了面向拥挤场景视频的基于连续帧信息的目标检测和人群计数方法。

本文的研究内容和主要成果包括：

（1）独自标注了拥挤场景视频帧2350帧（2fps，采集于webcam），其中2000张每秒标注两张图片作为训练集，350张每隔5s标注一张图片作为测试集，训练集和测试集取自不同时间段；与此同时，由于数据集较小，探索了模拟光照、模糊处理、噪声处理、颜色抖动等一系列数据增强方法，效果较为逼真，在一定程度上解决了数据量少的问题；

（2）设计并实现了基于连续帧信息的人物检测和计数方法的一系列模型。本文提出了两种有效连接时序信息的方式，一种是沿时序卷积，通过1x1的卷积核将连续帧的feature map进行线性组合；另一种是使用LSTM对当前帧进行预测，最后将预测得到的结果与当前帧的feature map进行线性组合；

（3）在标注数据集上进行训练测试，上述提出的两种方法均优于单独使用当前帧进行检测和计数，实验证明了连续帧信息对当前帧的目标检测和计数方法效果均有一定程度的提升。

# 第二章 相关研究现状

## 传统目标检测算法

### 算法概述

传统目标检测的方法一般分为三个阶段：首先在给定的图像上选择一些候选的区域，然后对这些区域提取特征，最后使用训练的分类器进行分类。下面我们对这三个阶段分别进行介绍。

(1) 区域选择：这一步是为了对目标的位置进行定位。由于目标可能出现在图像的任何位置，而且目标的大小、长宽比例也不确定，所以最初采用滑动窗口的策略对整幅图像进行遍历，而且需要设置不同的尺度，不同的长宽比。这种穷举的策略虽然包含了目标所有可能出现的位置，但是缺点也是显而易见的：时间复杂度太高，产生冗余窗口太多，这也严重影响后续特征提取和分类的速度和性能。

(2) 特征提取：由于目标的形态多样性，光照变化多样性，背景多样性等因素使得设计一个鲁棒的特征并不是那么容易。然而提取特征的好坏直接影响到分类的准确性。其中，这个阶段常用的特征有SIFT[3]、HOG[4]等。

(3) 分类器：主要有SVM, Adaboost等。

## 基于Region Proposal的深度学习目标检测算法



### 算法概述

传统目标检测存在的两个主要问题：一个是基于滑动窗口的区域选择策略没有针对性，时间复杂度高，窗口冗余；二是手工设计的特征对于多样性的变化并没有很好的鲁棒性。对于滑动窗口存在的问题，region proposal提供了很好的解决方案。region proposal（候选区域）是预先找出图中目标可能出现的位置。但由于region proposal利用了图像中的纹理、边缘、颜色等信息，可以保证在选取较少窗口（几千个甚至几百个）的情况下保持较高的召回率。比较常用的region proposal算法有Selective Search[5]和卷积神经网络，有了候选区域，剩下的工作实际就是对候选区域进行图像分类的工作。

### 代表性算法之R-CNN

首先输入一张图片，我们先通过selective search定位出2000个左右物体候选框，然后采用CNN网络提取每个候选框中图片的特征向量，特征向量的维度为4096维，接着采用svm算法对各个候选框中的物体进行分类识别。R-CNN[6]将深度学习引入检测领域，一举将PASCAL VOC上的检测率从35.1%提升到53.7%，相比于传统算法有了较大的提升。下图为算法流程：

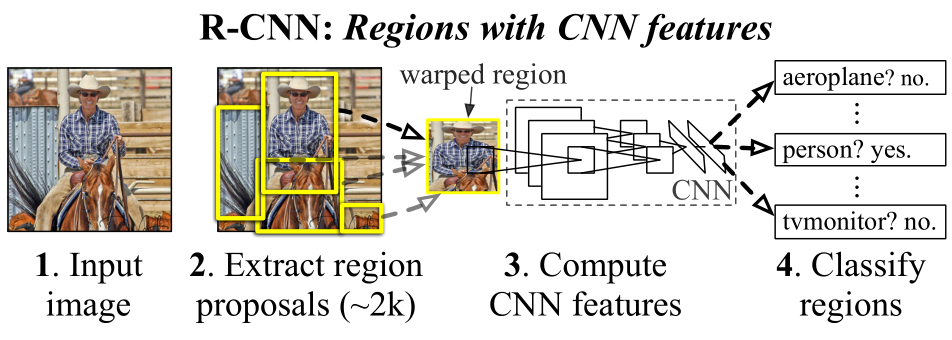


图2-1 R-CNN算法流程

### 代表性算法之Fast R-CNN

整体框架如图2-2，大致的训练过程可以理解为：首先通过selective search在一张图片中得到约2k个物体候选框；同时，将整张图像归一化后直接送入深度网络得到对应的feature map，然后加入物体候选框信息，对于每个物体候选框求取映射关系，在feature map上crop出对应的patch，并用一个单层的SPP layer（这里称为roi pooling layer）来统一到一样的尺度，即这些候选区域的前几层特征不需要再重复计算；最后，将roi pooling之后得到的feature map输入到后续的fc层，最后一层输出类别相关信息和4个bounding box的修正偏移量。

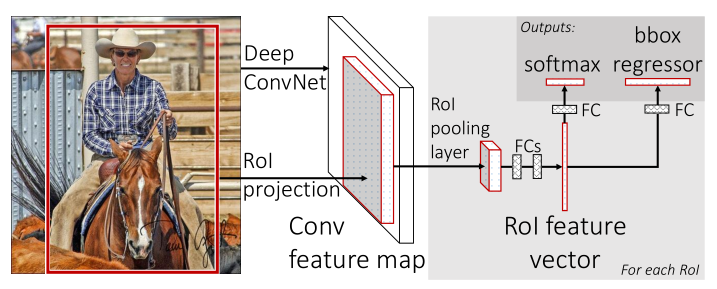


图2-2 Fast R-CNN[7]算法流程

### 代表性算法之Faster R-CNN

Faster R-CNN[8]可以简单地看做“区域生成网络+Fast R-CNN“的系统，用区域生成网络代替Fast R-CNN中的Selective Search方法。主要亮点是“区域生成网络”和“让区域生成网络和fast RCNN网络共享特征提取网络”，下面将进行逐一介绍：

（1）区域生成网络：基本设想是在提取好的特征图上，对所有可能的候选框进行判别。由于后续还有位置精修步骤，所以候选框实际比较稀疏。其中特征提取包含若干层conv+relu，直接套用ImageNet上常见的分类网络即可（常见的有VGG，GoogleNet等），额外添加一个conv+relu层，输出51\*39\*256维特征。接下来是候选区域，特征可以看做一个尺度51\*39的256通道图像，对于该图像的每一个位置，考虑9个可能的候选窗口：三种面积×三种比例，这些候选窗口称为anchors。对于每一个位置（anchor）来说，分类层从256维特征中输出属于前景和背景的概率；窗口回归层从256维特征中输出4个平移缩放参数。

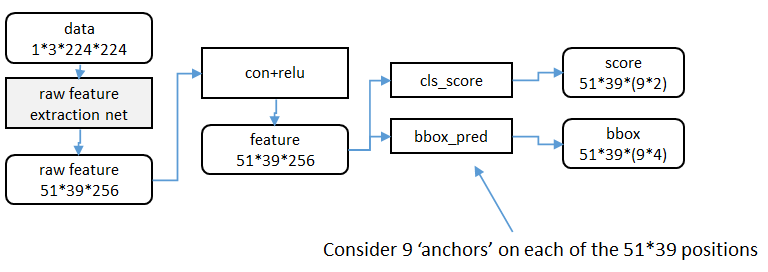


图2-3 区域生成网络结构

（2）特征提取网络共享：区域生成网络（RPN）和目标检测网络（Fast R-CNN）都需要一个原始特征提取网络，因此可以使用同一个特征进行特征提取，这样可以减少网络参数，同时可以提高Faster R-CNN的速度，即只需要进行一次特征提取即可。如图2-4所示，使用同一个特征提取网络进行特征提取，然后使用区域生成网络生成物体候选框，最后将候选框信息输入到目标检测网络。

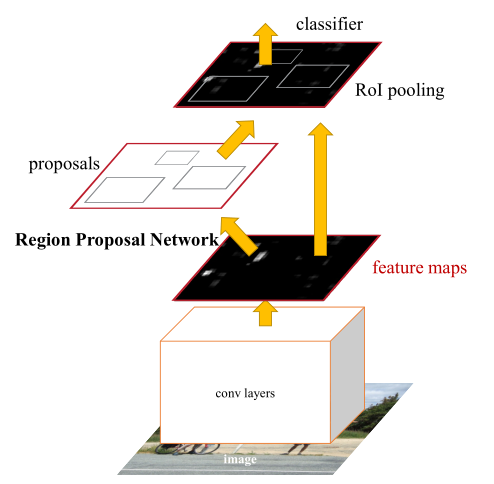


图2-4 Faster R-CNN算法流程

## 基于回归方法的深度学习目标检测算法

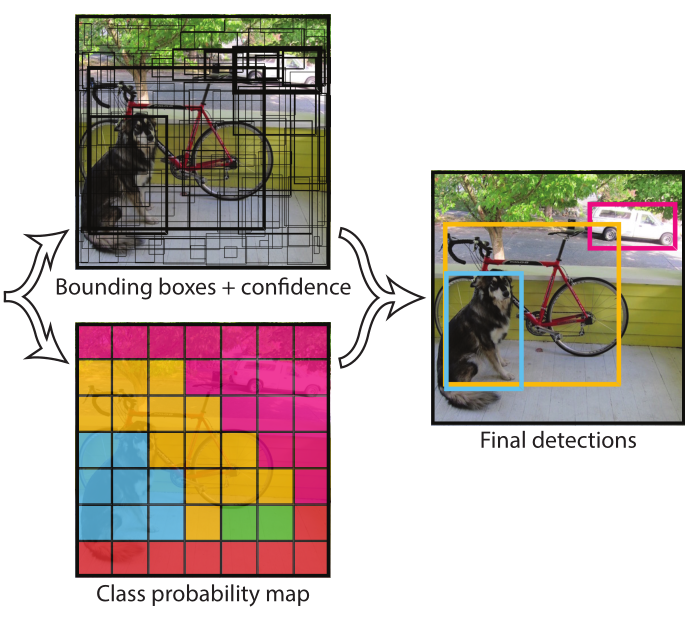


### 算法概述

Faster R-CNN的方法目前是主流的目标检测方法，但是速度上并不能满足实时的要求。YOLO一类的方法慢慢显现出其重要性，这类方法使用了回归的思想，既给定输入图像，直接在图像的多个位置上回归出这个位置的目标边框以及目标类别。

### 代表性算法之YOLO

YOLO[9]将输入图像分成SxS个格子，每个格子负责检测落入该格子的物体。若某个物体的中心位置的坐标落入到某个格子，那么这个格子就负责检测出这个物体。如下图所示，最左边图中物体狗的中心点（红色原点）落入第5行、第2列的格子内，所以这个格子负责预测图像中的物体狗。



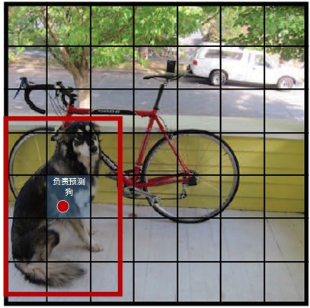


图2-5 YOLO算法流程

YOLO 具体算法流程： 给一个输入图像，首先将图像划分成7\*7的网格，其中对于每个网格，我们都预测2个边框（包括每个边框是目标的置信度以及每个边框区域在多个类别上的概率），最后根据上一步可以预测出7\*7\*2个目标窗口，然后根据阈值去除可能性比较低的目标窗口，最后NMS去除冗余窗口即可，具体网络结构如图2-6。

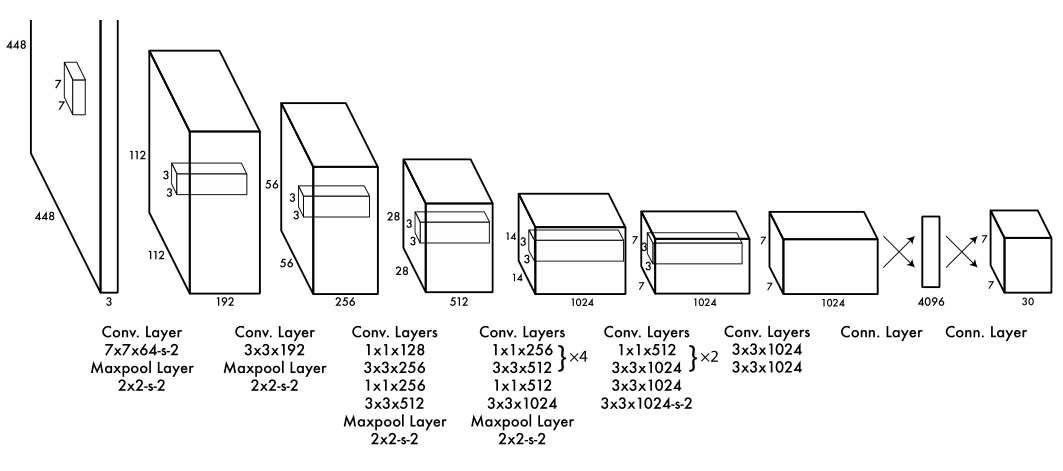


图2-6 YOLO网络结构

YOLO将目标检测任务转换成一个回归问题，大大加快了检测的速度，使得YOLO可以每秒处理45张图像。而且由于每个网络预测目标窗口时使用的是全图信息，使得false positive比例大幅降低（充分的上下文信息）。但是YOLO也存在问题：没有了region proposal机制，只使用7\*7的网格回归会使得目标不能非常精准的定位，这也导致了YOLO的检测精度并不是很高。

### 代表性算法之SSD

SSD[10]采用VGG16的基础网络结构，使用前面的前5层，然后利用astrous算法将fc6和fc7层转化成两个卷积层。再格外增加了3个卷积层，和一个average pool层。不同层次的feature map分别用于default box的偏移以及不同类别得分的预测，最后通过nms得到最终的检测结果。

如图2-7所示，假如某一层特征图（图b）大小是8\*8，那么就使用3\*3的滑窗提取每个位置的特征，然后这个特征回归得到目标的坐标信息和类别信息。对于4x4的特征图处理方法相同（图c）。

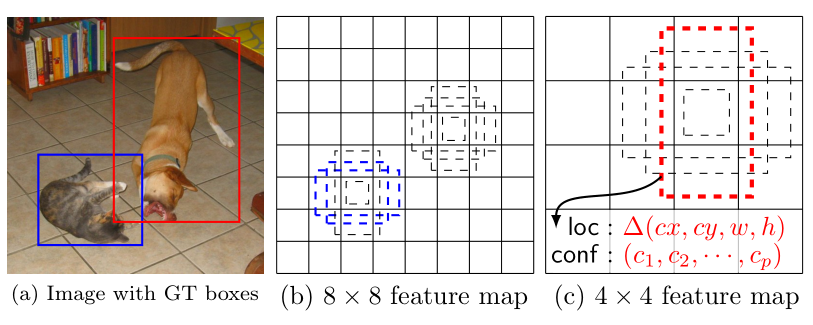


图2-7 SSD框架示例

SSD增加的卷积层的 feature map 的大小变化比较大，允许能够检测出不同尺度下的物体： 在低层的feature map,感受野比较小，高层的感受野比较大，在不同的feature map进行卷积，可以达到多尺度的目的。观察YOLO，后面存在两个全连接层，全连接层以后，每一个输出都会观察到整幅图像，并不是很合理。但是SSD去掉了全连接层，每一个输出只会感受到目标周围的信息，包括上下文。这样来做就增加了合理性。并且不同的feature map,预测不同宽高比的图像，这样比YOLO增加了预测更多的比例的box。

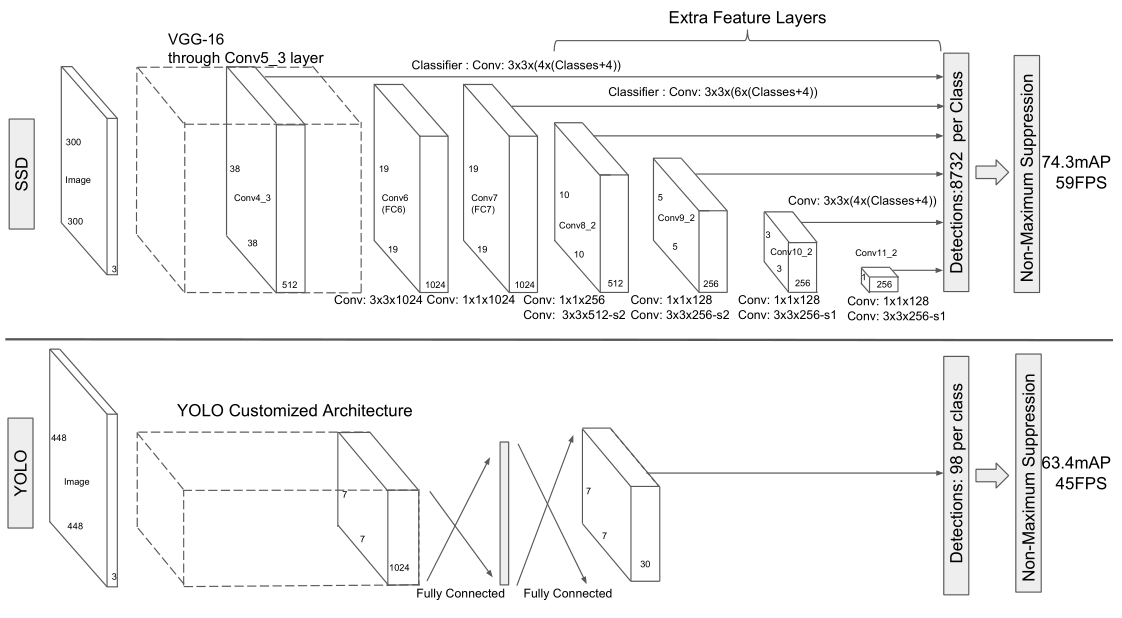


图2-8 SSD & YOLO网络结构比较

SSD结合了YOLO中的回归思想和Faster R-CNN中的anchor机制，使用全图各个位置的多尺度区域特征进行回归，既保持了YOLO速度快的特性，也保证了窗口预测的跟Faster R-CNN一样比较精准。SSD在VOC2007上mAP可以达到72.1%，速度在GPU上达到58帧每秒。总之，YOLO的提出给目标检测一个新的思路，SSD的性能则让我们看到了目标检测在实际应用中真正的可能性。

## 本章小结

本章主要介绍了目标检测的相关研究现状，首先介绍了基于传统的目标检测算法，接下来介绍了基于Region Proposal的深度学习目标检测算法并简要介绍了代表性的算法R-CNN、Fast R-CNN以及Faster R-CNN，最后介绍了基于回归方法的深度学习目标检测算法并简要介绍了代表性算法YOLO和SSD。

# 第三章 具体实验

## 实验理论基础

### GoogleNet

目前，卷积神经网络的发展趋势是增加网络层数和每层网络的大小，以提升其特征表达效果。然而，这一做法带来的直接问题就是快速增长的计算量和过拟合问题，而且如果大量权重的训练结果趋于0，则表明训练过程中浪费了大量的计算资源。这两个问题根本的解决办法是将全连接的网络结构变为稀疏连接，当然，仅仅改变卷积神经网络的全连接层远远不够，必须要在卷积层也做出这样的结构改变。另一方面，若从底层硬件的角度考虑，CPU和GPU擅长密集矩阵乘法运算，而对非一致（non-uniform）稀疏数据结构上的计算效率较低。LeCun、Bengio[11]等人在早期提出的传统做法是在卷积神经网络中使用随机的稀疏连接打破网络结构的对称性，以提高学习效率；而Krizhevsky[12]又重新使用全连接的结构，利用密集计算的高效性，采用更好的参数优化方法。

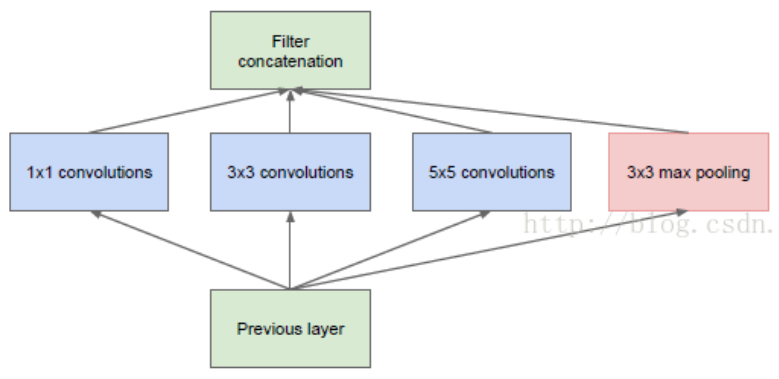


图3-1 一般的 inception 结构

GoogLeNet[13]提出了一种新的网络模块（inception module），它既能保持网络结构的稀疏性，又能利用密集矩阵的高计算性能。Inception的主要思想是用稠密的模板覆盖卷积神经网络中最优的局部稀疏结构，使之能够近似地表示原来的局部稀疏结构。Arora等人对网络结构的最后一层做相关性统计，将相关性高的聚合成簇，作为下一层的单元，并与上一层相连。较低的网络层（靠近输入图片）中的单元集中在某一些局部区域，也就是说，最终在单个区域内会有大量簇，可以在下一层通过 的卷积过滤器覆盖，当然，也可以用更大尺寸的卷积过滤器来覆盖更大的区域，从而减少过滤器的数量。图 3-2所示的 inception 结构采用了1x1，3x3和5x5的卷积过滤器，并额外地添加一个可选的池化操作，以增强效果。

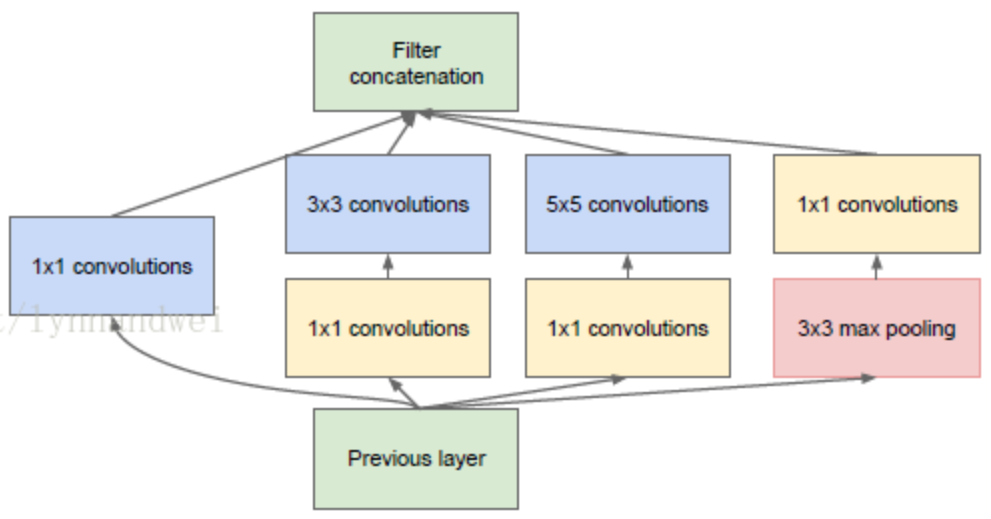
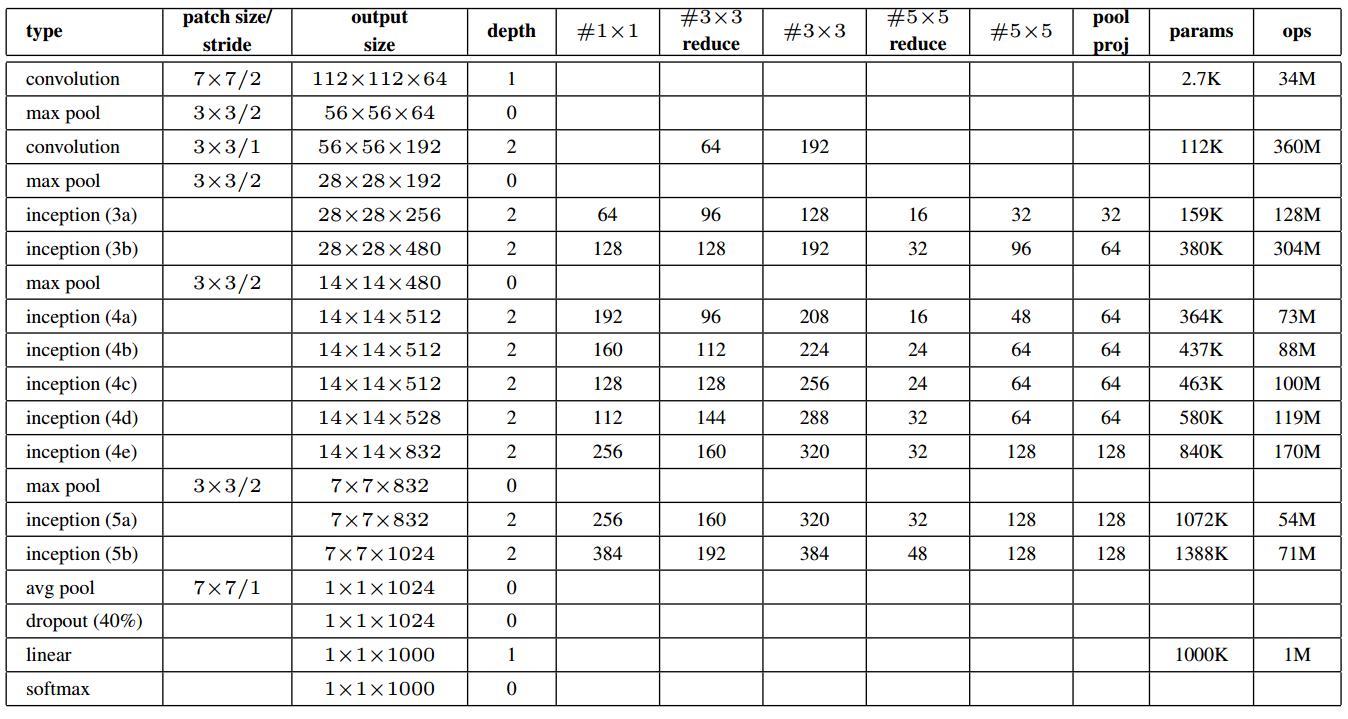


图3-2 GoogLeNet 中的 Inception

然而这种结构最大的问题是由于大量的滤波器存在，即使最后几个网络层中的卷积滤波器数量适中，计算量也会非常大。如果用这种结构覆盖最优稀疏结构，计算效率将非常低下。所以，GoogLeNet 采用的 inception 模块在上面的基础上做出了改进，在计算量较大的3x3和5x5的卷积过滤器之前，先用1x1的卷积过滤器降维，并修正线性特性。

GoogLeNet一共约为100层，其中带参数的层有22层，全连接层输出1024位的向量。对于浅层网络来说，GoogLeNet相对中间的层产生的特征非常具有辨识力。于是GoogLeNet的设计者在训练网络时，在一些中间层上增加了额外的分类器，以增强反向传播时的梯度信号和正则化，而在测试时，再将这些分类器去掉。GoogLeNet在ILSVRC 2014比赛中同时取得了分类任务和检测任务的第一名。

表格3-1 GoogLeNet网络结构说明



### RNN和LSTM

1. **循环神经网络（RNN）**

人类并不是每时每刻都从一片空白的大脑开始他们的思考。在你阅读这篇文章时候，你都是基于自己已经拥有的对先前所见词的理解来推断当前词的真实含义。我们不会将所有的东西都全部丢弃，然后用空白的大脑进行思考。我们的思想拥有持久性。传统的神经网络并不能做到这点，看起来也像是一种巨大的弊端。例如，假设你希望对电影中的每个时间点的时间类型进行分类。传统的神经网络应该很难来处理这个问题——使用电影中先前的事件推断后续的事件。RNN 解决了这个问题。RNN 是包含循环的网络，允许信息的持久化。循环神经网络一次处理一个输入序列元素，同时维护网络中隐式单元中包含的过去时刻序列元素的历史信息的“状态向量”，可以对序列信息建模，更加强大，也更加符合人类大脑皮层对信息的处理方式。

在训练过程中，前馈神经网络对网络层的误差求导以调节前一层的权重参数，而由于循环神经网络中的神经元没有分层、未形成有向非循环图，所以反向传播算法会根据神经元自身进行误差求导，权重学习变得十分复杂。所以，循环神经网络在训练时采用的策略是将循环神经网络在序列上“展开”，使之在形式上变成一个前馈神经网络。图 3-3 将一个简单的循环神经网络在时间序列上展开成前馈神经网络的示意图。

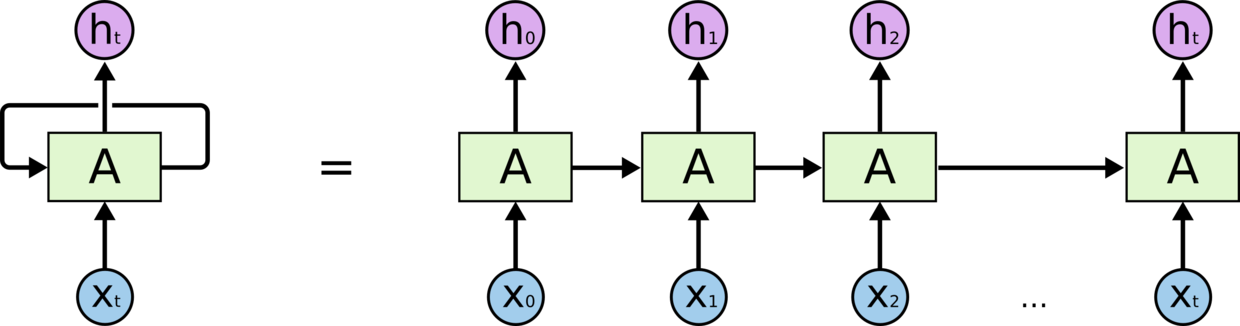
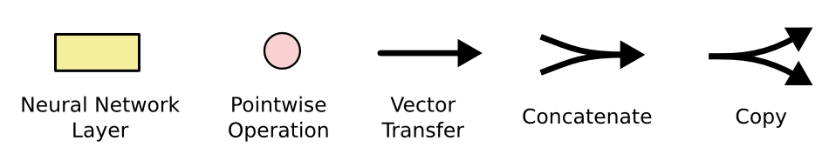


图3-3 循环神经网络的展开形式

1. **长短期记忆网络（LSTM）**

长短期记忆网络（Long-Short Term Memory networks, LSTMs）是一种特殊的循环神经网络，它独特的结构设计可以处理上文提到的梯度消失问题，有能力学习长时间的依赖关系。长短时间记忆网络由Hochreiter和Schmidhuber[14]于1997年首次提出，引入“门”的概念，用输入门和输出门控制网络单元的状态；1999 年Gers和 Schmidhuber[15]在网络中加入了遗忘门，使其功能更加完善；此后有很多科研人员尝试优化这种网络，使其能够在更多任务中有出色的表现。

长短期记忆网络的核心思想是记忆细胞（Memory Cell），它的读取、写入和保存由输入门（Input Gate）、输出门（Output Gate）和遗忘门（Forget Gate）这三个特殊的逻辑单元控制。记忆细胞是一个线性的神经元，与神经网络的其他部分相连，也与自身连接。记忆细胞的自连接，也被叫做常数误差传送带（Constant Error Carousel, CEC），它的激活值（avtivation，即激活函数的输出）即为细胞状态（Cell State），正是它解决了循环神经网络训练时的梯度消失问题，从而使学习长期依赖关系变成可能。门由含有激活函数（一般是sigmoid函数）的网络层和向量运算组成，负责在该记忆细胞与外界及自身的连接处设定权值，控制它所在的长短时间记忆单元接纳或者摒弃序列信息，而不影响神经网络的其他单元。顾名思义，输入门决定是否把当前的输入信息存进记忆细胞，遗忘门决定把怎样的历史信息从记忆细胞中抹去，输出门决定将记忆细胞中的哪一部分信息从记忆单元输出。与一般的循环神经网络类似，长短期记忆网络仍然可以表示为链式展开的形式，图3-4中3个绿色矩形是重复单元，它们内容完全一致，随着序列重复出现。



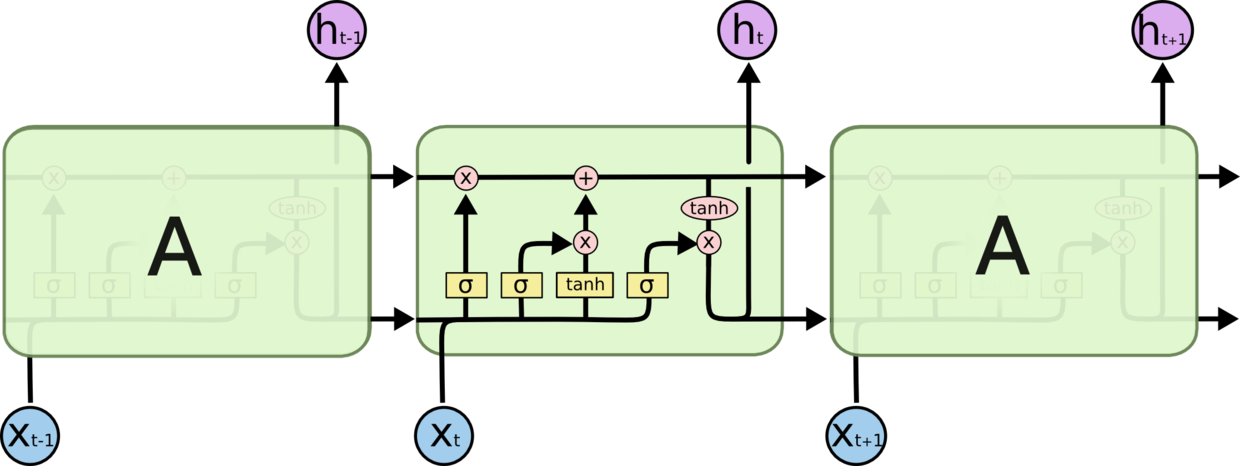


图3-4 长短期记忆网络

### 实验理论基础

本文实验的理论基础主要有以下几点：

（1）GoogLeNet 在 ILSVRC 2014 比赛中同时取得了分类任务和检测任务的第一名。由此可见，GoogLeNet能够较好的提取图片的特征，本文实验同样采用GoogLeNet来提取图片的特征；

（2）循环神经网络灵活而复杂，与非循环神经网络相比，它的结构更加符合人类大脑的神经网络结构。循环神经网络能够对序列信息进行操作，具有强大的功能和鲁棒性，在近几年的各种研究中也表现出不可思议的有效性。使用LSTM进行目标检测，可以充分利用检测上下文的信息，循环输出检测到的目标物体或人物；

（3）视频的连续帧信息在一定程度上反应了物体的运动状态，通过观察物体的运动状态，可以准确的识别出物体。且标注的视频帧的场景非常复杂和分辨率较低，需要根据前后帧的信息才能够完成准确标注。因此，从人为观察的角度可以发现，连续帧信息对当前帧的信息有一定的增益作用；

（4）通过沿时序卷积可以将连续帧信息以不同权重组合起来，通过训练过程产生一种最佳的线性组合方式。组合了连续帧信息之后的feature map能考虑到连续帧的信息，可以进行更好的检测；

（5）LSTM即长短期记忆模型，能够较为充分的利用时序信息，连续帧的信息通过使用LSTM预测当前帧的特征表示，最后将得到的预测特征图与当前帧的特征图进行线性组合，即考虑了时序信息，对当前帧的检测起到了一定的帮助；

（6）本实验的计数方法是在目标检测的基础上进行的，即通过计算检测框的数量进行人群计数，所以通过改善目标检测的结果在某种程度上也能改变最终的人群计数效果。

## 前期数据准备



### 数据采集和标注

训练集和测试集采集于网络摄像头的不同时间段，场景为菜市场。其中训练集采集了2000张，帧率为2fps，全部进行标注；测试集采集了3500张，帧率为2fps，每隔10帧标注一帧。即可用的训练集为2000张，测试集为350张，示例如下图：



图3-5 标注视频帧示例

标注的是人物的人头区域，可以看出视频帧的场景较为复杂，人头区域较小，分辨率较低且人物重叠现象较为严重。综合以上因素，标注的时候较为困难，有时候需要参考前后帧的信息才能准确判断出人头区域并进行标注。

### 数据增强

数据增强即对基础数据进行一系列操作，例如：放大，翻转，加噪声等，就得到了更多的数据。同时也增强了数据的多样性。一方面增强了数据在训练中的作用，另一方面也在一定程度上避免了过拟合，使得模型更加鲁棒。接下来我将介绍几种自己整理探索的数据增强方法：

（1）颜色抖动：即通过随机改变图片的颜色、对比度和亮度来实现图片的颜色变换，效果如下图：



图3-6 分别为原图、变换1、变换2

（2）模拟光照：即通过从上到下或者从左到右逐渐改变图片的亮度的方式来模拟四个方向的光照，效果如下图：



图3-7 分别为原图、变换1、变换2

（3）旋转抖动：通过对图像随机进行一定角度的旋转，即对应像素点进行相应的位置变换，来实现旋转抖动，效果如下图：



图3-8 分别为原图、变换1、变换2

（4）噪声处理：通过随机增加黑点和白点来添加椒盐噪声，通过对像素点的像素值进行高斯扰动来添加高斯噪声，效果如下图：



图3-9 分别为原图、椒盐噪声、高斯噪声

（5）镜像翻转，即对图像做简单的镜像操作或者翻转操作，效果如下图：



图3-10 分别为原图、镜像、翻转

（6）模糊处理：通过对图片进行反复resize处理，即扩大缩小，通过随机变换次数来得到不同的模糊程度，通过验证表明，多次resize的效果要比一次大尺度resize的效果逼真很多，效果如下图：



图3-11 分别为原图、变换1、变换2

## 基于当前帧的目标检测和计数方法的实现



### 算法概述

近几年，由于深度学习的兴起，不少学者利用卷积神经网络提取特征，进而用 softmax 或者 SVM 进行分类，再通过回归调整定位，从而完成检测。由于深度神经网络的结构和训练过程较为符合人类大脑的结构和自发的学习过程，与传统方法提取的特征相比，深度神经网络提取的特征具有更强的描述能力。为了克服前文提到的行人检测的困难，获得更好的检测效果，考虑利用深度神经网络GoogLeNet 作为行人检测的特征提取方法。另外，以往的目标检测方法都是通过滑动窗口或者预选区域机制生成大量矩形框（候选结果），产生检测结果的过完备集，再做某种形式的合并或者非极大值抑制等后处理得到最终检测结果。

非极大值抑制和其他合并矩形框的算法仅仅从矩形框的性质（比如：距离，重叠等）判断是否保留该矩形框或者将该矩形框与其他矩形框合并，而完全没有用到图片本身的信息，这是后处理的明显缺陷，然而，几乎所有的目标检测方法都用到了这种后处理算法。对于目标单独出现的图片，这种方法具有较好的效果，但对于多个目标大量出现在同一张图中，并且他们相互遮挡时，经过后处理算法得到的检测结果的准确率和召回率都有较大损失。所以，当目标相互遮挡时，利用图片信息决定矩形框的位置和个数显得十分必要。

Stewart和Andriluka[16]提出的端到端的目标检测方法，避免了非极大值抑制的过程，在遮挡较为严重场景中，取得了较好的检测效果，网络结构如下图所示。

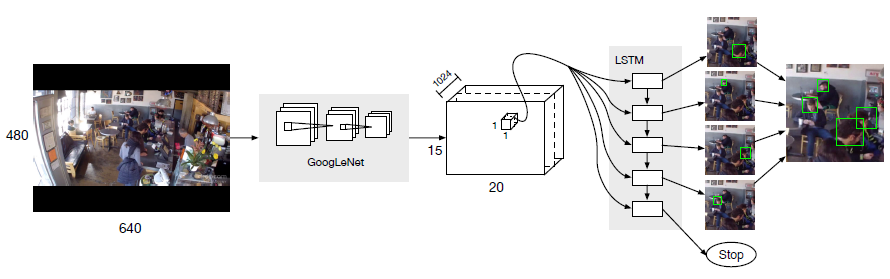


图3-12 端到端的目标检测方法的网络结构

该网络结构将卷积神经网络与循环神经网络相连，其中卷积神经网络用于提取具有丰富描述能力的特征，循环神经网络作为检测子，生成长度（对应图片中行人个数）可变的行人检测结果。行人检测的过程可以理解为：先用卷积神经网络对图像编码，生成包含丰富信息的特征描述子；然后用循环神经网络解码描述子，生成一个矩形框（行人检测结果）的集合。该网络结构使用长短时间记忆单元组成循环神经网络，作为预测检测结果的核心机制。

卷积神经网络GoogLeNet将整个图片划分为 300（15x20）个cell，每个cell对应原图一个固定大小的感受野。GoogLeNet将每个cell编码为一个1024维的高层特征描述子，此向量刻画了感受野中的图片内容，包含了目标位置在内的丰富信息。每个cell都与一个与之对应的长短期记忆网络相连，因此，对于整张图片，一共存在300个平行的LSTM。

长短期记忆网络以GoogLeNet生成的1024维的特征向量作为输入，每次循环将解码一个检测结果，其中表示矩形框的坐标和尺寸，表示检测结果的分数，即置信度，置信度值越大，表示当前检测结果越可靠。测试时，当长短时间记忆模型输出的检测结果的置信度小于阈值（比如：0.5）时，忽视这次检测结果。最后，根据300个cell的感受野的位置，将300个长短时间记忆网络的检测结果缝合在一起，得到最终的行人检测结果。

### 损失函数

首先，将行人检测的标注集（set of ground truth）记作。为了调节网络权重，在训练过程中引入了匈牙利匹配算法，建立从标注集到候选检测结果集的单射函数，例如，表示标号为*i*的标注矩形框对应的候选检测结果的标号。在匈牙利匹配算法中，我们根据人体检测的任务需要，定义标注结果与候选检测结果的比较函数为：

函数 返回一个三元组，其中表示惩罚因子，当候选检测结果的矩形中心未落在标注集中任何标注矩形框内时，即没有标注矩形与候选结果匹配，取 1，否则取 0，表示在长短期记忆单元输出的候选检测结果序列中的排序，是两个矩形框的范式距离。当评价两个候选检测结果时，我们把由比较函数产生的两个三元组按照字典序比较：先根据两个三元组的值排序，如果值相同时，再比较，以此类推。

已知f，定义损失函数L如下：

其中，用 计算标注矩形框到对应的候选检测结果的距离，表示候选检测结果置信度（即该候选检测结果是否能够匹配到标注集中的矩形）的交叉熵损失。定义为，即当候选检测结果在标注集有与之匹配的标注时，为1。损失函数中的系数用于平衡定位偏差和置信度误差。

## 基于连续帧的目标检测和计数方法的实现



### 算法概述

到目前为止，几乎主流的目标检测方法都只是考虑当前被检测图片的信息，其中主要是因为如下几个原因：

（1）当前开放的目标检测相关数据集大部分没有标注连续的视频帧，即大部分图片都是相互独立的。部分Tracking相关的数据集虽然标注了连续的视频帧，但是由于是用来做Tracking相关的算法研究，大部分图片没有将所有的目标进行标注，所以不能直接用来做目标检测。由于数据集的缺失，部分基于连续帧信息的目标检测的研究可能无法正常进行；

（2）到目前为止，没有较好的连接连续帧信息的方法，本文到目前为止尝试了很多种连接连续帧信息的方式，如串联、线性组合、使用LSTM等等。最终实验表明，使用1x1卷积核来实现连续帧信息的线性组合和使用LSTM来预测当前帧特征并与当前帧特征进行线性组合两种方式取得了较好的结果。

算法主要思路为：输入连续的视频帧（九帧，当前帧和连续的前八帧），对每一帧用GoogLeNet网络提取特征，然后使用不同的方式来连接连续帧的信息，即通过不同的方式组合连续帧的特征。最后使用与只使用当前帧进行目标检测同样的策略进行当前帧的目标检测，接下来将进行具体的描述。

### 通过1x1卷积核来连接连续帧信息

1. **1x1卷积核**

1x1卷积核即size为1x1的卷积核，1×1的卷积层（可能）引起人们的重视是在NIN[17]的结构中，论文中林敏师兄的想法是利用MLP代替传统的线性卷积核，从而提高网络的表达能力。文中同时利用了跨通道pooling的角度解释，认为文中提出的MLP其实等价于在传统卷积核后面接cccp层（cccp层等价于1×1卷积层），从而实现多个feature map的线性组合，实现跨通道的信息整合。CNN里的卷积大都是多通道的feature map和多通道的卷积核之间的操作（输入的多通道的feature map和一组卷积核做卷积求和得到一个输出的feature map），如果使用1x1的卷积核，这个操作实现的就是多个feature map的线性组合，可以实现feature map在通道个数上的变化。

除此之外，1x1卷积核能够进行通道数的降维和升维。在GoogLeNet里，对于每一个Inception模块（如下图3-13），原始模块是左图，右图中是加入了1×1卷积进行降维的。虽然左图的卷积核都比较小，但是当输入和输出的通道数很大时，乘起来也会使得卷积核参数变的很大，而右图加入1×1卷积后可以降低输入的通道数，卷积核参数、运算复杂度也就跟着降下来了。GoogLeNet利用1×1的卷积降维后，得到了更为紧凑的网络结构，虽然总共有22层，但是参数数量却只是8层的AlexNet的十二分之一。

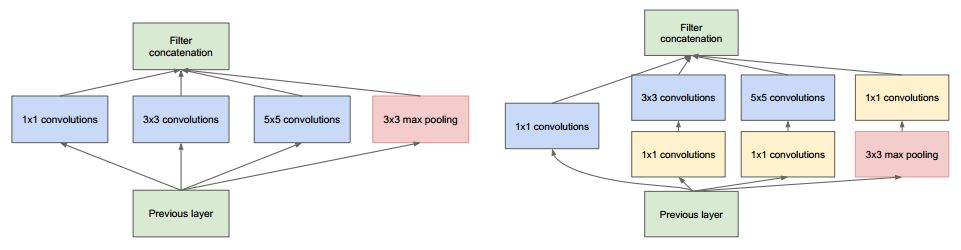


图3-13 左图为一般的Inception模块，右图为GoogLeNet中的Inception模块

ResNet[18]同样也利用了1×1卷积，并且是在3×3卷积层的前后都使用了，不仅进行了降维，还进行了升维，使得3x3卷积层的输入和输出的通道数都减小，参数数量进一步减少，如下图的结构。

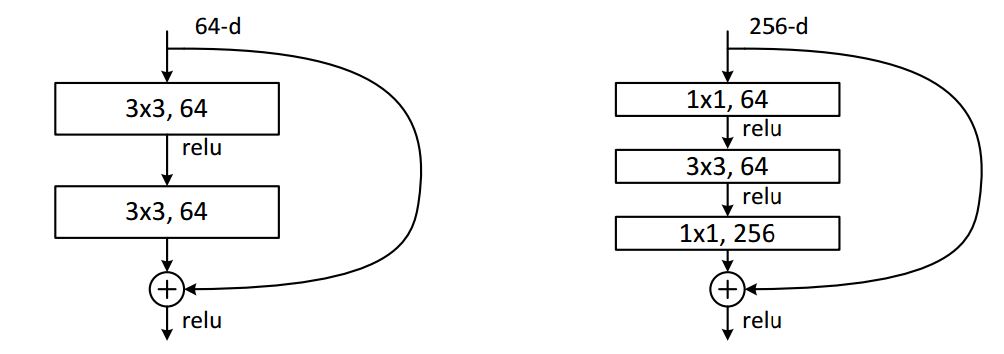


图3-14 ResNet结构示例

1. **设计思路和网络结构**

首先输入连续的视频帧九帧，分别用GoogLeNet网络对图像编码，生成包含丰富信息的高层特征描述子，与本章3.1小节相同，卷积神经网络GoogLeNet仍然将整个图片划分为 300（20x15）个感受野相互重叠的cell，将每个cell编码为一个1024维的高层特征描述子，此向量描述了感受野中的图片的丰富信息。

然后使用1x1卷积核（具体介绍见本章4.2.1）对每张图片经过编码生成的高层特征进行1x1的卷积操作，将卷积生成的高层feature map直接进行相加即可，参数数量为9x1x1。具体实现为沿batch\_size进行卷积，将batch\_size与channels互换维度，后面连接上1x1的卷积层，输入通道数为9，即连续帧的帧数，输出的通道数为1，这样就能实现不同帧的feature map的线性组合。

最后，将1x1卷积操作之后的batch\_size和channels互换维度，即得到的feature map仍然是将20x15x1024，即将整个图片划分为 300（20x15）个感受野相互重叠的cell，将每个cell编码为一个1024维的高层特征描述子，每个cell都与一个与之对应的长短期记忆网络LSTM相连，每个LSTM的长度为5，即输出5个检测目标和对应的置信度。

具体网络结构分为参数共享和参数不共享两种特征提取方式，首先介绍参数共享的网络结构：连续的九帧图片使用同一个GoogLeNet网络来提取特征，即打包成一个batch输入到同一个GoogLeNet网络，这样可以大幅度减少网络的参数量，如下图3-15所示。

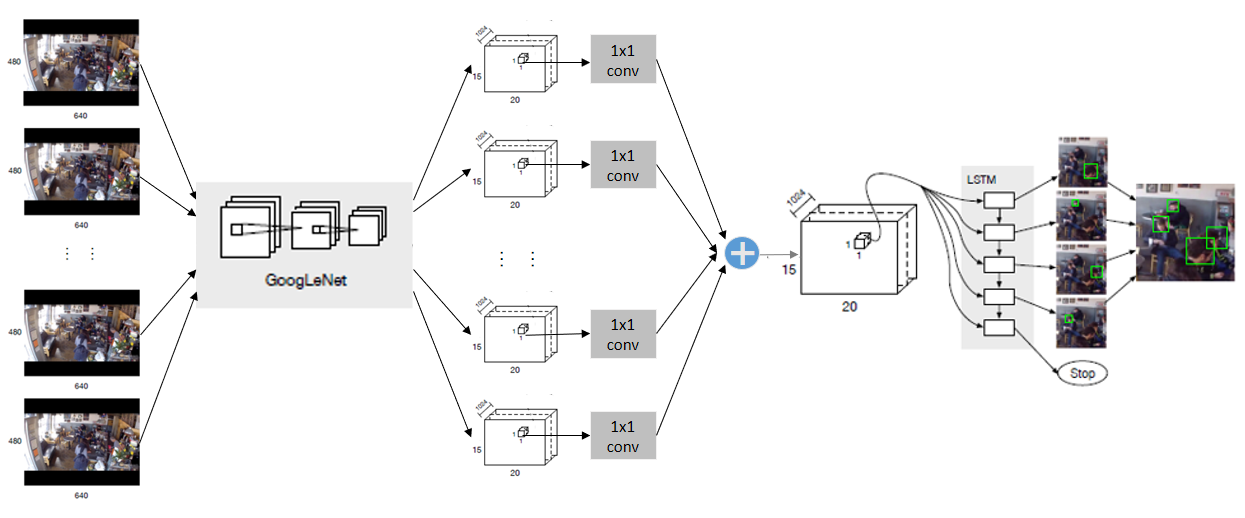


图3-15 1x1卷积连接连续帧信息网络结构（参数共享）

接下来是参数不共享的情况，网络结构如图3-16所示，每一帧图片都输入到一个不同的GoogLeNet网络中提取特征，这样GoogLeNet网络就相当于复制了九份，提取特征这一部分参数数量相应增加了9倍。

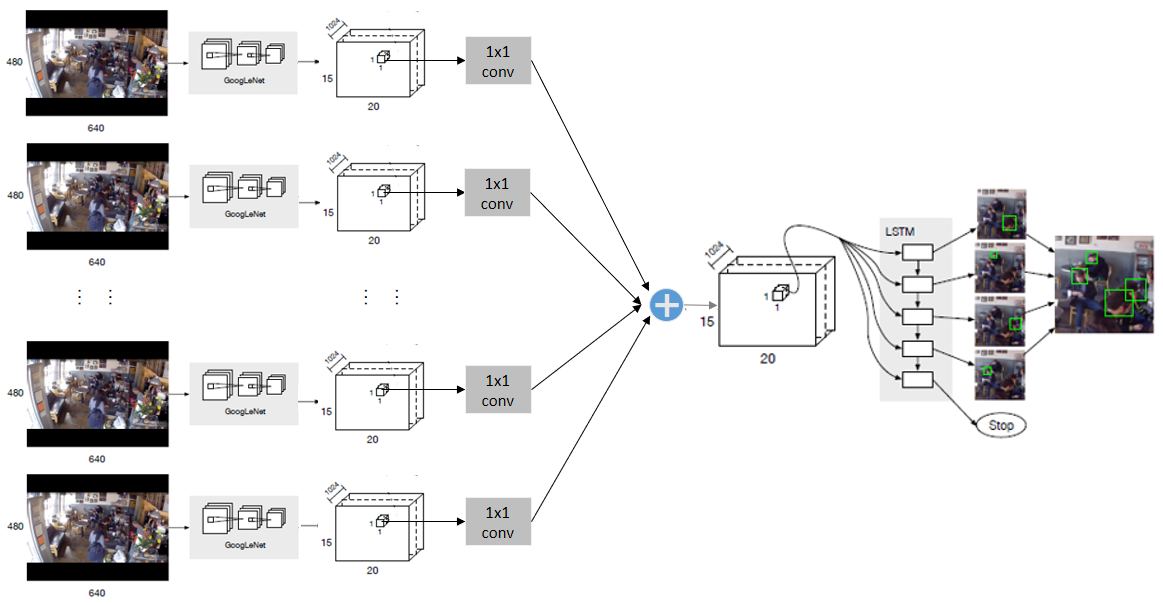


图3-16 1x1卷积连接连续帧信息网络结构（参数不共享）

### 通过LSTM来连接连续帧信息

1. **设计思路和网络结构**

与本章4.2相同，首先输入连续的视频帧九帧，分别用GoogLeNet网络对图像编码，生成包含丰富信息的高层特征描述子，卷积神经网络GoogLeNet仍然将整个图片划分为 300（20x15）个感受野相互重叠的cell，将每个cell编码为一个1024维的高层特征描述子。

然后使用LSTM（具体介绍见本章1.2.2）来连接连续帧信息，通过前几帧的feature map预测当前帧的feature map，其中LSTM的size为1024，即针对每一个格子区域进行LSTM预测操作，可以降低整个网络的参数量，同时也取得了较好的效果。每次循环，我们都将 GoogLeNet 提取的特征与LSTM 上一次循环的输出串联起来，作为 LSTM 当前循环的输入，以此加强特征描述。

最后，还使用1x1卷积对LSTM预测出feature map和当前帧的feature map做线性组合，得到的feature map仍然是将20x15x1024，即将整个图片划分为 300（20x15）个感受野相互重叠的cell，将每个cell编码为一个1024维的高层特征描述子，每个cell都与一个与之对应的长短期记忆网络LSTM相连，每个LSTM的长度为5，即输出5个检测目标和对应的置信度。

具体网络结构仍然分为参数共享和参数不共享两种特征提取方式，首先介绍参数共享的网络结构，参数共享的具体操作和使用1x1卷积核的情况类似，整体网络结构如下图3-17所示。

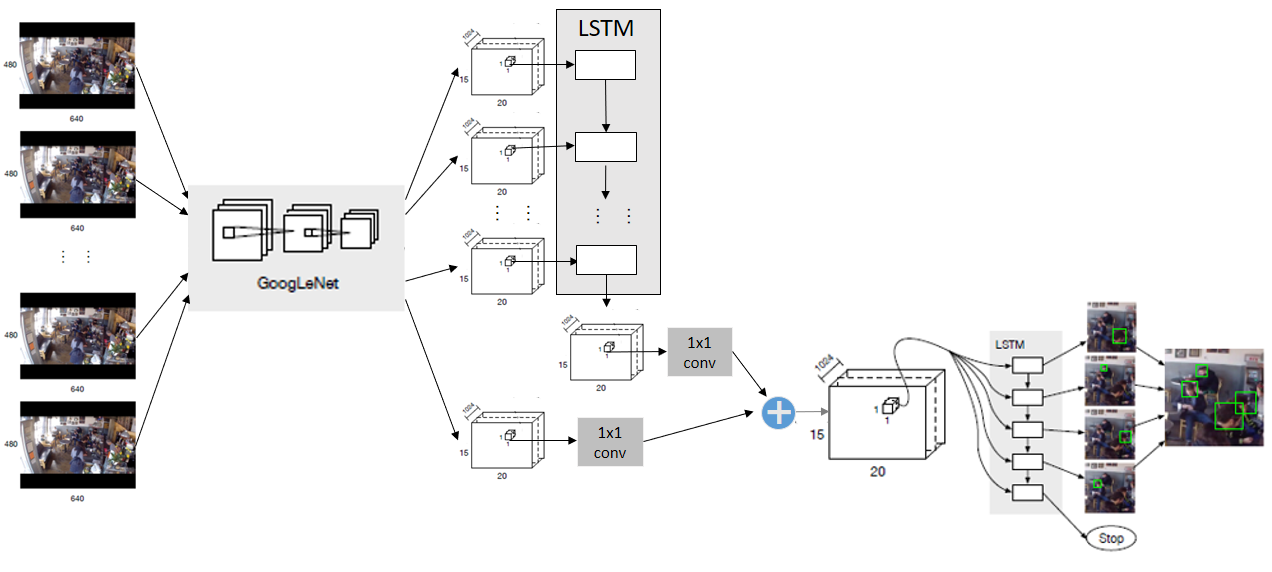


图3-17 LSTM连接连续帧信息网络结构（参数共享）

接下来是参数不共享的情况，具体操作不再详细描述，与本章4.2.2类似，网络结构如图3-18所示。

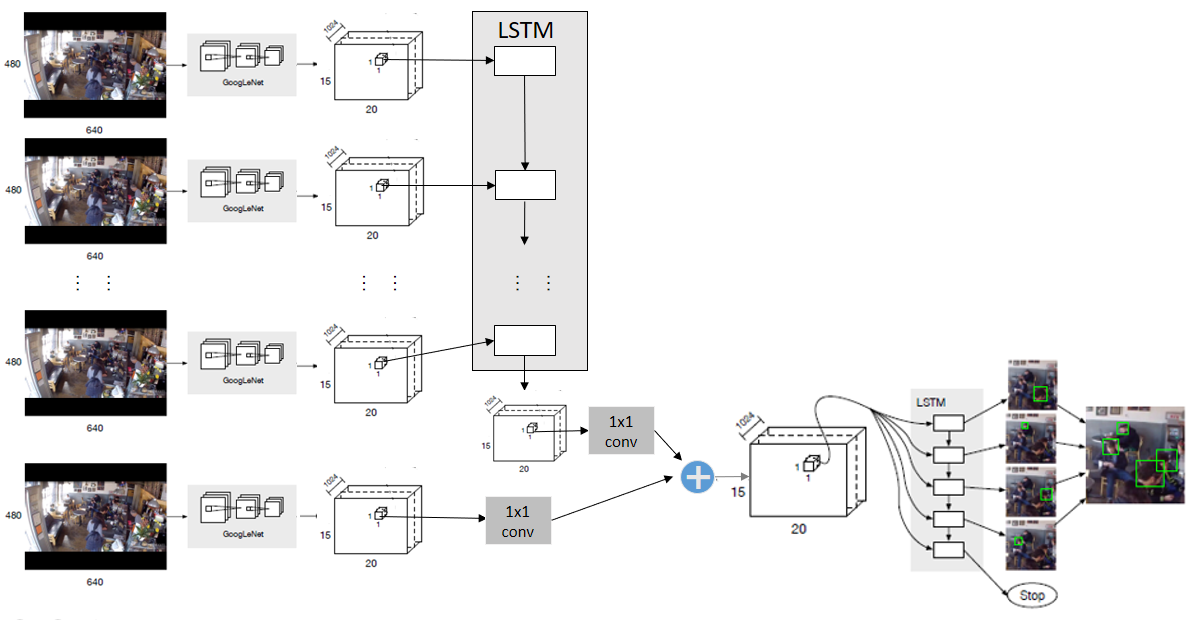


图3-18 LSTM连接连续帧信息网络结构（参数不共享）

## 本章小结

本章首先介绍了论文实验的理论基础，主要是GoogLeNet、循环神经网络以及其他相关理论基础，然后介绍了数据的采集和标注，并着重探究了几种数据增强的方式，接下来在实验理论的基础上，详细阐述了基于当前帧的目标检测和人群计数网络的设计思路以及优化目标。最后，在基于当前帧（单帧）的目标检测网络的基础上，提出了基于连续帧信息的目标检测和人群计数的网络结构，输入连续几帧图片来进行当前帧的目标检测，本文采用了两种连接连续帧的方式，分别是使用1x1卷积核和使用LSTM。

# 第四章 实验结果与分析

## 数据集和评测标准

### 数据集

本文采用的数据集为自己手动标注的视频帧，具体介绍见第三章“前期数据准备”，其中在训练时采用了数据增强的策略来扩增数据集。原始训练集为2000张视频帧，每帧人数大概为20人左右，多集中在某一较小区域，较为拥挤。测试集为350张视频帧，共标注人头6117个，测试用的视频帧和训练用的视频帧采集于同一场景的不同时间段。

### 评测标准

本文采用的目标检测评测指标为EER（Equal Error Rate），如图4-1所示，即PR曲线和对角线的交点，此时Recall和Precision相同，错误接受率和错误拒绝率相同，即准确率和召回率的折衷。

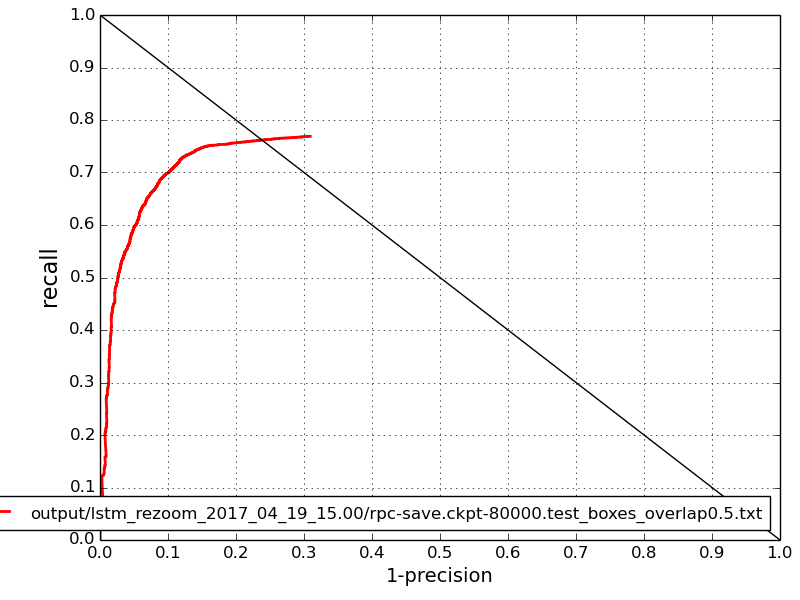


图4-1 实验效果评测示例

人群计数采用的评测标准为MAE（Mean Absolute Error），即平均绝对误差。平均绝对误差的定义如下：

## 实验设置

本文使用TensorFlow开源深度学习框架来实现目标检测和人群计数方法。在实验中，我们规定输入人群计数模型的图片大小为480x640，从而可以推出每个cell对应原图一个大小为139x139的感受野。训练行人检测模块时，我们在每个cell感受野中央的64x64的区域进行检测，因为通过实验发现，64x64的区域已经足够捕捉局部的遮挡关系。

在本文实验中，我们将初始的学习率设置为0.01，每迭代30,000次，学习率乘以0.1，最大迭代次数为500,000次。GoogLeNet网络使用ImageNet数据集上训练的模型进行初始化，然后在该基础上进行finetune使其适合本文的数据集。用于连接连续帧信息的LSTM的记忆细胞状态是一个1024维的向量，最后用于目标检测的LSTM的记忆细胞状态是一个500维的向量。所有LSTM的权重初始化为[0.1,0.1]的均匀分布。

## 实验结果

实验结果如表4-1所示，共分为三列，其中第一列为采用的目标检测和人群计数的方法，第二列为对应方法的EER值，第三列为对应方法的MAE值。由表格可知，共比较了5种方法，UNI-GoogLeNet表示基于当前帧（单帧）的目标检测和人群计数算法，CONV-GoogLeNet表示使用1x1卷积核去连接连续帧信息的目标检测和人群计数算法，分为参数共享（Reuse）和参数不共享（Not Reuse）两种情况，LSTM-GoogLeNet表示使用LSTM连接连续帧信息的目标检测和人群计数算法，同样分为参数共享（Reuse）和参数不共享（Not Reuse）两种情况。以上方法全部使用GoogLeNet来提取特征，具体实验结果如下表：

表4-1 实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| Method |  | EER | MAE |
| UNI-GoogLeNet  CONV-GoogLeNet, Reuse  CONV-GoogLeNet, Not Reuse  LSTM-GoogLeNet, Reuse  LSTM-GoogLeNet, Not Reuse |  | 74.73% 75.47%  76.20%  76.53%  75.77% | 1.73  1.70  1.65  1.66  1.68 |

除此之外，在相同数据集上使用Faster R-CNN训练了对应的目标检测和人群计数模型，由于目标检测的衡量指标不一样，就没有进行比较。其中人群计数的的MAE指标为1.89。上面算法明显优于Faster R-CNN。

最后，使用ResNet网络替换GoogLeNet来提取特征，并比较了两种特征提取方式的实验结果，其中UNI-GoogLeNet的实验效果如表4-1所示，UNI-ResNet的EER为82.33%，MAE为1.59。

## 实验结果分析

根据实验结果，具体分析如下：

（1）本次实验所用数据集的图片分辨率较低，且场景较为拥挤。标注的时候需要看前后帧的信息才能准确完成标注，而且实验结果表明连续帧信息对当前帧的检测有一定的影响。通过尝试多种连接连续帧信息的方式，发现通过使用1x1卷积核或者LSTM连接连续帧信息能够取得较好的结果；

（2）通过LSTM连接连续帧信息要稍好于通过1x1卷积实现连续帧信息线性组合的方式。使用1x1卷积核实现连续帧信息的线性连接较为简单，而LSTM经常被用来学习时序信息，用LSTM来连接连续帧信息应该能够取得更好的实验效果；

（3）通过设置参数共享和参数不共享两种状态，可以看出使用1x1卷积核对连续帧信息实现线性组合比较适合参数不共享的状态，即使用不同的GoogLeNet提取特征，而使用LSTM连接连续帧信息明显适合参数共享的情况，也同时从侧面说明LSTM在某种程度上能够连接连续帧信息，因为用同一个GoogLeNet网络提取特征才能找出规律预测下一帧，如果特征提取的方式不同，那么就很难找出相邻帧之间的规律，无法预测下一帧的特征表示；

（4）本实验的主要目的在于探索连续帧信息对检测结果的影响，除此之外，通过和Faster R-CNN的比较可以看出使用LSTM来替代NMS能够在拥挤场景下取得较好的实验效果。LSTM在一定程度上能够处理人头区域重叠较严重的情况，而NMS在人头区域重叠严重的情况下，会排除掉一些正样本；

（5）最后通过使用ResNet替换GoogLeNet表明，ResNet能够更好的提取图片的特征，能够取得较好的实验效果。同理，如果用更好的网络继续替换ResNet，在很大程度上能够取得更好的结果。

## Demo展示

开发环境：Ubuntu 16.04.1，16G内存，2.8GHz；GPU为GTX 1060，6G显存；程序语言为Python 2.7。

接下来简要介绍一下目标检测和人群计数系统的相关实现细节，首先后端目标检测和人群计数采用的是UNI-ResNet算法模型，每帧图片的处理速度大概为200ms，然后将获取到的视频帧经过检测统计之后展示在视频播放框内，这里采集的视频的帧率为2fps，也可以直接获取电脑摄像头的场景视频，除此之外，设置了几个功能键，start开始播放和检测键，stop暂停播放和检测键，reset重头开始播放视频键，clear开始不进行检测。

目标检测和人群计数系统如图4-2所示：

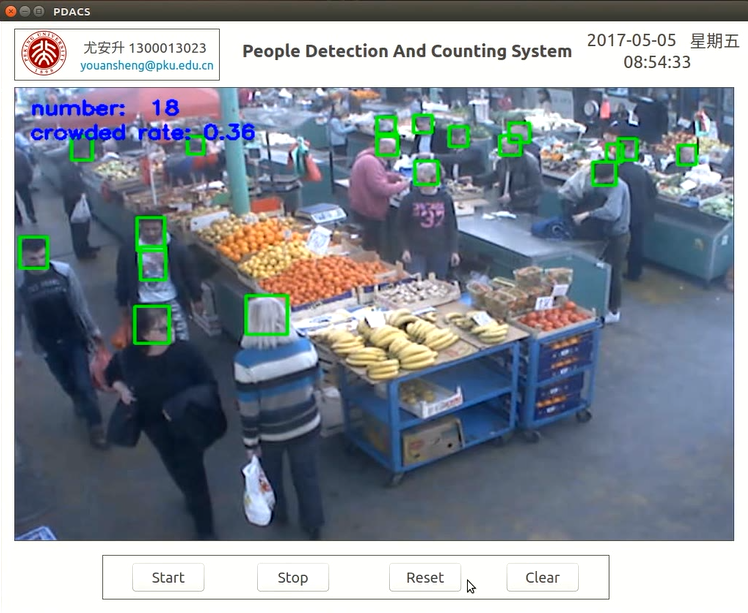


图4-2 目标检测和人群计数系统

## 本章小结

本章首先介绍了实验所用的数据集以及相关的评测标准（EER和MAE），然后着重介绍了实验设置并给出了各种算法下的实验效果，接下来对实验效果进行了一定程度的分析，最后还以图片加叙述的形式进行了目标检测和人群计数系统的Demo展示。

# 第五章 总结和展望

## 本文工作总结

目标检测和人群计数是图像处理和计算机视觉学科的重要分支，也是智能监控系统的核心部分，具有重要的社会意义。目标检测广泛应用于机器人导航、智能视频监控、工业检测、航空航天等诸多领域，同时也是泛身份识别领域研究的关键性基础；人群计数作为智能视频监控系统的重要课题之一，对高密度人群的智能检测和计数，能够减少对人力资本的消耗，预防人群安全事故，有助于保障重点区域（比如：广场、体育场、人行道、机场等）人群的人身安全。

本文在广泛调研国内外目标检测和人群计数方法的基础上，通过考虑视频帧的连续帧信息，即时序信息，设计并实现了面向拥挤场景视频的基于连续帧信息的目标检测和人群计数方法。首先，独自标注了拥挤场景视频帧2350帧（2fps，采集于webcam），其中2000张每秒标注两张图片作为训练集，350张每隔5s标注一张图片作为测试集，与此同时，由于数据集较小，探索了模拟光照、模糊处理、噪声处理、颜色抖动等一系列数据增强方法；然后设计并实现了基于连续帧信息的人物检测和计数方法的一系列模型，提出了两种有效连接时序信息的方式，一种是沿时序卷积，通过1x1的卷积核将连续帧的feature map进行线性组合；另一种是使用LSTM对当前帧进行预测，最后将预测得到的结果与当前帧的feature map进行线性组合。

本文在自己标注的数据集上进行实验，验证了本文方法的有效性和对真实数据的处理能力。实验证明，连续帧信息对当前帧的目标检测和人群计数效果有一定的提升，LSTM替代NMS能够在一定程度上解决重叠较为严重的情况，以及使用较好的特征提取网络能够大幅度提高检测和计数的效果。

## 改进意见和展望

尽管深度学习在近几年取得的成功标志着我们向人工智能迈进了一大步，但

是目前的科技水平仍未达到真正的智能化。在建设健全的智能视频监控系统的过

程中，仍然存在许多困难和挑战。本文着眼于智能视频监控系统的基本任务“目标检测和人群计数”，设计并实现了基于连续帧信息的人物检测和计数方法的一系列模型，提出了两种有效连接时序信息的方式，并在标注数据集上验证了方法的有效性。在下一步的工作中，我们可以从以下几点进行改进：

（1）进一步探索更好的连接时序信息的方式，争取能够找出较好的方法来考虑视频的时序信息，进一步提升目标检测和计数的结果；

（2）进一步扩大数据集，尝试用不同场景下人群更加密集的数据集，提高模型的泛化能力，最终将模型运用到实际的场景；

（3）在不损失目标检测和计数精度的情况下，尽量精简网络的结构，同时探索其他能够进行模型加速的办法，使之在速度上能够满足实时的要求。

我们相信，在未来，人工智能技术将推进视频监控系统将达到更加智能化的水平，从而改善人们的生活，为人类社会带来福利。

# 参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] Vishwakarma S, Agrawal A. A survey on activity recognition and behavior understanding in video surveillance [J]. The Visual Computer, 2012: 1-27.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[5] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. IJCV, 2013.

[6] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR 2014.

[7] Girshick R. Fast R-CNN [J]. Computer Science, 2015.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Neural Information Processing Systems (NIPS), 2015.

[9] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2016:779-788.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2016.

[11] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1106–1114, 2012.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” CoRR, vol. abs/1409.4842, 2014.

[14] Hochreiter, Sepp, and Jurgen Schmidhuber. Long Short-Term Memory. Neural Computation 9.8 (1997): 1735. Psychology and Behavioral Sciences Collection. Web. 2 May 2016.

[15] Gers, F. A., J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with LSTM. Neural Computation 2.10(1999):2451-2471.

[16] Stewart, Russell, and M. Andriluka. End-to-end people detection in crowded scenes. Computer Science (2015).

[17] Lin M, Chen Q, Yan S. Network In Network [J]. Computer Science, 2014.

[18] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." Computer Vision and Pattern Recognition IEEE, 2015:770-778.

# 致谢

时光荏苒，四年的大学生活悄悄接近了尾声，感谢北京大学四年以来对我的辛苦培育，让我在园子里这四年来遇到很多优秀的老师和同学，学到很多东西，非常感谢信息科学技术学院为我提供了良好的学习环境、感谢领导、老师们四年来对我无微不至的关怀和指导，让我能够健康自信的成长。在此，我还要感谢在班里同学和朋友，感谢你们在我遇到困难的时候帮助我，给我支持和鼓励，感谢你们。

在完成论文的过程中，感谢实验室丁雨辰师兄一直以来的指导和帮助，提出了很多宝贵的建议。同时，感谢清华代旭师兄、北大王洋师兄在技术上的指导，在和你们讨论的过程中，加深了我对论文研究领域的理解。特别感谢我的指导老师童云海副教授，感谢您对我的论文的指导和帮助。您严谨的治学态度、高涨的工作热情，还有您的包容、友善和耐心，我都将受益终身，非常期待研究生阶段能继续聆听您的教诲。感谢百忙之中抽出时间的各位评审老师，由于本人知识有限，敬请各位老师多多指教！

最后，我要感谢我的父母，你们对我二十多年的培养和爱护，还有一直以来

的理解与支持，使我能够在成长的道路上安心、踏实地不断前行。你们是我最大

的幸运和幸福。