



# 计算机科学与技术学院

## 毕业设计

论文题目	<u>基于 SSD 网络模型的房屋瓦片损害检测</u>		
学校导师	<u>刘立</u>	职称	<u>教授</u>
企业导师	<u>刘立</u>	职称	<u>教授</u>
学生姓名	<u>李开运</u>	学号	<u>20144330106</u>
专业班级	<u>物联网</u>	班级	<u>14 级 01 班</u>
系主任	<u>毛宇</u>	院长	<u>刘振宇</u>
起止时间	<u>2017 年 6 月 5 日至 2018 年 5 月 22 日</u>		

2018 年 3 月 8 日

# 目录

<b>第一章 绪论</b>	<b>6</b>
1.1 课题背景及研究意义	6
1.1.1 研究背景	6
1.1.2 研究意义	7
1.2 研究现状及发展难点	7
1.2.1 研究现状	7
1.2.2 发展难点	8
1.3 研究内容及章节安排	8
<b>第二章 目标检测相关算法</b>	<b>10</b>
2.1 目标检测算法概述	10
2.2 Viola-Jones 人脸检测器	10
2.3 可变形部件模型 (DPM)	11
2.4 R-CNN 系列	11
2.4.1 R-CNN	12
2.4.2 SPP Net	13
2.4.3 Fast R-CNN	14
2.4.4 Faster R-CNN	16
2.4.5 Mask R-CNN	19
2.5 YOLO 系列	19
2.5.1 YOLO	20
2.5.2 SSD	20
2.5.3 YOLO9000	20
2.6 本章小结	20
<b>第三章 瓦片损害检测算法设计</b>	<b>21</b>
3.1 瓦片损害检测流程	21
3.2 SSD 算法核心思想	21
3.3 SSD 模型结构	23
3.4 损失函数	24
3.4.1 SSD 中采用的损失函数	24
3.4.2 改进损失函数-Focus Loss	24
3.5 SSD 模型训练	26
3.6 算法改进方案	27
3.6.1 Soft-NMS	27
3.6.2 数据增强	28
3.7 本章小结	28
<b>第四章 瓦片损害检测算法实现</b>	<b>29</b>
4.1 图像预处理	29
4.1.1 标注工具	29
4.1.2 数据集	29
4.1.3 Pytorch 介绍	29
4.2 网络模型实现	30
4.3 损失函数	33
4.4 模型训练/测试	34
4.5 瓦片检测	34

<b>第五章 实验结果与分析</b>	35
5.1 改进 ssd 算法的实验结果 . . . . .	35
5.2 改进 ssd 对瓦片损害检测的准确率和召回率的实验 . . . . .	35
5.3 ssd 与修改了 focus loss 的算法进行 mAP、速度的比较 . . . . .	35
<b>第六章 总结及展望</b>	36
<b>第七章 致谢</b>	37

## 基于 SSD 网络模型的房屋瓦片损害检测

**摘 要：** 近年来，自然灾害的频发导致美国房屋理赔行业急需一套无损害、低成本的对房屋瓦片损害的检测方案。随着商用无人机的发展和基于深度学习的各类检测算法的出现，使得通过无人机代替工作人员进行房屋瓦片检测成为现实。本研究通过比较近年来主要的目标检测算法，考虑到瓦片的特征，采用基于 SSD: Single Shot MultiBox Detector [5] 改进算法进行瓦片损害检测，实验结果满足了保险行业的要求。

**关键词：** 目标检测，深度学习，SSD

## House tile damage detection based on SSD network model

## 第一章 绪论

### 1.1 课题背景及研究意义

#### 1.1.1 研究背景

**自然灾害频发：**美国中文网根据今日美国报道，美国国家海洋和大气管理局 (NOAA) 周一宣布，由于三次强大的飓风和凶猛的野火，2017 年是美国遭受自然灾害最为严重的一年。这些自然灾害让美国遭受了 3060 亿美元的损失。

2017 年，美国历经 16 次天气和气候灾害，每一项灾害的损失都超过了 10 亿美元。总损失约为 3060 亿美元，创下了新纪录。它打破了 2005 年的纪录，当年飓风卡特里娜等灾害给美国造成了 2150 亿美元的损失。NOAA 说，去年的灾难共造成全美 (包括波多黎各) 362 人死亡。然而，NOAA 气候学家亚当·史密斯 (Adam Smith) 表示，死亡人数可能会随着波多黎各的后续报告而增加。史密斯说，这也是有史以来破坏最强的飓风季，损失达到 2650 亿美元。也是有史以来损失最大的野火季，损失达到了 180 亿美元。飓风哈维总共造成了 1250 亿美元的损失，在近 30 年来造成的破坏仅次于飓风卡特里娜。飓风玛丽亚和艾玛分别造成了 900 亿美元和 500 亿美元的损失。这个消息是在德州奥斯汀的美国气象学会年会上公布的。美国大陆和阿拉斯加在 2017 年的气温也是连续第三年高于平均水平。亚利桑那州、乔治亚州、新墨西哥州、北卡罗莱纳州和南卡罗莱纳州 5 个州在 2017 年都经历了有记录以来最温暖的一年。包括阿拉斯加在内的 32 个州 2017 年的气温也创下有记录以来的前十高温。<sup>1</sup>

**无人机的商用：**从手掌大小的微型飞行器到可用于检查输电线路的商用无人机，目前市面上在售的无人机种类和数量都在迅速增加。较小的最低 40 美元就可以买到，但高端无人机的价格至少也要数千美元 (军用无人机的成本更加高昂)。消费类无人机的用途主要是娱乐和拍摄，大的可以执行任务的无人机则开始用于商业投递。

说起无人机技术，大家可以想到层面会是多种多样，有军用无人机技术领域的“全球鹰”、“捕食者”，有民用无人机领域的测绘、航拍甚至快递。那个距离我们最近的“智能科技”，未来发展之路是如何呢？可以说，无人机技术是智能科技皇冠上的一颗璀璨宝石，因为它有最小的体积，集成了最高的人类科技——1981 年，第一台商用 GPS 接收机诞生，重达 50 磅，价格高达 10 万美元。现在 GPS 仅重 0.3 克，芯片成本也大幅下降，无人机将 GPS 技术集纳；1976 年，柯达推出了第一款数码相机，像素只有 10 万，重量为 3.75 磅，价格超过 1 万美元。而无人机将最新的数字相机技术整合，并根据用途，细分了数种功能性镜头；除此之外，计算机技术、蓝牙通讯技术更是无人机的必备功能，有了这些智能抗美科技的“加持”，说无人机是一枚宝石并不为过。

**机器视觉的发展：**随着中国制造业的蓬勃发展，机器视觉行业也在中国市场度过了发展的最初时期，不仅国际知名品牌纷纷在中国开展业务，中国本土的企业也

<sup>1</sup>美国中文网,2018-1-8

逐渐兴起，机器视觉已为广大客户所熟知，应用范围也逐步扩大，由起初的电子制造业和半导体生产企业，发展到了包装，汽车，交通和印刷等多个行业。

近几年深度学习发展迅猛，更是由于前段时间的谷歌的 AlphaGo 而轰动一时，国内也开始迎来这一技术的研究热潮，深度学习目前还处于发展阶段，不管是理论方面还是实践方面都还有许多问题待解决，不过由于我们处在了一个“大数据”时代，以及计算资源的大大提升，新模型、新理论的验证周期会大大缩短。人工智能时代的开启必然会很大程度的改变这个世界，无论是从交通，医疗，购物，军事等方面，或许我们正处于最好的年代。

### 1.1.2 研究意义

针对传统的房屋瓦片检测方法存在着操作复杂、耗时、高危和具有破坏性等缺点，本项研究尝试利用计算机视觉技术对瓦片进行快速无损检测。本文提出了一种利用基于 SSD(Single Shot MultiBox Detector) [5] 改进算法进行瓦片损害检测的方案，为房屋检测行业提供利用机器视觉处理传统检测问题的高效手段。其意义有三：

1、**无损检测**：传统的房屋瓦片损害检测工作，需要工作人员爬上房屋拍照，将照片带回工作室进行损害鉴定，这种操作无疑会对瓦片带来人为的破坏；与此同时，工作人员因操作不当受伤甚至致死的报道也时有发生，本项研究利用无人机替代工作人员的拍照工作，利用无人机的图像处理模块，实时进行损害的检测，将结果和图片一并送到系统中进行信息的汇总和检测报告的生成。做到的对瓦片和对工作人员的两个“无损”。

2、**缩短检测周期**：在美国，遇到自然灾害致使房屋受损后，参保的家庭会联系保险公司进行理赔，按照现在的处理水平，一栋房屋平均会耗时 7 个星期，对于偏远的地方会更久。采用无人机对受损房屋瓦片进行检测会将这个时长缩短到 1 个星期。大大节约了成本。同时实验也表明有更好的检测效果。

3、**节约人力成本**：经融危机和通货膨胀造成了人力成本的极大提高。传统的损害检测十分依赖工作人员的经验，所以人力成本一直居高不下，采用搭载了检测算法的无人机进行检测，摆脱了对工作人员经验的依赖，将成本从 100 美元降到了 10 美元。

## 1.2 研究现状及发展难点

本文主要是利用 SSD 改进算法对房屋瓦片损害进行检测，涉及到目标检测系列算法，对于此系列算法的研究是深度学习方向的研究热点。定位和分类可以迭代起来，最终在一张图片汇总对多个对象进行检测和分类。目标检测是在图像上发现和分类一个变量的问题。目标检测与定位、分类相比，重要的区别是这个“变量”。对象检测的输出长度是可变的，因为检测到的对象的数量会根据图像的不同而变化。

### 1.2.1 研究现状

在深度学习正式介入之前，传统的“目标检测”方法都是区域选择、提取特征、分类回归三部曲，这样就有两个难以解决的问题；其一是区域选择的策略效果差、时间复杂度高；其二是手工提取的特征鲁棒性较差。云计算时代来临后，“目标检测”

算法大家族主要划分为两大派系。

基于“Proposal + Classification”的 Object Detection 的方法，RCNN 系列 (RCNN [3]、SPPnet [4]、Fast R-CNN [2] 以及 Faster R-CNN [7]) 取得了非常好的效果，因为这一类方法先预先回归一次边框，然后再进行骨干网络训练，所以精度要高，这类方法被称为“two stage”的方法。但也正是由于此，这类方法在速度方面还有待改进。由此，YOLO [6] 应运而生，YOLO 系列只做了一次边框回归和打分，所以相比于 RCNN 系列被称为“one stage”的方法，这类方法的最大特点就是速度快。但是 YOLO 虽然能达到实时的效果，但是由于只做了一次边框回归并打分，这类方法导致了小目标训练非常不充分，对于小目标的检测效果非常的差。简而言之，YOLO 系列对于目标的尺度比较敏感，而且对于尺度变化较大的物体泛化能力比较差。

针对 YOLO 和 Faster R-CNN 的各自不足与优势，WeiLiu 等人提出了 Single Shot MultiBox Detector，简称为 SSD [5]。SSD 整个网络采取了“one stage”的思想，以此提高检测速度。并且网络中融入了 Faster R-CNN [7] 中的 anchors 思想，并且做了特征分层提取并依次计算边框回归和分类操作，由此可以适应多种尺度目标的训练和检测任务。SSD 的出现使得大家看到了实时高精度目标检测的可行性。

### 1.2.2 发展难点

**可变数量的对象** (Variable number of objects)。在训练机器学习模型时，通常需要将数据表示为固定大小的向量。但是，由于图片中对象的数量事先不知道，所以我们不知道正确的输出维度。因此需要一些后期处理，这增加了模型的复杂性。一般使用滑动窗口的方法来处理可变数量的对象，通过滑动固定大小的窗口，在所有的地方生成固定大小的特征。在得到这些被过滤后的特征之后，一些被丢弃，另一些被合并以生成最终的结果。

**调整对象检测窗口大小** (Resizing)。另一个巨大的挑战是各种可能的对象大小，即在进行分类时，既希望占图片大部分的对象进行分类，又想要找到一些可能只有 12 个像素、或者是原始图像一小部分的小对象。使用不同尺寸的滑动窗口可以解决这个问题，但效率很低。

**建模**。第三个挑战是同时解决两个问题——如何用一个简单的模型解决两种不同的需求，即定位和分类。

## 1.3 研究内容及章节安排

第一章：绪论。本章概要阐述本文主要内容，及研究背景及意义。

第二章：目标检测相关算法。本章按照两条主线系统讲解国内外对于目标检测算法的研究。

第三章：瓦片损害检测算法设计。通过第二章的综述，使我们了解到现有的目标检测相关算法。在第三章吸收了各种算法的优劣，我对 SSD 算法进行了两点改进：

1、Loss 函数 2、Soft-NMS

第四章：瓦片损害检测算法实现。本章展示具体的算法实现，基于 pytorch。

第五章：实验结果及分析。本章结合原始算法与改进算法进行准确率和召回率



的对比，并展示该算法对于瓦片损害检测的效果

第六章：总结及展望。

第七章：致谢。

## 第二章 目标检测相关算法

### 2.1 目标检测算法概述

目标检测一直是计算机视觉的基础问题，在 2010 年左右就开始停滞不前了。自 2013 年一篇论文的发表，目标检测从原始的传统手工提取特征方法变成了基于卷积神经网络的特征提取，从此一发不可收拾。根着历史的潮流，简要地探讨“目标检测”算法的两种思想和这些思想引申出的算法，主要涉及那些主流算法。

在深度学习正式介入之前，传统的“目标检测”方法都是区域选择、提取特征、分类回归三部曲，这样就有两个难以解决的问题；其一是区域选择的策略效果差、时间复杂度高；其二是手工提取的特征鲁棒性较差。云计算时代来临后，“目标检测”算法大家族主要划分为两大派系，一个是 R-CNN 系“two stage”，另一个则是以 YOLO 为代表的“one stage”。下面分别解释一下“two stage”和“one stage”。

Two stage: 顾名思义，分两步解决问题：

- 1、生成可能区域 (Region Proposal)& CNN 提取特征
- 2、放入分类器分类并修正位置

这一流派的算法都离不开 Region Proposal，即是优点也是缺点，主要代表人物就是 R-CNN 系。

One stage: 顾名思义，一步解决问题，直接对预测的目标物体进行回归。回归解决问题简单快速，但是太粗暴了，主要代表人物是 YOLO 和 SSD。

无论“two stage”还是“one stage”，他们都是在同一个天平下选取一个平衡点、或者选取一个极端——要么准，要么快。”two stage”的天平主要倾向准，“one stage”的天平主要倾向快。但最后万剑归宗，大家也找到了自己的平衡，平衡点的有略微的不同。接下来我们花开两朵各表一支，一朵“two stage”的前世今生，另一朵“one stage”的发展历史。

### 2.2 Viola-Jones 人脸检测器

在 2001 年, Viola 和 Jones 发表了经典的 Rapid Object Detection using a Boosted Cascade of Simple Features [8] 和 Robust Real-Time Face Detection [9], 在 AdaBoost 算法的基础上, 使用 Haar-like 小波特征 (简称类 haar 特征) 和积分图方法进行人脸检测, 他俩不是最早使用提出小波特征的, 但是他们设计了针对人脸检测更有效的特征, 并对 AdaBoost 训练出的强分类器进行级联。这可以说是人脸检测史上里程碑式的一笔了, 也因此当时提出的这个算法被称为 Viola-Jones 检测器。

物体检测在整个计算机领域里, Viola-Jones 人脸检测器是早期比较成功的一个例子, 其使得物体检测相比而言成了一项较为成熟的技术。这个方法基本的思路就是滑动窗口式的, 用一个固定大小的窗口在输入图像进行滑动, 窗口框定的区域会被送入到分类器, 去判断是人脸窗口还是非人脸窗口。滑动的窗口其大小是固定的, 但是人脸的大小则多种多样, 为了检测不同大小的人脸, 还需要把输入图像缩放到不同大小, 使得不同大小的人脸能够在某个尺度上和窗口大小相匹配。这种滑动窗

口式的做法有一个很明显的问题，就是有太多的位置要去检查，去判断是人脸还是非人脸。

判断是不是人脸，这是两个分类问题，在 2000 年的时候，采用的是 AdaBoost 分类器。进行分类时，分类器的输入用的是 Haar 特征，这是一种非常简单的特征，在图上可以看到有很多黑色和白色的小块，Haar 特征就是把黑色区域所有像素值之和减去白色区域所有像素值之和，以这个差值作为一个特征，黑色块和白色块有不同的大小和相对位置关系，这就形成了很多个不同的 Haar 特征。AdaBoost 分类器是一种由多个弱分类器组合而成的强分类器，Viola-Jones 检测器是由多个 AdaBoost 分类器级联组成，这种级联结构的一个重要作用就是加速。

2000 年人脸检测技术开始成熟起来之后，就出现了相关的实际应用，例如数码相机中的人脸对焦的功能，照相的时候，相机会自动检测人脸，然后根据人脸的位置把焦距调整得更好。

### 2.3 可变形部件模型 (DPM)

Viola-Jones 人脸检测器之后，在 2009 年出现了另外一个比较重要的方法：deformable part model(DPM) [1]，即可变形部件模型。就人脸检测而言，人脸可以大致看成是一种刚体，通常情况下不会有非常大的形变，比方说嘴巴变到鼻子的位置上去。但是对于其它物体，例如人体，人可以把胳膊抬起来，可以把腿翘上去，这会使得人体有非常多非常大的非刚性变换，而 DPM 通过对部件进行建模就能够更好地处理这种变换。刚开始的时候大家也试图去尝试用类似于 Haar 特征 + AdaBoost 分类器这样的做法来检测行人，但是发现效果不是很好，到 2009 年之后，有了 DPM 去建模不同的部件，比如说人有头有胳膊有膝盖，然后同时基于局部的部件和整体去做分类，这样效果就好了很多。DPM 相对比较复杂，检测速度比较慢，但是其在人脸检测还有行人和车的检测等任务上还是取得了一定的效果。后来出现了一些加速 DPM 的方法，试图提高其检测速度。DPM 引入了对部件的建模，本身是一个很好的方法，但是其被深度学习的光芒给盖过去了，深度学习在检测精度上带来了非常大的提升，所以研究 DPM 的一些人也快速转到深度学习上去了。

### 2.4 R-CNN 系列

R-CNN 其实是一个很大的家族，自从 Ross Girshick 发表 Rich feature hierarchies for accurate object detection and semantic segmentation [3]，子孙无数、桃李满天下。在此，我们只探讨 R-CNN 直系亲属，他们的发展顺序如下：



他们在整个家族进化的过程中，一致暗埋了一条主线：充分榨干 feature maps 的价值。

### 2.4.1 R-CNN

Region CNN(RCNN) [3] 可以说是利用深度学习进行目标检测的开山之作。作者 Ross Girshick 多次在 PASCAL VOC 的目标检测竞赛中折桂，2010 年更带领团队获得终身成就奖，如今供职于 Facebook 旗下的 FAIR。图 3.1 这个模型，是利用卷积神经网络来做“目标检测”的开山之作，其意义深远不言而喻。

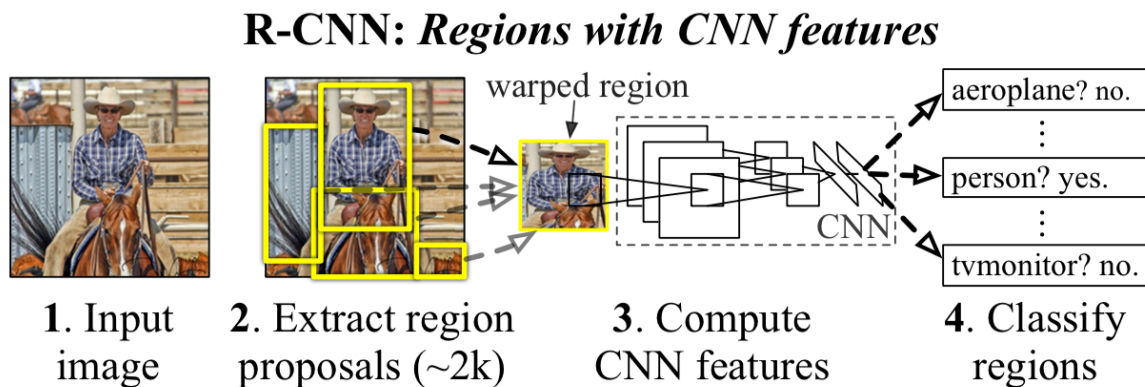


图 2.1: RCNN 算法框架

本文解决了目标检测中的两个关键问题 **解决问题一、速度**。传统的区域选择使用滑窗，每滑一个窗口检测一次，相邻窗口信息重叠高，检测速度慢。R-CNN 使用一个启发式方法 (Selective search)，先生成候选区域再检测，降低信息冗余程度，从而提高检测速度。

**解决问题二、训练集**。经典的目标检测算法在区域中提取人工设定的特征 (Haar, HOG)。传统的手工提取特征鲁棒性差，限于如颜色、纹理等低层次 (Low level) 的特征。使用 CNN(卷积神经网络) 提取特征，可以提取更高层面的抽象特征，从而提高特征的鲁棒性。

该方法将 PASCAL VOC 上的检测率从 35.1% 提升到 53.7%，提高了好几个量级。虽然比传统方法好很多，但是从现在的眼光看，只能是初窥门径。

#### 算法流程:

- > 一张图像生成 1K 至 2K 个候选区域
- > 对每个候选区域，使用深度网络提取特征
- > 特征送入每一类的 SVM 分类器，判别是否属于该类
- > 使用回归器精细修正候选位置

1、**候选区域生成**: 使用 Selective Search 方法从一张图像生成约 2000-3000 个候选区域。基本思路如下:

- 使用一种分割手段，将图像分割成小区域

- 查看现有小区域，合并可能性最高的两个区域。重复直到整张图像合并成一个区域位置

- 输出所有曾经存在过的区域，所谓候选区域

2、**特征提取**: 借鉴 Hinton 2012 年在 Image Net 上的分类网络，提取特征。

3、**类别判断**: 对每一类目标，使用一个线性 SVM 二类分类器进行判别。输入为深度网络输出的 4096 维特征，输出是否属于此类。由于负样本很多，使用 hard negative mining 方法。

4、**位置精修**: 目标检测问题的衡量标准是重叠面积：许多看似准确的检测结果，往往因为候选框不够准确，重叠面积很小。故需要一个位置精修步骤。

## 2.4.2 SPP Net

R-CNN 提出后的一年，以何恺明、任少卿为首的团队发表了 Spatial pyramid pooling in deep convolutional networks for visual recognition(SPP Net) [4]，这才是真正摸到了卷积神经网络的脉络。也不奇怪，毕竟这些人鼓捣出了 ResNet 残差网络，对神经网络的理解是其他人没法比的。尽管 R-CNN 效果不错，但是他还有两个硬伤：

**硬伤一、算力冗余**。先生成候选区域，再对区域进行卷积，这里有两个问题：其一是候选区域会有一定程度的重叠，对相同区域进行重复卷积；其二是每个区域进行新的卷积需要新的存储空间。何恺明等人意识到这个可以优化，于是把先生成候选区域再卷积，变成了先卷积后生成区域。“简单地”改变顺序，不仅减少存储量而且加快了训练速度。

**硬伤二、图片缩放**。无论是剪裁 (Crop) 还是缩放 (Warp)，在很大程度上会丢失图片原有的信息导致训练效果不好，如图 2.2所示。直观的理解，把车剪裁成一个门，人看到这个门也不好判断整体是一辆车；把一座高塔缩放成一个胖胖的塔，人看到也没很大把握直接下结论。人都做不到，机器的难度就可想而知了。

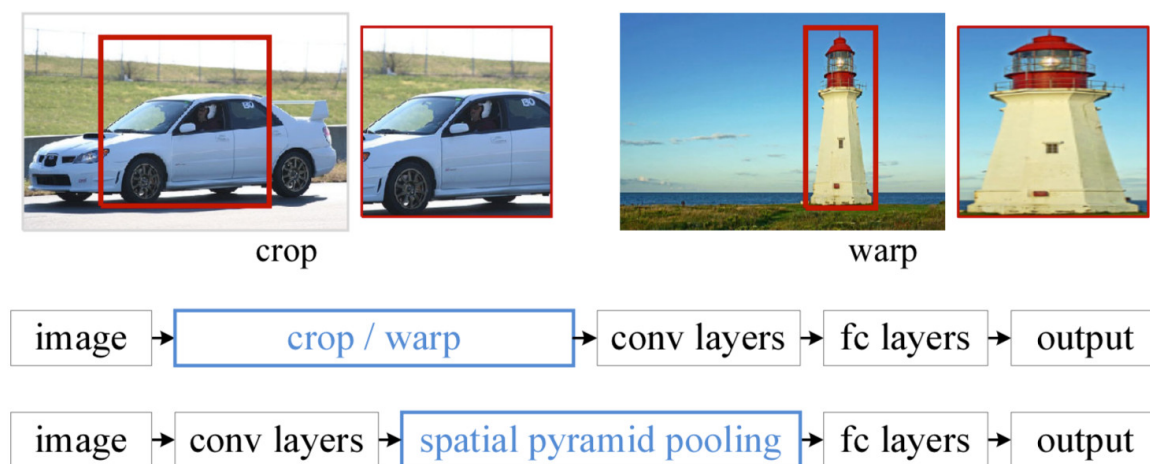


图 2.2: 因剪裁和缩放导致的信息丢失

何恺明等人发现了这个问题，于是思索有什么办法能不对图片进行变形，将图

片原汁原味地输入进去学习。最后，他们发现问题的根源是 FC Layer(全连接层) 需要确定输入维度，于是他们在输入全连接层前定义一个特殊的池化层，将输入的任意尺度 feature maps 组合成特定维度的输出，这个组合可以是不同大小的拼凑，如同拼凑七巧板般。举个例子，我们要输入的维度  $64 \times 256$ ，那么我们可以这样组合  $32 \times 256 + 16 \times 256 + 8 \times 256 + 8 \times 256$ 。

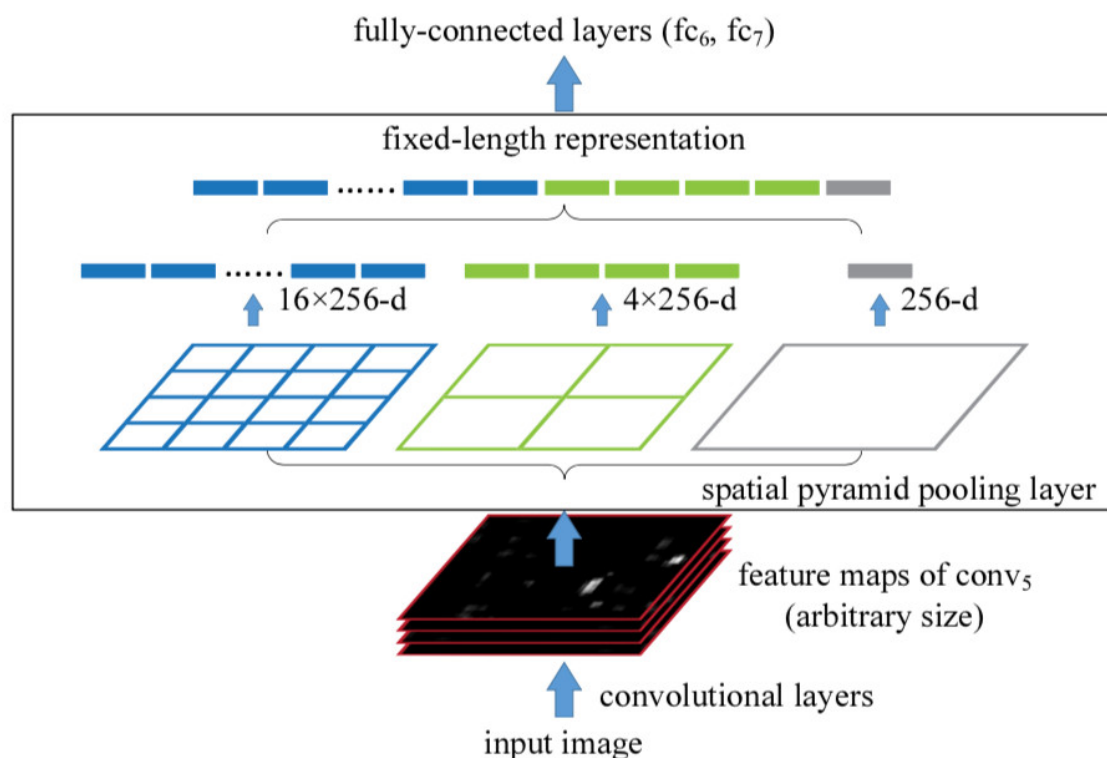


图 2.3: 输入维度的组合方式

SPP Net 的出现是如同一道惊雷，不仅减少了计算冗余，更重要的是打破了固定尺寸输入这一束缚，让后来者享受到这一缕阳光。

### 2.4.3 Fast R-CNN

继 2014 年的 RCNN 之后，Ross Girshick 在 15 年推出 Fast RCNN，构思精巧，流程更为紧凑，大幅提升了目标检测的速度。2015 年 Girshick, Ross 发表了 Fast r-cnn [2]，在这篇论文中，引用了 SPP Net 的工作，并且致谢其第一作者何恺明的慷慨解答。纵观全文，最大的建树就是将原来的串行结构改成并行结构。同样使用最大规模的网络，Fast RCNN 和 RCNN 相比，训练时间从 84 小时减少为 9.5 小时，测试时间从 47 秒减少为 0.32 秒。在 PASCAL VOC 2007 上的准确率相差无几，约在 66%-67% 之间。其算法框架如图 2.4

原来的 R-CNN 是先对候选框区域进行分类，判断有没有物体，如果有则对 Bounding Box 进行精修、回归。这是一个串联式的任务，那么势必没有并联的快，所以 Ross Girshick 就将原有结构改成并行——在分类的同时，对 Bbox 进行回归。这一改变将 Bbox 和 Clf 的 loss 结合起来变成一个 Loss 一起训练，并吸纳了 SPP Net



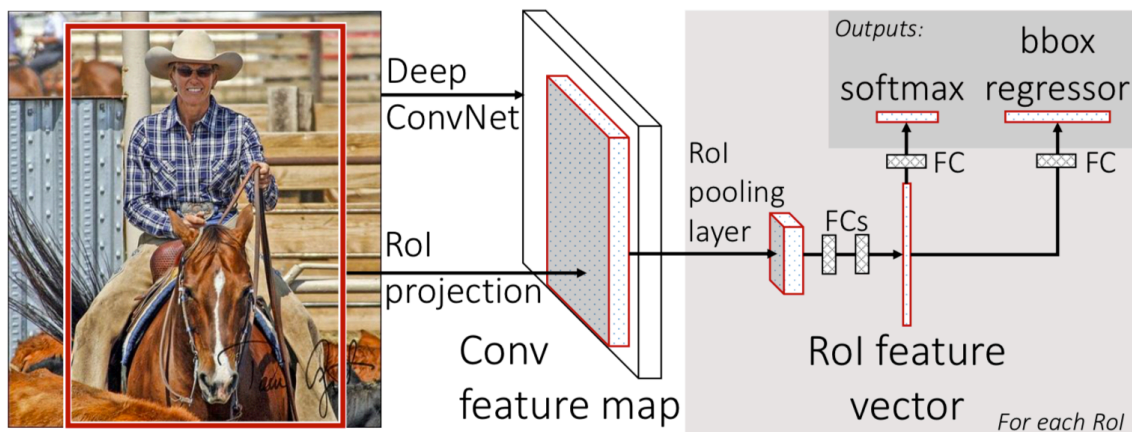


图 2.4: Fast R-CNN 算法框架

的优点，最终不仅加快了预测的速度，而且提高了精度。

简单来说，RCNN 使用以下四步来实现目标检测：

- > 在图像中确定 1000-2000 个候选框
- > 对于每个候选框内图像块，使用深度网络提取特征
- > 对候选框中提取出的特征，使用分类器判别是否属于一个特定类
- > 对于属于某一特征的候选框，用回归器进一步调整位置

Fast RCNN 方法解决了 RCNN 方法三个问题：

**问题一、测试时速度慢。**RCNN 一张图像内候选框之间大量重叠，提取特征操作冗余，本文将整张图像归一化后直接送入深度网络。在邻接时，才加入候选框信息，在末尾的少数几层处理每个候选框。

**问题二、训练时速度慢。**原因同问题一，在训练时，本文先将一张图像送入网络，紧接着送入从这幅图像上提取出的候选区域。这些候选区域的前几层特征不需要再重复计算。

**问题三、训练所需空间大。**RCNN 中独立的分类器和加归器需要大量特征作为训练样，本文把类别判断和位置精调统一用深度网络实现，不再需要额外存储。其网络模型如图 2.5

**损失函数：**loss\_cls 层评估分类代价。由真实分类  $u$  对应的概率决定：

$$L_{cls} = -\log p_u$$

loss\_bbox 评估检测框定位代价。比较真实分类对应的预测参数  $t^u$  和真实平移缩放为  $v$  的差别：

$$L_{loc} = \sum_{i=1}^4 g(t_i^u - v_i)$$

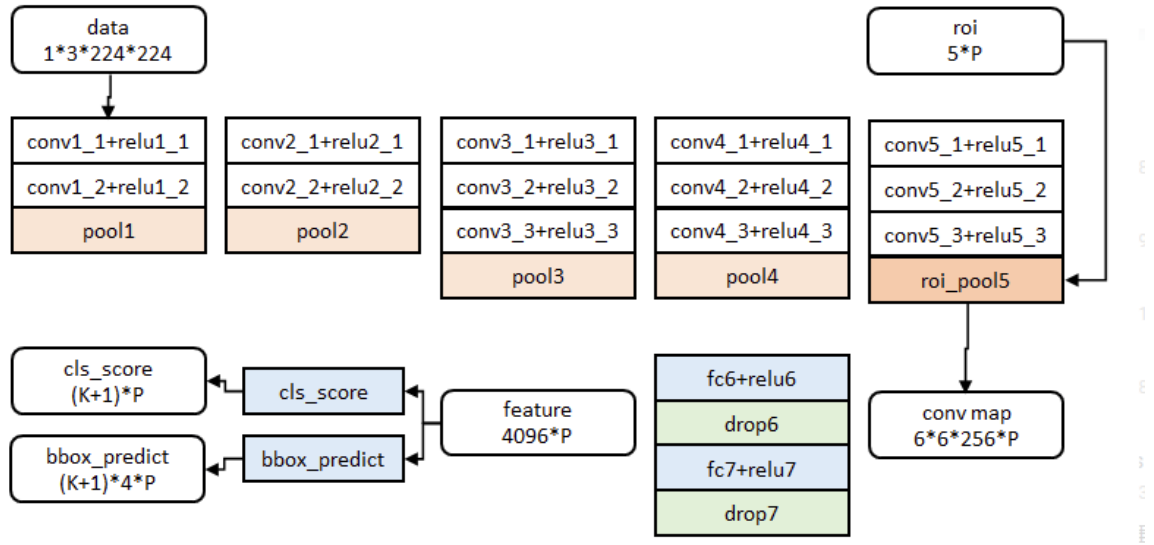


图 2.5: Fast R-CNN 网络模型

$g$  为 Smooth L1 误差，对 outlier 不敏感：

$$g(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (1)$$

总代价为两者加权和，如果分类为背景则不考虑定位代价：

$$L = \begin{cases} L_{cls} + \lambda L_{loc} & u \\ L_{cls} & u \end{cases} \quad (2)$$

#### 2.4.4 Faster R-CNN

继 RCNN, Fast RCNN 之后，目标检测界的领军人物 Ross Girshick 团队在 2015 年的又一力作。简单网络目标检测速度达到 17fps，在 PASCAL VOC 上准确率为 59.9%；复杂网络达到 5fps，准确率 78.8%。在 Faster R-CNN 前，我们生产候选区域都是用的一系列启发式算法，基于 Low Level 特征生成区域。这样就有两个问题：

**第一个问题**是生成区域的靠谱程度随缘，而“two stage”算法正是依靠生成区域的靠谱程度——生成大量无效区域则会造成算力的浪费、少生成区域则会漏检；

**第二个问题**是生成候选区域的算法是在 CPU 上运行的，而我们的训练在 GPU 上面，跨结构交互必定会有损效率。

那么怎么解决这两个问题呢？于是乎，任少卿等人提出了一个 Region Proposal Networks 的概念，利用神经网络自己学习去生成候选区域。其结构如图 2.6。

这种生成方法同时解决了上述的两个问题，神经网络可以学到更加高层、语义、抽象的特征，生成的候选区域的可靠程度大大提高；可以从上图看出 RPNs 和 RoI



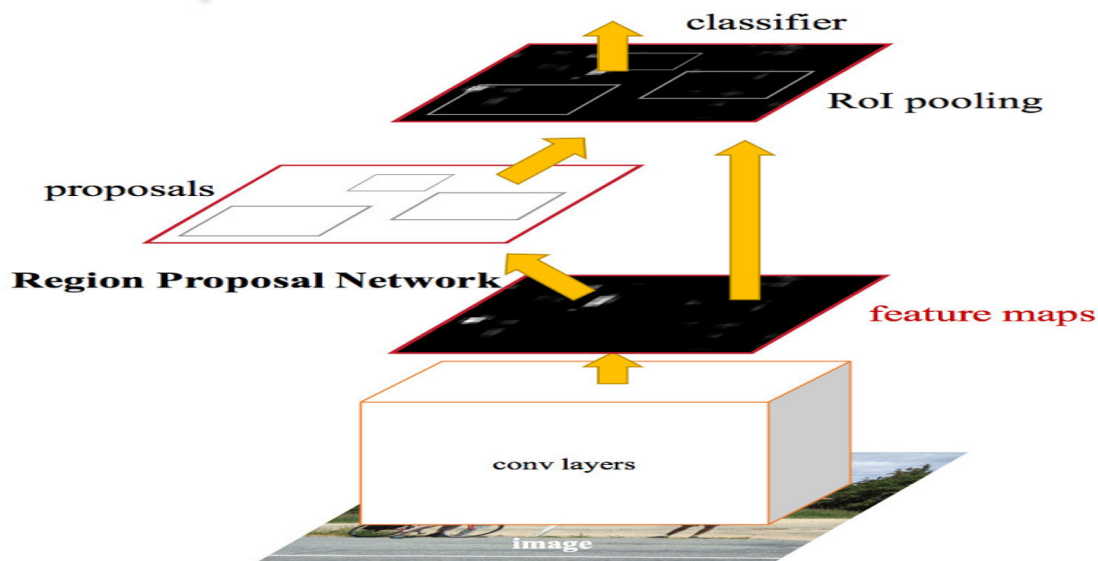


图 2.6: Faster RCNN 算法架构

Pooling 共用前面的卷积神经网络——将 RPNs 嵌入原有网络，原有网络和 RPNs 一起预测，大大地减少了参数量和预测时间。在 RPNs 中引入了 anchor 的概念 feature map 中每个滑窗位置都会生成 k 个 anchors，然后判断 anchor 覆盖的图像是前景还是背景，同时回归 Bbox 的精细位置，预测的 Bbox 更加精确。

从 RCNN 到 fast RCNN，再到本文的 faster RCNN，目标检测的四个基本步骤（候选区域生成，特征提取，分类，位置精修）终于被统一到一个深度网络框架之内。所有计算没有重复，完全在 GPU 中完成，大大提高了运行速度。

Faster RCNN 着重解决了三个问题：

- > 如何设计区域生成网络
- > 如何训练区域生成网络
- > 如何让区域生成网络和 Faster RCNN 网络共享特征提取网络

### 区域生成网络 (RPN): 结构

其结构如图 2.7

1、**特征提取**：原始特征提取（上图灰色方框）包含若干层 conv+relu，直接套用 ImageNet 上常见的分类网络即可。本文试验了两种网络：5 层的 ZF<sup>2</sup>，16 层的 VGG-16，具体结构不再赘述。额外添加一个 conv+relu 层，输出 51\*39\*256 维特征 (feature)。

2、**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 128<sup>2</sup>, 256<sup>2</sup>, 512<sup>2</sup> × 三种比例 1:1, 1:2, 2:1。

<sup>2</sup>介绍一下这个网络

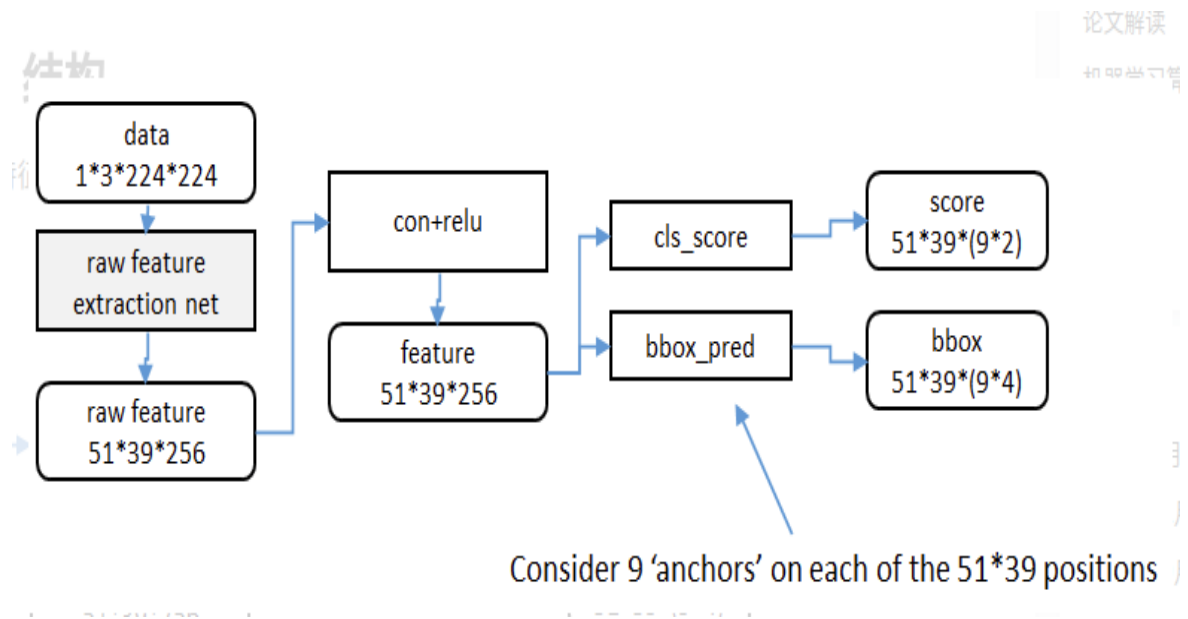


图 2.7: RPN 结构

这些候选窗口称为 anchors。下图示出 51\*39 个 anchor 中心，以及 9 种 anchor 示例 2.8。

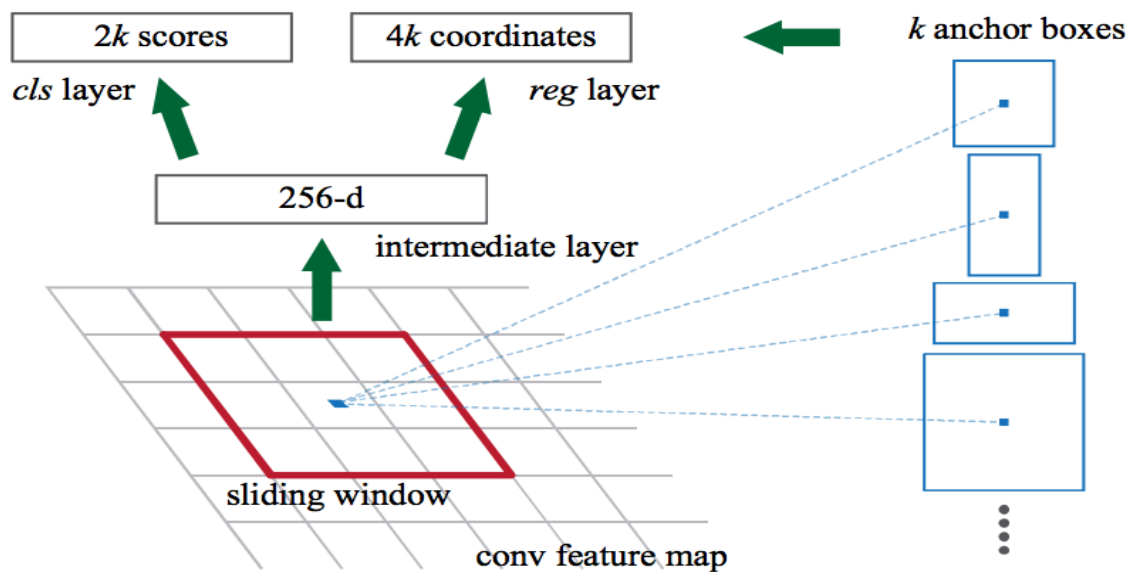


图 2.8: Faster RCNN anchor 示意图

**3、窗口分类和位置精修:** 分类层 (cls\_score) 输出每一个位置上, 9 个 anchor 属于前景和背景的概率; 窗口回归层 (bbox\_pred) 输出每一个位置上, 9 个 anchor 对应窗口应该平移缩放参数。对于每一个位置来说, 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。

### 2.4.5 Mask R-CNN

时隔一年，何恺明团队再次更新了 R-CNN 家族，改进 Faster R-CNN 并使用新的 backbone 和 FPN 创造出了 Mask R-CNN。

**加一条通道。**我们纵观发展历史，发现 SPP Net 升级为 Fast R-CNN 时结合了两个 loss，也就是说网络输入了两种信息去训练，结果精度大大提高了。何恺明他们就思考着再加一个信息输入，即图像的 Mask，信息变多之后会不会有提升呢？于是乎 Mask R-CNN 就这样出来了，不仅可以做“目标检测”还可以同时做“语义分割”，将两个计算机视觉基本任务融入一个框架。没有使用什么 trick，性能却有了较为明显的提升，这个升级的版本让人们不无啧啧惊叹。作者称其为 meta algorithm，即一个基础的算法，只要需要“目标检测”或者“语义分割”都可以使用这个作为 Backbone。

## 2.5 YOLO 系列

You only look once: Unified, real-time object detection(YOLO) [6] 是单阶段方法的开山之作。它将检测任务表述成一个统一的、端到端的回归问题，并且以只处理一次图片同时到位置和分类而得名。“one stage”的想法就比较暴力，给定一张图像，使用回归的方式输出这个目标的边框和类别。“one stage”最核心的还是利用了分类器优秀的分类效果，首先给出一个大致的范围（最开始就是全图）进行分类，然后不断迭代这个范围直到一个精细的位置，



图 2.9: YOLO

如图 2.9从蓝色的框框到红色的框框。这就是“one stage”回归的思想，这样做的优点就是快，但是会有许多漏检。

### 2.5.1 YOLO

YOLO 就是使用回归这种做法的典型算法。首先将图片 Resize 到固定尺寸，然后通过一套卷积神经网络，最后接上 FC(全连接层) 直接输出结果，这就他们整个网络的基本结构。更具体地做法，是将输入图片划分成一个  $S \times S$  的网格，每个网格负责检测网格里面的物体是啥，并输出 Bbox Info 和置信度。这里的置信度指的是该网格内含有何物体和预测这个物体的准确度。

更具体的是如下定义：

$$Pr(Class_i|Object) * Pr(Object) * IOU_{pred}^{truth} = Pr(Class_i) * IOU_{pred}^{truth}$$

这个想法其实就是一个简单的分而治之想法，将图片卷积后提取的特征图分为  $S \times S$  块，然后利用优秀的分类模型对每一块进行分类，将每个网格处理完使用 NMS(非极大值抑制) 的算法去除重叠的框，最后得到我们的结果。

### 2.5.2 SSD

YOLO 这样做的确非常快，但是问题就在于这个框有点大，就会变得粗糙——小物体就容易从这个大网中漏出去，因此对小物体的检测效果不好。

所以 SSD 就在 YOLO 的主意上添加了 Faster R-CNN 的 Anchor 概念，并融合不同卷积层的特征做出预测。

第三章将详细阐述 SSD 算法

### 2.5.3 YOLO9000


到了 SSD，回归方法的目标检测应该一统天下了，但是 YOLO 的作者升级做了一个 YOLO9000——号称可以同时识别 9000 类物体的实时监测算法。讲道理，YOLO9000 更像是 SSD 加了一些 Trick，而并没有什么本质上的进步：

加了 BN 层，扩大输入维度，使用了 Anchor 训练的时候数据增强，SSD 和 YOLO9000 可以归为一类。

## 2.6 本章小结

回顾过去，从 YOLO 到 SSD，人们兼收并蓄把不同思想融合起来。YOLO 使用了分治思想，将输入图片分为  $S \times S$  的网格，不同网格用性能优良的分类器去分类。SSD 将 YOLO 和 Anchor 思想融合起来，并创新使用 Feature Pyramid 结构。但是 Resize 输入，必定会损失许多的信息和一定的精度，这也许是“one stage”快的原因。无论如何，YOLO 和 SSD 这两篇论文都是让人不得不赞叹他们想法的精巧，让人受益良多在目标检测中有两个指标：快 (Fast) 和准 (Accurate)。“one stage”代表的是快，但是最后在快和准中找到了平衡，第一是快，第二是准。“two stage”代表的是准，虽然没有那么快但是也有 6 FPS 可接受的程度，第一是准，第二是快。两类算法都有其适用的范围，比如说实时快速动作捕捉，“one stage”更胜一筹；复杂、多物体重叠，“two stage”当仁不让。没有不好的算法，只有不合适的使用场景。我相信 Mask R-CNN 并不是“目标检测”的最终答案，展望未来。

### 3.1 瓦片损害检测流程

- 
- ```
graph LR; A[瓦片检测流程图] --> B[无人机拍照]; A --> C[将图像切割成小图]; A --> D[将小图输入网络模型中送检]; A --> E[输出检测结果];
```
- 瓦片检测流程图
- 无人机拍照
  - 将图像切割成小图
  - 将小图输入网络模型中送检
  - 输出检测结果

### 3.2 SSD 算法核心思想

**SSD**

Image: 300x300

VGG-16 through Conv5\_3 layer

Classifier: Conv: 3x3x(4x(Classes+4))

Classifier: Conv: 3x3x(6x(Classes+4))

Extra Feature Layers

Conv#\_2: 10x10x512

Conv#\_3: 5x5x256

Conv#\_4: 3x3x256

Conv#\_5: 3x3x256

Detections: 8732 per Class

Non-Maximum Suppression

74.3mAP  
59FPS

---

**YOLO**

Image: 448x448

YOLO Customized Architecture

7x7 Conv: 1024

7x7 Conv: 30

Detections: 88 per class

Non-Maximum Suppression

63.4mAP  
45FPS

下面将 SSD 核心设计思想总结为以下三点：

- 第 21 页 共 38 页



所谓多尺度采用大小不同的特征图，CNN 网络一般前面的特征图比较大，后面会逐渐采用 stride=2 的卷积或者 pool 来降低特征图大小，如图 3.3所示

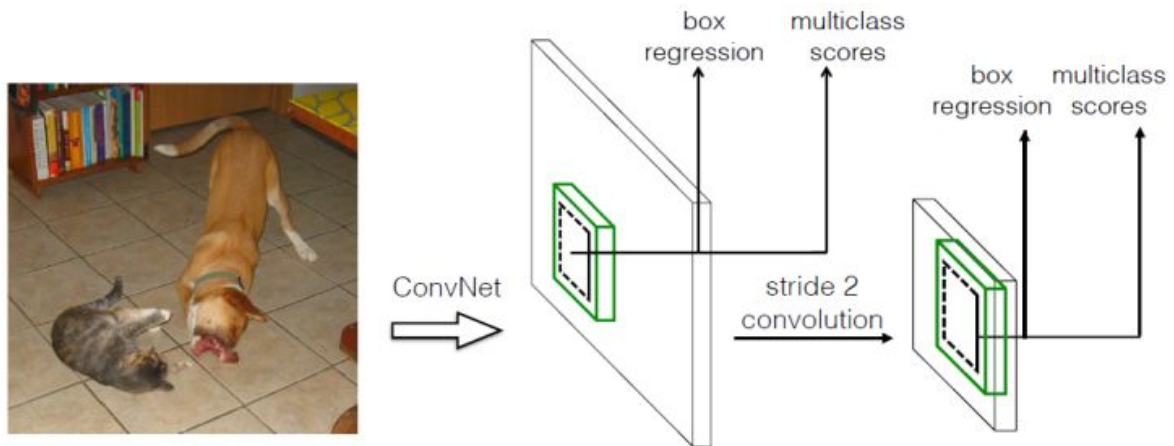


图 3.3: 采用多尺度用于检测

一个比较大的特征图和一个比较小的特征图，它们都用来做检测。这样做的好处是比较大的特征图来用来检测相对较小的目标，而小的特征图负责检测大目标，如图 3.4所示，8x8 的特征图可以划分更多的单元，但是其每个单元的先验框尺度比较小。

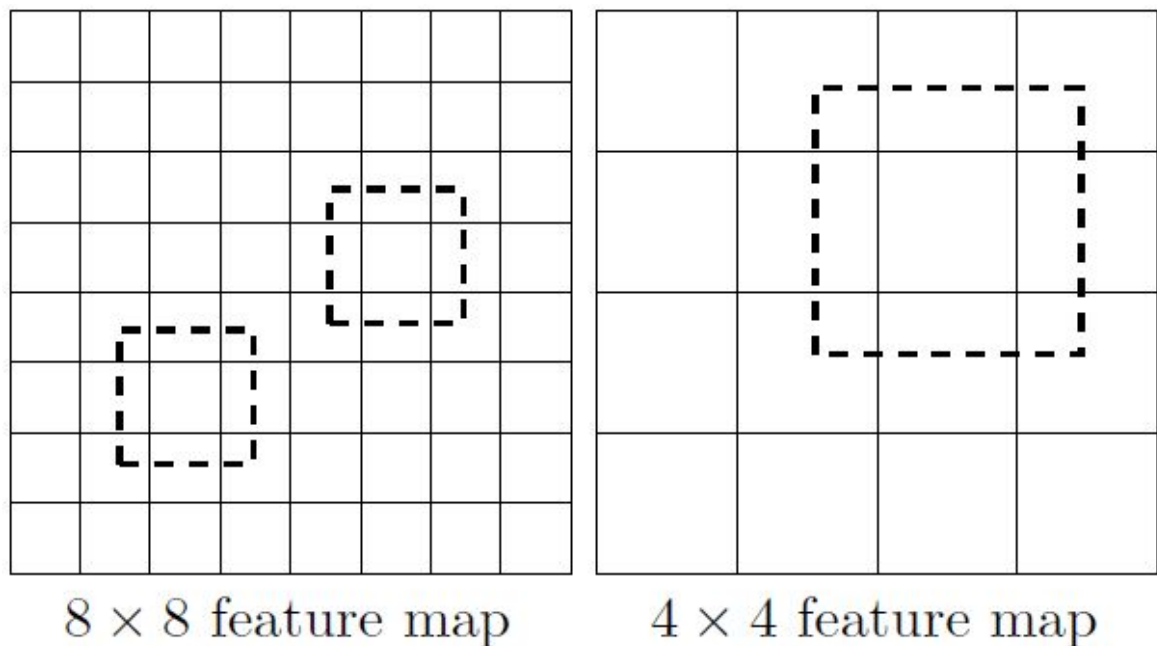


图 3.4: 8x8 与 4x4 的特征图

## 2、采用卷积进行检测

与 YOLO 最后采用全连接层不同，SSD 直接采用卷积对不同的特征图来进行提取检测结果。对于形状为  $m \times n \times p$  的特征图，只需要采用  $3 \times 3 \times p$  这样比较小的卷积核得到检测值。

### 3、设置先验框

在 YOLO 中，每个单元预测多个边界框，但是其都是相对这个单元本身 (正方块)，但是真实目标的形状是多变的，YOLO 需要在训练过程中自适应目标的形状。而 SSD 借鉴了 Faster R-CNN 中 anchor 的理念，每个单元设置尺度或者长宽比不同的先验框，预测的边界框 (bounding boxes) 是以这些先验框为基准的，在一定程度上减少训练难度。一般情况下，每个单元会设置多个先验框，其尺度和长宽比存在差异，如图 3.5 所示，可以看到每个单元使用了 4 个不同的先验框，图片中猫和狗分别采用最适合它们形状的先验框来进行训练，后面会详细讲解训练过程中的先验框匹配原则。

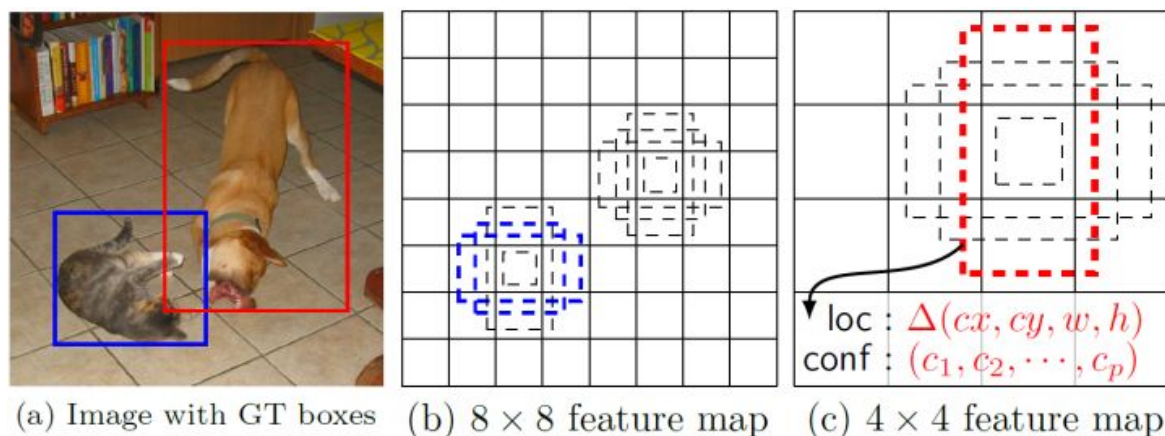


图 3.5: SSD 算法中的先验框

### 3.3 SSD 模型结构

SSD 网络主体设计的思想是特征分层提取，并依次进行边框回归和分类。因为不同层次的特征图能代表不同层次的语义信息，低层次的特征图能代表低层语义信息 (含有更多的细节)，能提高语义分割质量，适合小尺度目标的学习。高层次的特征图能代表高层语义信息，能光滑分割结果，适合对大尺度的目标进行深入学习。所以作者提出的 SSD 的网络理论上能适合不同尺度的目标检测。

所以 SSD 网络中分为了 6 个 stage，每个 stage 能学习到一个特征图，然后进行边框回归和分类。SSD 网络以 VGG16 的前 5 层卷积网络作为第 1 个 stage，然后将 VGG16 中的 fc6 和 fc7 两个全连接层转化为两个卷积层 Conv6 和 Conv7 作为网络的第 2、第 3 个 stage。接着在此基础上，SSD 网络继续增加了 Conv8、Conv9、Conv10 和 Conv11 四层网络，用来提取更高层次的语义信息。如下图 3.1 所示就是 SSD 的网络结构。在每个 stage 操作中，网络包含了多个卷积层操作，每个卷积层操作基本上都是小卷积。

骨干网络：SSD 前面的骨干网络选用的 VGG16 的基础网络结构，如上图所示，虚线框内的是 VGG16 的前 5 层网络。然后后面的 Conv6 和 Conv7 是将 VGG16 的后两层全连接层网络 (fc6, fc7) 转换而来。另外：在此基础上，SSD 网络继续增加了 Conv8 和 Conv9、Conv10 和 Conv11 四层网络。图中所示，立方体的长高表示特征

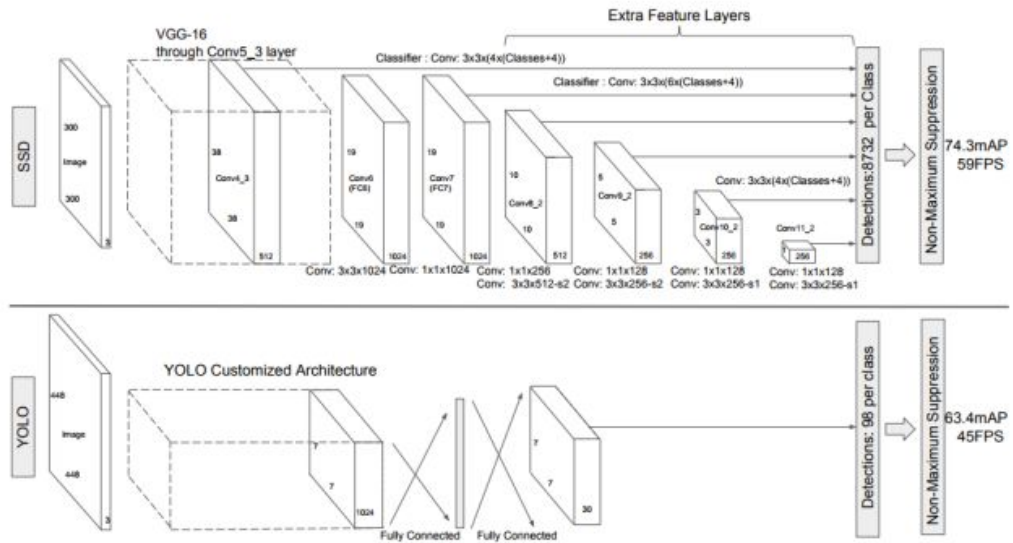


图 3.6: RCNN

图的大小，厚度表示是 channel。

### 3.4 损失函数

#### 3.4.1 SSD 中采用的损失函数

**联合 LOSS FUNCTION:**SSD 网络对于每个 Stage 输出的特征图都进行边框回归和分类操作。SSD 网络中作者设计了一个联合损失函数

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + \alpha L_{loc}(x, l, g))$$

其中：

- 1、 $L_{conf}$  代表的是分类误差，采用的是多分类的 softmax loss
- 2、 $L_{loc}$  代表的是回归误差，采用的是 Smooth L1 loss
- 3、 $\alpha$  取 1

回归 loss,smoothL1:

$$L_{loc}(x, l, g) = \sum_{i \in Pos} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k smooth_{L1}(l_i^m - g_j^m)$$

$$g_j^{cx} = (g_j^{cx} - d_i^{cx}) / d_i^w$$

#### 3.4.2 改进损失函数-Focus Loss

我们知道 object detection 的算法主要可以分为两大类：two-stage detector 和 one-stage detector。前者是指类似 Faster RCNN，RFCN 这样需要 region proposal 的检测算法，这类算法可以达到很高的准确率，但是速度较慢。虽然可以通过减少



proposal 的数量或降低输入图像的分辨率等方式达到提速，但是速度并没有质的提升。后者是指类似 YOLO，SSD 这样不需要 region proposal，直接回归的检测算法，这类算法速度很快，但是准确率不如前者。作者提出 focal loss 的出发点也是希望 one-stage detector 可以达到 two-stage detector 的准确率，同时不影响原有的速度。

既然有了出发点，那么就要找 one-stage detector 的准确率不如 two-stage detector 的原因，作者认为原因是：样本的类别不均衡导致的。我们知道在 object detection 领域，一张图像可能生成成千上万的 candidate locations，但是其中只有很少一部分是包含 object 的，这就带来了类别不均衡。那么类别不均衡会带来什么后果呢？引用原文讲的两个后果：

(1) training is inefficient as most locations are easy negatives that contribute no useful learning signal;

(2) en masse, the easy negatives can overwhelm training and lead to degenerate models.

什么意思呢？负样本数量太大，占总的 loss 的大部分，而且多是容易分类的，因此使得模型的优化方向并不是我们所希望的那样。其实先前也有一些算法来处理类别不均衡的问题，比如 OHEM (online hard example mining)，OHEM 的主要思想可以用原文的一句话概括：In OHEM each example is scored by its loss, non-maximum suppression (nms) is then applied, and a minibatch is constructed with the highest-loss examples。OHEM 算法虽然增加了错分类样本的权重，但是 OHEM 算法忽略了容易分类的样本。因此针对类别不均衡问题，作者提出一种新的损失函数：focal loss，这个损失函数是在标准交叉熵损失基础上修改得到的。这个函数可以通过减少易分类样本的权重，使得模型在训练时更专注于难分类的样本。为了证明 focal loss 的有效性，作者设计了一个 dense detector：RetinaNet，并且在训练时采用 focal loss 训练。实验证明 RetinaNet 不仅可以达到 one-stage detector 的速度，也能有 two-stage detector 的准确率。

### 标准交叉熵损失函数

原来的分类 loss 是各个训练样本交叉熵的直接求和，也就是各个样本的权重是一样的。如下图所示：

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases} \quad (3)$$

其中：

CE 表示 cross entropy,  $p$  表示预测样本属于 1 的概率， $y$  表示 label， $y$  的取值为 +1, -1，这里仅仅以二分类为例，多分类以此类推。为了表示简便，我们用  $p_t$  表示样本属于 true class 的概率。所以上式可以写成

$$CE(p, y) = CE(p_t) = -\log(p_t)$$

### 平衡交叉熵

既然“one stage detector”在训练的时候正负样本的数量差距很大，那么一种常见的做法就是给正负样本加上权重，负样本出现的频次多，那么就降低负样本的权重，正样本数量少，就相对提高正样本的权重，如下式所示：

$$CE(p_t) = -\alpha \log(p_t)$$

### focal-loss 定义

作者实际上解决的是 easy example 和 hard example 不均衡的问题，这个和训练时候正负样本不均衡是两码事，因为正负样本里面都会有简单的样本和容易分错的样本。作者提出的 focal loss，相当于是对各个样本加上了各自的权重，这个权重是和网络预测该样本属于 true class 的概率相关的，显然，如果网络预测的该样本属于 true class 的概率很大，那么这个样本对网络来说就属于 easy(well-classified) example。如果网络预测的该样本属于 true class 的概率很小，那么这个样本对网络来说就属于 hard example。为了训练一个有效的 classification part，显然应该降低绝大部分 easy example 的权重，相对增大 hard example 的权重。作者提出的 focal loss 如下式所示：

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t)$$

参数  $\gamma$  大于 0。当参数  $\gamma = 0$  的时候，就是普通的交叉熵，作者的实验中发现  $\gamma = 2$  效果最高。可以看到，当一定的时候，比如等于 2，一样 easy example( $p_t = 0.9$ ) 的 loss 要比标准的交叉熵 loss 小 100+ 倍，当  $p_t = 0.968$  时，要小 1000+ 倍，但是对于 hard example( $p_t < 0.5$ )，loss 最多小了 4 倍。这样的话 hard example 的权重相对就提升了很多。实际实验中，作者和 3.2 一样，对正负样本又做了一个 reweighting

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

## 3.5 SSD 模型训练

训练 SSD 和基于 region proposal 方法的最大区别就是：SSD 需要精确的将 ground truth 映射到输出结果上。这样才能提高检测的准确率。文中主要采取了以下几个技巧来提高检测的准确度。

**匹配策略** Default boxes 生成器 Hard Negative Mining Data Augmentation

### 1. 匹配策略

这里的匹配是指的 ground truth 和 Default box 的匹配。这里采取的方法与 Faster R-CNN 中的方法类似。主要是分为两步：第一步是根据最大的 overlap 将 ground truth 和 default box 进行匹配 (根据 ground truth 找到 default box 中 IOU 最大的作为正样本)，第二步是将 default boxes 与 overlap 大于某个阈值 (目标检测中通常选取 0.5，Faster R-CNN 中选取的是 0.7) 的 ground truth 进行匹配。

## 2. Default Boxes 生成器

3. Hard Negative Mining 经过匹配策略会得到大量的负样本，只有少量的正样本。这样导致了正负样本不平衡，经过试验表明，正负样本的不均衡是导致检测正确率低下的一个重要原因。所以在训练过程中采用了 Hard Negative Mining 的策略，根据 Confidence Loss 对所有的 box 进行排序，使得正负样本的比例控制在 1:3 之内，经过作者实验，这样做能提高 4% 左右的准确度。

## 4. Data Augmentation

为了模型更加鲁棒，需要使用不同尺度目标的输入，作者对数据进行了增强处理。

1、使用整张图像作为输入

2、使用 IOU 和目标物体为 0.1、0.3、0.5、0.7 和 0.9 的 patch，这些 patch 在原图的大小的  $[0.1, 1]$  之间，相应的宽高比在  $[1/2, 2]$  之间。

3、随机采取一个 patch

## 3.6 算法改进方案

### 3.6.1 Soft-NMS

NMS 算法的大致过程可以看原文这段话：

First, it sorts all detection boxes on the basis of their scores. The detection box M with the maximum score is selected and all other detection boxes with a significant overlap (using a pre-defined threshold) with M are suppressed. This process is recursively applied on the remaining boxes.

那么传统的 NMS 算法存在什么问题呢？可以看 Figure1。在 Figure1 中，检测算法本来应该输出两个框，但是传统的 NMS 算法可能会把 score 较低的绿框过滤掉（如果绿框和红框的 IOU 大于设定的阈值就会被过滤掉），导致只检测出一个 object（一个马），显然这样 object 的 recall 就比较低了。可以看出 NMS 算法是略显粗暴（hard），因为 NMS 直接将和得分最大的 box 的 IOU 大于某个阈值的 box 的得分置零，那么有没有 soft 一点的方法呢？这就是本文提出 Soft NMS。那么 Soft-NMS 算法到底是什么样呢？简单讲就是：

An algorithm which decays the detection scores of all other objects as a continuous function of their overlap with M.

换句话说就是用稍低一点的分数来代替原有的分数，而不是直接置零。另外由于 Soft NMS 可以很方便地引入到 object detection 算法中，不需要重新训练原有的模型，因此这是该算法的一大优点。

**这里插入一个图片**

传统的 NMS 处理方法可以通过以下的分数重置函数 (Rescoring Function) 来

表示：

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ 0, & iou(M, b_i) \geq N_t \end{cases} \quad (4)$$

在这个公式中，NMS 采用了硬阈值来判断相邻检测框是否保留。但是，换一种方法，假设我们对一个与 M 高度重叠的检测框  $b_i$  的检测分数进行衰减，而非全部抑制。如果检测框  $b_i$  中包含不同于 M 中的物体，那么在检测阈值比较低的情况下，该物体并不会错过检测。但是，如果  $b_i$  中并不包含任何物体，即使在衰减过后， $b_i$  的分数仍然较高，它还是会产生一个假阳性的结果，因此，在使用 NMS 做物体处理的时候，需要注意以下几点：

1、相邻检测框的检测分数应该被降低，从而减少假阳性结果，但是，衰减后的分数仍然应该比明显的假阳性结果要高。

2、通过较低的 NMS 重叠阈值来移除所有相邻检测框并不是最优解，并且很容易导致错过被检测物体，特别是在物体高度重叠的地方。

3、当 NMS 采用一个较高的重叠阈值时，平均准确率可能会相应降低。

值得注意的是，soft-NMS 也是一种贪心算法，并不能保证找到全局最优的检测框分数重置。但是，soft-NMS 算法是一种更加通用的非最大抑制算法，传统的 NMS 算法可以看做是它的一个采用不连续二值权重函数的特例。除了以上这两种分数重置函数，我们也可以考虑开发其他包含更多参数的分数重置函数，比如 Gompertz 函数等。但是它们在完成分数重置的过程中增加了额外的参数。

### 3.6.2 数据增强

采用数据扩增 (Data Augmentation) 可以提升 SSD 的性能，主要采用的技术有水平翻转 (horizontal flip)，随机裁剪加颜色扭曲 (random crop & color distortion)，随机采集块域 (Randomly sample a patch) 获取小目标训练样本

## 3.7 本章小结

## 第四章 瓦片损害检测算法实现

### 4.1 图像预处理

#### 4.1.1 标注工具

**labelImg 介绍:** 图片标注主要是用来创建自己的数据集，方便进行深度学习训练。本文使用 labelImg



图 4.1: labelImg 界面

#### 4.1.2 数据集

**VOC 数据集介绍:** PASCAL VOC 为图像识别和分类提供了一整套标准化的优秀的数据集。

其目录结构

- JPEGImage
- Annotations
- ImageSet

#### 4.1.3 Pytorch 介绍

PyTorch 是一个比较新的深度学习框架，正在研究人员中迅速普及。和深度学习框架 Chainer 类似，PyTorch 支持动态计算图，这个功能使 PyTorch 对使用文本与时间序列的研究者和工程师很有吸引力。TensorFlow 解决了质量控制和包装的问题。它提供了一种 Theano 风格的编程模式，所以它是一种非常底层的深度学习框架。由于 TensorFlow 本身是非常底层的框架，所以许多基于 TensorFlow 的前端框架应运而生，比如 TF-slim 和 Keras。这样的框架目前有 10 到 15 个，仅 Google 就可能有四五个。

Torch 的思想和 Theano 一直略有不同。TensorFlow 是一个更好的 Theano 风格的框架，并且我们在 Torch 中注入了强制的思想，这意味着你可以立即运行你的计

算。调试应该是平稳顺利的，用户在调试过程中不应当遇到难题，无论使用 Python 调试器还是像 GDB 或其他类似的东西。

## 4.2 网络模型实现

Listing 1 SSD 网络模型

```
1: class SSD300(nn.Module):
2:     input_size = 300
3:
4:     def __init__(self):
5:         super(SSD300, self).__init__()
6:
7:         # model
8:         self.base = self.VGG16()
9:         self.norm4 = L2Norm2d(20)
10:
11:         self.conv5_1 = nn.Conv2d(
12:             512, 512,
13:             kernel_size=3,
14:             padding=1,
15:             dilation=1
16:         )
17:         self.conv5_2 = nn.Conv2d(
18:             512, 512,
19:             kernel_size=3,
20:             padding=1,
21:             dilation=1
22:         )
23:         self.conv5_3 = nn.Conv2d(
24:             512, 512,
25:             kernel_size=3,
26:             padding=1,
27:             dilation=1
28:         )
29:
30:         self.conv6 = nn.Conv2d(
31:             512, 1024,
32:             kernel_size=3,
33:             padding=6,
34:             dilation=6
35:         )
36:
37:         self.conv7 = nn.Conv2d(
```

```

38:         1024, 1024,
39:         kernel_size=1
40:     )
41:
42:     self.conv8_1 = nn.Conv2d(
43:         1024, 256,
44:         kernel_size=1
45:     )
46:     self.conv8_2 = nn.Conv2d(
47:         256, 512,
48:         kernel_size=3,
49:         padding=1,
50:         stride=2
51:     )
52:
53:     self.conv9_1 = nn.Conv2d(
54:         512, 128,
55:         kernel_size=1
56:     )
57:     self.conv9_2 = nn.Conv2d(
58:         128, 256,
59:         kernel_size=3,
60:         padding=1,
61:         stride=2
62:     )
63:
64:     self.conv10_1 = nn.Conv2d(
65:         256, 128,
66:         kernel_size=1
67:     )
68:     self.conv10_2 = nn.Conv2d(
69:         128, 256,
70:         kernel_size=3
71:     )
72:
73:     self.conv11_1 = nn.Conv2d(
74:         256, 128,
75:         kernel_size=1
76:     )
77:     self.conv11_2 = nn.Conv2d(
78:         128, 256,
79:         kernel_size=3

```

```
80:         )
81:
82:     # multibox layer
83:     self.multibox = MultiBoxLayer()
84:
85:     def forward(self, x):
86:         hs = []
87:         h = self.base(x)
88:         hs.append(self.norm4(h)) # conv4_3
89:
90:         h = F.max_pool2d(
91:             h, kernel_size=2,
92:             stride=2, ceil_mode=True
93:         )
94:
95:         h = F.relu(self.conv5_1(h))
96:         h = F.relu(self.conv5_2(h))
97:         h = F.relu(self.conv5_3(h))
98:         h = F.max_pool2d(
99:             h, kernel_size=3,
100:             padding=1, stride=1,
101:             ceil_mode=True
102:         )
103:
104:         h = F.relu(self.conv6(h))
105:         h = F.relu(self.conv7(h))
106:         hs.append(h) # conv7
107:
108:         h = F.relu(self.conv8_1(h))
109:         h = F.relu(self.conv8_2(h))
110:         hs.append(h) # conv8_2
111:
112:         h = F.relu(self.conv9_1(h))
113:         h = F.relu(self.conv9_2(h))
114:         hs.append(h) # conv9_2
115:
116:         h = F.relu(self.conv10_1(h))
117:         h = F.relu(self.conv10_2(h))
118:         hs.append(h) # conv10_2
119:
120:         h = F.relu(self.conv11_1(h))
121:         h = F.relu(self.conv11_2(h))
```



```

122:         hs.append(h) # conv11_2
123:
124:         loc_preds, conf_preds = self.multibox(hs)
125:         return loc_preds, conf_preds
126:
127:     def VGG16(self):
128:         '''VGG16 layers.'''
129:         cfg = [
130:             64, 64, 'M',
131:             128, 128, 'M',
132:             256, 256, 256,
133:             'M', 512, 512, 512
134:         ]
135:         layers = []
136:         in_channels = 3
137:         for x in cfg:
138:             if x == 'M':
139:                 layers += [
140:                     nn.MaxPool2d(kernel_size=2,
141:                                   stride=2, ceil_mode=True)
142:                 ]
143:             else:
144:                 layers += [
145:                     nn.Conv2d(in_channels,
146:                               x, kernel_size=3, padding=1),
147:                     nn.ReLU(True)
148:                 ]
149:                 in_channels = x
150:         return nn.Sequential(*layers)

```

### 4.3 损失函数

Listing 2 损失函数

```

1: #####
2: # loc_loss = SmoothL1Loss(pos_loc_preds, pos_loc_targets)
3: #####
4: pos_mask = pos.unsqueeze(2).expand_as(loc_preds)
5: pos_loc_preds = loc_preds[pos_mask].view(-1,4)
6: pos_loc_targets = loc_targets[pos_mask].view(-1,4)
7: loc_loss = F.smooth_l1_loss(
8:     pos_loc_preds,
9:     pos_loc_targets,
10:    size_average=False

```

```
11: )
12:
13: #####
14: # conf_loss = CrossEntropyLoss(pos_conf_preds, pos_conf_targets)
15: #         + CrossEntropyLoss(neg_conf_preds, neg_conf_targets)
16: #####
17: conf_loss = self.cross_entropy_loss(
18:     conf_preds.view(-1, self.num_classes),
19:     conf_targets.view(-1)
20: )
21:
22: neg = self.hard_negative_mining(conf_loss, pos)
23:
24: pos_mask = pos.unsqueeze(2).expand_as(conf_preds)
25: neg_mask = neg.unsqueeze(2).expand_as(conf_preds)
26: mask = (pos_mask+neg_mask).gt(0)
27:
28: pos_and_neg = (pos+neg).gt(0)
29: preds = conf_preds[mask].view(-1, self.num_classes)
30: targets = conf_targets[pos_and_neg]
31: conf_loss = F.cross_entropy(
32:     preds,
33:     targets,
34:     size_average=False
35: )
36:
37: loc_loss /= num_matched_boxes
38: conf_loss /= num_matched_boxes
39: print(' %f %f' % (loc_loss.data[0], conf_loss.data[0]), end=' ')
40: return loc_loss + conf_loss
```

---

#### 4.4 模型训练/测试

#### 4.5 瓦片检测

## 第五章 实验结果与分析

- 5.1 改进 ssd 算法的实验结果
- 5.2 改进 ssd 对瓦片损害检测的准确率和召回率的实验
- 5.3 ssd 与修改了 focus loss 的算法进行 mAP、速度的比较

## 第六章 总结及展望

目前业界出现的目标检测算法有以下几种：

- 1、传统的目标检测算法：Cascade + Haar / SVM + HOG / DPM ；
  - 2、候选窗 + 深度学习分类：通过提取候选区域，并对相应区域进行以深度学习方法为主的分类的方案，如：RCNN / SPP-net/ Fast-RCNN / Faster-RCNN / R-FCN 系列方法；
  - 3、基于深度学习的回归方法：YOLO / SSD / DenseBox 等方法；
  - 4、结合 RNN 算法的 RRC detection；结合 DPM 的 Deformable CNN 等方法；
- 基于深度学习方法的几个可能的方向：
- 1、从原始图像、低层的 feature map 层，以及高层的语义层获取更多的信息，从而得到对目标 bounding box 的更准确的估计；
  - 2、对 bounding box 的估计可以结合图片的一些由粗到细（coarse-to-fine）的分割信息；
  - 3、对 bounding box 的估计需要引入更多的局部的 content 的信息；
  - 4、目标检测数据集的标注难度非常大，如何把其他如 classification 领域学习到的知识用于检测当中，甚至是将 classification 的数据和检测数据集做 co-training（如 YOLO9000）的方式，可以从数据层面获得更多的信息；
  - 5、更好的启发式的学习方式，人在识别物体的时候，第一次可能只是知道这是一个单独的物体，也知道 bounding box，但是不知道类别；当人类通过其他渠道学习到类别的时候，下一次就能够识别了；目标检测也是如此，我们不可能标注所有的物体的类别，但是如何将这种快速学习的机制引入，也是一个问题；
  - 6、RRC，deformable cnn 中卷积和其他的很好的图片的操作、机器学习的思想的结合未来也有很大的空间；
  - 7、语意信息和分割的结合，可能能够为目标检测提供更多的有用的信息；
  - 8、场景信息也会为目标检测提供更多信息；比如天空不会出现汽车等等。

## 第七章 致谢

感谢国家感谢党，感谢学校，感谢老师

## 参考文献

- [1] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.
- [2] Ross Girshick. Fast r-cnn. arXiv preprint arXiv:1504.08083, 2015.
- [3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In european conference on computer vision, pages 346–361. Springer, 2014.
- [5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.
- [6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.
- [7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.
- [8] Paul Viola and Michael Jones. 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, volume 1, pages I–I. IEEE, 2001.
- [9] Paul Viola and Michael J Jones. Robust real-time face detection. International journal of computer vision, 57(2):137–154, 2004.