中国科学技术大学

学士学位论文

基于深度学习与视觉注意的遥感图像目标检测

作者姓名： 揭一新

学科专业： 计算机科学与技术

导师姓名： 万寿红教授

完成时间： 二〇二〇年五月二十四日

University of Science and Technology of China

A dissertation for bachelor's degree

Remote Sensing Image Object Detection Based on Deep Learning and Visual Attention

Author: YiXin Jie

Speciality: Computer Science and Technology

Supervisor: Prof. ShouHong Wan

Finished time: May 24, 2020

致　谢

本文是我的本科毕业论文，是在我的导师万寿红副教授的悉心指导下完成的。感谢万老师经常挤出时间与我们讨论毕业设计的完成情况以及解答我们在学习生活中所遇到的问题，在此由衷地感谢万老师在我做毕业设计的过程中所给予的教诲和支持。

感谢在毕业设计讨论会中与我们共同探讨的各位师兄师姐，他们在讨论会中解决了很多我在学习过程中遇到的问题。感谢谢佳学长的帮助，在完成毕业设计的过程中遇到和解决的诸多问题，都得益于谢学长的指导与帮助、提出论文的修改意见等。

感谢班主任的认真负责，在我的学习和生活上都提供了巨大的帮助。

感谢中国科学技术大学计算机科学与技术学院的所有老师，他们爱岗敬业的治学精神使我受益匪浅。我将把我在本科四年学习到的知识和学习方法延续到今后的学习与工作中。

感谢参与论文评阅和答辩的老师们的辛苦工作，感谢老师们在百忙之中审阅这篇论文，感谢老师们对我的指导。

最后再次感谢在大学生活期间遇见的所有老师和同学，感谢各位长久以来的陪伴和关心！

目 录

中文内容摘要

随着深度学习与计算机视觉的飞速发展，目标检测成为近年来的一个热门研究方向。目标检测算法分为传统检测方法与基于深度学习的检测方法，而基于深度学习的目标检测算法分为一阶段和两阶段的方法。随着遥感技术的飞速发展，遥感图像的数量与质量得到了极大的提升。目前，遥感已广泛应用于农业、林业、军事侦察及环境监测等领域，遥感正以其强大的生命力展现出广阔的发展及应用前景。ASSD 是一阶段使用候选框的目标检测器，它是基于 SSD 进行改进的，使用特征融合机制来丰富语义，使得小目标的识别更加准确，并引入了注意力机制来对提取到的特征图进行处理，对图片上感兴趣的区域进行着重分析。本实验研究 ASSD 的各层结构，并测试其在遥感图像数据集上的检测效果。

关键词：目标检测；遥感图像；一阶段目标检测器；注意力机制；特征融合机制

Abstract

With the rapid development of Deep Learning and Computer Vision, Object Detection has become a hot research direction. Object Detection algorithm counts traditional detection methods and methods based on Deep Learning, while Object Detection based on Deep Learning is divided into one-stage and two-stage methods. The quantity and quality of Remote Sensing Image have been greatly improved with the help of the development of Remote Sensing Technology. At present, Remote Sensing Technology has been widely used in agriculture, forestry, military reconnaissance, environmental monitoring and other fields. With its strong vitality, Remote Sensing Technology is showing broad development and application prospects. ASSD is the One-Stage Object Detector based on Anchor.It is improved from SSD,using Feature Fusion to enrich the semantics which makes it have a greater detection accuracy on small scale objects.It also use Attention Mechanism to process Feature Map in order to focus on analysing the concerned region. This experiment study the structure of ASSD, and observe its detection effect on Remote Sensing Image.

Key Words: Object Detection; Remote Sensing Image; One-Stage Detector; Attention Mechanism; Feature Fusion

第一章　绪论

第一节　研究背景和意义

一、背景研究

Hinton 课题组在 2012 年参加 ImageNet 图像识别比赛时使用 AlexNet[1] 卷积神经网络获得冠军，使得神经网络备受关注。深度学习使用复杂的计算模型对抽象的数据表示进行学习，其能够从大量输入数据中分析并归纳出整体结构并加以分类概括。如今，在计算机视觉等领域的诸多分类问题上深度学习有着广泛的应用。针对目标运动的分析在计算机视觉领域中大致分为三个方面: 运动分割，目标检测；目标跟踪；动作识别，行为描述。

其中，目标检测不仅是计算机视觉领域需要处理的重要内容，而且它在视频追踪技术中也有着非常重要的地位。由于图像或视频中的观测目标很可能会出现诸如被背景环境遮掩、部分细节或部分未出现在观测画面上、拍摄时的光线位置等因素导致会影响检测效果。故目标检测在各种技术相对成熟的现在仍是诸多课题中很具有挑战性的方向，在未来会有可预见的提升空间。

遥感影像是指记录各种地物电磁波大小的胶片或照片，它主要分为航空像片和卫星相片。

随着遥感技术的飞速发展，近几年遥感图像的数量与质量得到了极大的提升。遥感图像能够描述诸如飞机、轮船、建筑设施等的地球表面的各种物体，如图1.1所示。

与自然场景下的普通图像相比，遥感图像有以下特点：角度是从空中拍摄；图像涵盖范围广，小目标占比很大，分布密集；旋转不变性；周围环境信息更重要; 训练数据少，但单张图片像素可能极大；图片可能包含其他各种信息，如波段，地理坐标；对机器的要求更高。遥感卫星影像获取信息快，受限制条件少，更新周期短，具有用途广、效益高、动态监测等特点。遥感图像拍摄时通常为瞬时成像，能及时获取所测目标物的最新资料，不仅便于及时更新信息，而且有利于对动态变化的资料进行分析、处理和研究。

目前，遥感在农业、林业、水文、气象、军事侦察等领域有着广泛的应用，并且应用领域也在不断扩大。遥感技术有着它强大的生命力，未来有着广阔的发展、应用前景。

由于目标检测在图像处理领域起着重要作用，广泛的应用于精准农业、智能检测和地理信息系统更新等领域，在对分析和理解遥感图像以进行智能观测的强烈需求下，对于遥感图像的目标检测问题变得越来越重要。

图 1.1　遥感图像

二、研究意义

目标检测，是利用计算机深度学习算法找出图像中是否存在感兴趣目标 (形状、位置)，同时对目标进行分类，判决目标具体是什么类别，即检测目标和识别目标。

目标检测为计算机视觉的一个热门方向，其在视频监控、机器人导航、视频监控、航空航天等许多领域有着广泛的应用。由于目标检测通过使用计算机视觉的相关算法在计算机上运行，减少了人力成本的消耗，具有着重要的现实意义。因此，近几年来目标检测逐渐成为理论与应用的研究热点。它不仅是图像处理与计算机视觉学科的重要分支，也是智能监控技术的核心组成部分；同时目标检测也是识别领域的一个基础性算法，对最近很火的人脸识别算法、视频追踪、实体

分割等技术起着非常重要的作用。

由于目前遥感图像数据集有着图像数目较少、图像中研究目标的信息缺乏、图像多样性匮乏等不足之处，虽然目前已经有着成熟的高效率高准确性的基于深度学习的目标检测算法，但是对于遥感图像数据集中的识别检测还是非常困难，有着很大的发展空间。

第二节　国内外研究现状

自深度学习的概念被提出以来，其吸引了非常多的个人与企业对这个领域进行深入研究。2015 年在《自然》杂志上名为《Deep Learning》[2] 的文章发表之后，深度学习在全球范围内掀起了浪潮。

Google、Amazon、Microsoft、Apple 等国际大公司在深度学习方面的研究都很深入，做出的很多相关产品在日常生活中有着广泛的应用，比如 Microsoft 的交互机器人“Cortana”，Google 的质能推荐服务，Apple 的语音助手“siri”等。我国对于深度学习的研究同样也有长足的发展，很多如百度、腾讯、阿里等科技公司逐渐开始全方位多层次地涉及相关内容，并有着与国际企业并驾齐驱的势头，将该技术应用在智能设备、医疗机械、推荐系统等方面。

一、传统的目标检测算法的发展

目标检测、物体分类的研究在计算机视觉学科里有着非常重要的作用。对人来说，将看到的音频、图片上面的物体的进行分类相当简单。而计算机缺乏泛化分类的处理能力，不能理解具体目标的抽象概念；在收到输入像素信息后，并不能直接处理得到输入文件中包含哪些物体。很多时候输入的内容也会出现噪声、物体不连续、目标尺寸过小等问题，导致计算机处理分析难上加难。

传统的目标检测算法包括产生目标建议框、对每个建议框中的特征进行提取和根据特征进行分类回归三个阶段。

（1）产生目标建议框

对于输入的图片，计算机只能知道具体像素点的信息，无法直接感知整体的特征。需要在原图上进行穷举建议框来预测出包含物体的建议框，即用不同大小尺度的滑动窗口依次在整张图片上进行扫描。显然，这种方法不仅运算量很大，而且大部分都在计算都是相重复的，导致了整体执行的效率极低。

（2）对每个建议框中的特征进行提取

比如传统的目标检测算法中的 HOG 算法[3] 使用直方图统计来对物体的边缘进行编码，这样有着较强的识别表达能力，不过该算法需要人工指定特征设计，可靠性并不高。

（3）根据特征进行分类回归

机器学习领域里面有着较多优秀的分类器，但是其在速度与精度方面并不能达到物体识别的要求。并且随着深度学习算法在计算机视觉学科上面优异表现的冲击，传统识别算法逐渐退出舞台。

二、基于深度学习的目标检测技术发展

从 AlexNet 网络在 ImageNet 数据集上表现出了前所未有的极好效果开始，深度学习方法便逐渐应用到了目标检测领域上。之后的几年中，各种结构的模型陆续出现，不断刷新对于如 VOC、COCO 等公认的标准数据集的检测准确率，性能方面也在逐渐提升。

如今，高性能与质量的目标检测算法都是基于深度学习的。最早首次使用深度学习模型来提取图像特征的 R-CNN（Region-based CNN）[4]，它的准确率达到了 49.6%，使得目标检测算法实现了划时代的进步。但是早期基于深度学习的目标检测，在生成目标建议框时也都是用滑动窗口的方式，本质上与穷举法无异，即仍未解决重复计算的问题。

为了解决冗余计算这个问题，Fast R-CNN[5] 被提出来了。为了合并 R-CNN 的训练和测试过程，Fast R-CNN 在卷积层后全连接层前加了一个简化的 SPP 层。虽然重复计算的问题被解决了，但是其速度仍满足不了实时检测的需求。

而 Faster R-CNN[6] 中的 RPN(Region Proposal Networks) 网络的表现要好了很多。Faster R-CNN[6] 则直接使用 RPN 网络来生成目标候选框。RPN 网络对输入的原始图像进行处理，输出一批包含目标置信度与坐标位置信息的矩形区域。从 R-CNN 到 Fast R-CNN 再到 Faster R-CNN 是一个不断改进优化的过程，它使用一个深度学习的网络模型将传统目标检测的三个步骤整合起来。

随着以 YOLO[7] 和 SSD[8] 方法为代表的一阶段目标检测算法的提出与优化，目标检测领域又到达了一个新的高度，它们真正的做到了实时效果。一阶段目标检测算法在速度上有着显著的优势，并且随着模型的不断发展，一阶段的目标检测器的精度越来越高，能够达到检测需求。

第三节　研究内容

实验研究一阶段的基于 anchor 的神经网络和注意力机制发展在遥感图像上的应用。ASSD[9] 是近年来结合上述两个研究点先进性网络，我们将研究与现存的一些基于深度学习的遥感图像目标检测方法相比，ASSD 是否可以在速度和精度上取得一个更好的平衡。

第四节　组织结构

全文分为五章，整体组织结构如下：

第一章为绪论部分，主要介绍了现在目标检测算法研究的内容和意义所在，同时介绍了国内外的研究现状。

第二章相关技术介绍，主要介绍了本实验中利用到的各项主要概念、技术与原理；包括卷积神经网络、Resnet 网络、目标检测器的分类及特点、注意力机制以及遥感图像的介绍。

第三章为本实验的网络模型 ASSD 的结构分析，介绍了其网络结构、候选框机制、特征融合机制与注意力单元。

第四章是本实验的实验过程与结果分析，主要介绍了本文实验的环境，实验室用的数据集以及对实验结果的评估等内容。

第五章是论文的总结与展望，总结了本文的主要工作并展望该方向今后的研究工作。

第二章　相关技术介绍

第一节　卷积神经网络

深度学习中一种常见的网络架构就是卷积神经网络，其是在生物自然视觉认知机制的启发下发展而来。1959 年，Hubel Wiesel 发现了视觉系统的信息处理，可视皮层是分级的。比如人眼在观察一个物体时，从瞳孔摄入信息开始，先做初步处理：大脑皮层的相应功能部分对物体的边缘和方向进行识别，然后再进行抽象：大脑判定，眼前的物体的形状等信息，最后进一步抽象处理，大脑进一步分析判定该物体的类别。

现代的 CNN 结构是由 20 世纪 90 年代，LeCun 等人发表论文确立的，后来学者又对其进行完善。他们当时设计了一种叫做 LeNet-5 的可以对手写数字做分类的多层的神经网络，见[10]。

CNN 能够从输入的原始像素中，识别空间局部的特征。然而当时既缺乏大量的训练数据又缺乏算力强的计算机，这导致了 LeNet-5 对于复杂问题的处理效果很不理想。为了解决难以训练深度卷积神经网络的困难，很多专家学者先后提出了很多方法。其中，最著名的是 Krizhevsky et al. 提出的 AlexNet[1]，其在图像识别任务上取得了重大突破，同时也让卷积神经网络的研究进入了一个新的高度。在那之后，研究人员又提出了一系列新的改进方法，其中最著名的要数 ZFNet[11], VGGNet[12], GoogleNet[13] 和 ResNet[14] 这四种。从结构看，CNN 发展的一个方向就是层数变得更多，ILSVRC 2015 冠军 ResNet[14] 是 AlexNet[1] 的 20 多倍，是 VGGNet[12] 的 8 倍多。虽然通过增加深度，同时使用非线性激活函数 ReLu 与 Dropout 的方法，利用增加的非线性网络能够得出与目标函数的近似结构。但是，这样做也使得网络整体变得更复杂，并且很难去优化。

卷积神经网络一般由两层组成：特征提取层与特征映射层。

在特征提取层中，为了能够提取局部特征，每个神经元的输入与前一层的输出相连。而且它与其它特征间的关系在局部特征被提取后也会被确定下来；

在特征映射层中，若干个特征映射组成网络的计算层，其中每个特征映射是一个平面，并且上面所有神经元的权值均相等。特征映射结构采用非线性的 sigmoid 激活函数。特征映射层上的神经元的权值共享机制，也减少了网络中参数的个数。这种每个卷积层都跟着一个计算层的提取结构减小了特征分辨率。

一、卷积神经网络的基本结构

卷积层、池化层、全连接层、激活层与回归层构成卷积神经网络的主体结构。它的输入通常先经过卷积层与激活层，然后再通过池化层输出作为其中的一步结果，这种结构堆叠若干次后 (每堆叠一次即增加一层深度)，再接上全连接层、回归层运算得出最终的输出结果。这样便构成了卷积神经网络的完整网络结构，上述步骤的简化模型如图2.1所示。

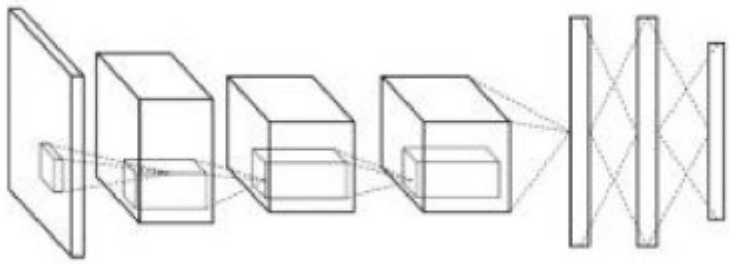


图 2.1　卷积神经网络模型

卷积神经网络的输入内容通常为图像 (由像素矩阵构成的位图) 或其他多维的数据，输出为如分类标签或者目标位置的检测目标值。接下来对各层的作用进行具体说明。

卷积层（(Convolutions layer)：用来学习特征 (对于接受输入的数据)。卷积层由很多的卷积核 (convolutional kernel) 组成，不同的卷积核用来计算不同尺度的特征图 (feature map)。卷积层的本质就是许多不同尺度的滤波器。每个滤波器由三维的矩阵构成。其中前两维是滤波器的尺寸 (即卷积核的大小)，第三维为其深度，它是由输入的特征图的深度决定其值，其中初始输入图像的深度为颜色通道数 (位图为 RGB-3 个通道)。并且在使用同一个卷积核的情况下，特征图可以共享该卷积核的权值，这样就能减少很多的网络参数，使运算更为简化。

激活层 (activation layer)：由于卷积操作过程是线性的，而实际图像的特征通常都是非线性的。因此，想要提高网络结构的非线性的表达能力就需要在卷积神经网络中加入激活函数。引入了非线性后，卷积神经网络便能够匹配期望的目标函数。深度学习中通常使用的激活函数有 sigmoid、tanh、ReLU 等。

池化层 (Pooling layer)：用来降低卷积层输出的特征向量的维度。这样不仅可以减少神经网络参数的数量，而且还能够改善运算结果，使网络结构能过正常工作，同时还可以让网络不易出现过拟合现象。池化层的本质是二维滤波器，常见的池化层的操作有平均池化和最大化池化等。

网络可以通过堆叠卷积层与池化层从而获得图像的更多抽象特征。

全连接层 (Full connected layer)：卷积层和池化层堆叠之后，就能够形成一层或多层全连接层，能够实现整合能力，起到分类器的作用。全连接层能够从前面卷积池化层学到的特征图映射到与之对应的样本空间。全连接层的每一个节点都和上一层的所有节点相连，获得一维向量的输出。全连接层常用 sigmoid/tanh 非线性函数。

二、卷积神经网络的训练过程

卷积神经网络的训练过程主要包括前向传播和误差反向传播，它实际上就是让网络各层的神经元节点更新它们的权值，以得到较好的模型。

1. 前向传播

前向传播是指卷积神经网络将提取到的特征在模型中的以正向的次序进行传递。特征依次按照顺序经过卷积神经网络的各层结构。这个过程中包含了局部感知、权重共享和下采样这三项主要内容。

（1）局部感知

使用小于输入特征图尺寸的卷积核对其进行扫描，然后在每个位置都进行卷积运算是局部感知的主要操作。由于在图像中，局部联系通常比全局联系更密切，并且符合人感知视觉从局部到全局的特点，因此只感受局部信息而不是全局的信息就可以得出有效的特征。故在这种方式下，层中的每个节点只需要与图像上某个局部的像素点相连，这样做也减少了训练时的权值参数，提高了计算性能。

（2）权重共享

权重共享指的是对同一张图像均使用同一个卷积核进行单次扫描。由于在同一张图像中，每一个局部的统计特性具有共性，故共享卷积核的操作能够减少权值参数，便于计算。在每个卷积核遍历接受输入的特征图时，每个位置处均是共享一组权重的，因此不需要去单独学习图像中各个位置的权重。

（3）下采样

它本质上就是池化过程。由于经过池化过程能够缩小图像的规模 (卷积神经网络中，池化操作并不会损失重要的局部信息)，因此可以减少计算量，并且减少了训练参数，不易出现过拟合现象。

以上三种方法每步都能减少计算步骤，从而组合来使用不但可以极大地减少各层节点之间的连接数量，而且还能够有效的提升网络的计算处理性能。

2. 反向传播

反向传播是指卷积神经网络输出的特征和标准特征 (*Ground Truth*) 之间存在误差，我们可以利用这些误差信息根据一些特定的计算方法在模型中按照反向的次序调整各层神经元的权值，从而实现对网络整体的各参数进行调整。常见的损失函数有交叉熵、均方差、对数似然等。

3. 整体过程

如图2.2所示，一幅图像通过卷积神经网络的过程就是：

（1）前向传播过程

首先对其进行卷积的操作，输出特征图。得到特征图后，对特征图进行池化操作，然后使用激活函数处理。反复重复这个过程多次之后，再通过全连接层与目标结果建立映射关系，得出训练结果。

（2）反向传播过程

在反向传播过程中，网络需要对训练结果与真实值 (*Ground Truth*) 使用某种损失函数进行误差计算，然后根据运算结果与网络实际结构综合来更新各层神经元的权值。

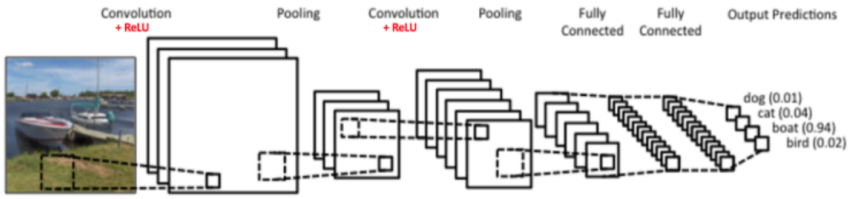


图 2.2　卷积神经网络训练过程

三、深度网络的退化问题

网络的深度对模型的总体性能有着重要作用。理论上说，网络层数增加后，网络可以对更加复杂的特征进行提取，所以应该可以取得更好的结果。但是更深的网络结构的性能不一定会更好。实验结果发现增加网络深度时会出现退化问题：当网络层数增加时，网络训练结果的准确度会出现饱和，深度过大时甚至出现准确度下降的现象。这个现象可以在图2.3中直观看出来：56 层的网络比 20 层得网络效果还要差。因为 56 层网络的训练误差同样高，所以这也不会是过拟合得问题。由于深层网络很可能会出现梯度消失或者梯度爆炸的问题，这就使得深度学习模型得训练过程会很艰难。虽然现在已经存在一些如 BatchNorm 的技术手段来缓解这个问题，但是效果仍然不是很理想，仍需要一个更加优秀的网络模

型来解决这一问题。

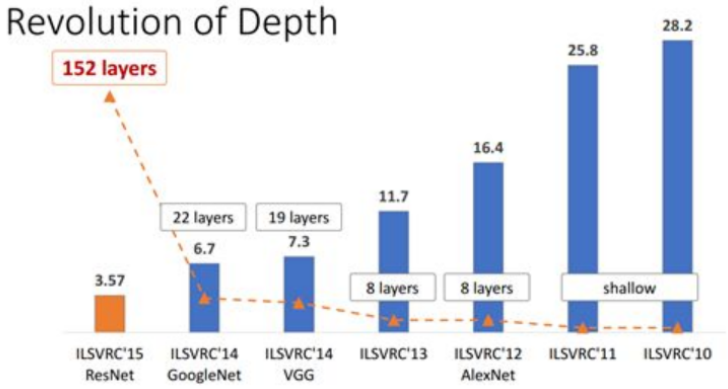


图 2.3　神经网络的退化问题

四、残差学习

最近几年深度学习发展迅速，出现了很多优秀的 CNN 网络结构，例如 AlexNet、VGG、GoogLeNet、ResNet 等。

VGG[12] 是 Oxford 的 Visual Geometry Group 的组提出的。该网络说明了网络深度的增加在一定程度上会影响网络的表现性能。VGG[12] 有两种结构，分别是 VGG16 和 VGG19，二者只是在网络深度上不同，并没有本质上的区别。VGG16 相对于 AlexNet[1] 的其中一个改进是将 AlexNet 中的较大卷积核 (11×11，7×7，5×5) 用连续的几个 3x3 的卷积核来代替。由于多层非线性层可以增加网络深度来保证学习更复杂的模式，因此采用堆积的小卷积核在给定的感受野的情况下优于采用大的卷积核，提升了网络的深度，在一定程度上提升了神经网络的效果。并且其参数较 AlexNet 更少，计算开销更小。但是 VGG 仍然耗费了很多的计算资源，仍使用了较多的参数，导致内存占用较大，整体效果不是特别理想。

ResNet[14] 是微软团队开发的网络。它的特征在于具有比以前的网络更深的结构。虽然加深层对于提升性能很重要。但是，可能会出现上面提到的退化问题，导致最终性能不佳。为了解决这样的问题，ResNet 提出了一种“快捷结构”。使用了这种快捷结构后，随着层数的加深网络在一定限度内就可以而不断提升性能了。

快捷结构，将输入的 x 加到输出中，跳过了输入数据的卷积层。如图2.4所示，在连续的卷积层中，输入的 x 一方面进入卷积池化层进行运算，另一方面存储下来直接加到前面运算的结果上。即通过里的快捷结构，使原来的输出由 F(x) 变为了 F(x)+x。由于采用这种结构后，反向传播时的信号可以无衰减地进行传

递，防止了由于深度增加而引起的梯度消失或爆炸问题。因此，即便是加深层数，Renet 网络也能高效地学习。

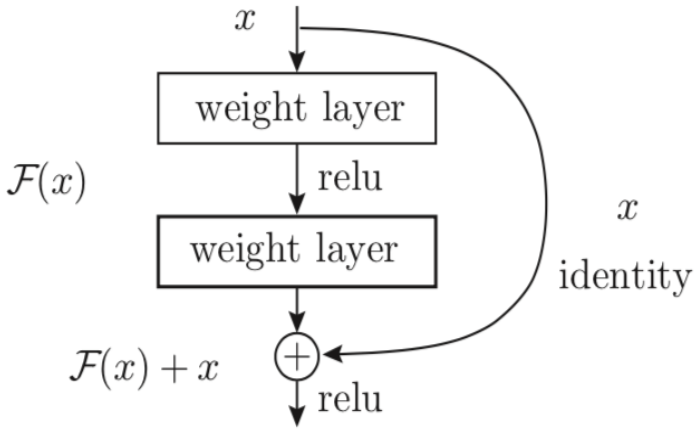


图 2.4　Resnet 中快捷结构示意图

ResNet 以 VGG 网络为基础，引入快捷结构以加深层并且能够避免退化问题，其结构如图2.5所示。

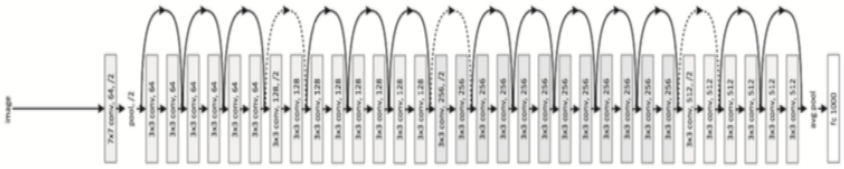


图 2.5　Resnet 模型结构示意图

ResNet 使用跳跃式的连接方法来加深层 (以两个卷积层为间隔)。实验结果显示，即使层数以及达到了 150 以上，检测结果的准确性也会持续提高。在 ILSVRC 大赛中，ResNet 的 Top5 错误识别率为 3.5%，如此优秀的表现真是令人称奇。

Resnet 在精度、速度上都比 VGG 表现好很多，因此，现阶段大部分模型都选择将 Resnet 作为主干网络进行训练。

第二节　目标检测器

通常，目标检测器可以分为两类：一阶段目标检测器和两阶段目标检测器。两阶段目标检测器，如 Mask R-CNN[15]、Fast R-CNN[5]、Faster R-CNN[6] 是提取感兴趣的区域进行检测和识别，由于其包含一个用于区域建议的预处理步骤，使得整体流程是两阶段式的。由于处理的步骤多以及阶段耦合的原因，两阶段的目标检测器效率很低，不能够满足实时任务的要求。一阶段目标检测器，如

YOLO[7]、DES[16] 和 SSD[8] 等直接训练产生出物体的置信度和坐标位置，将所有计算 (划分候选框与识别) 封装在一个网络中，从而在根本上很大程度的提高了检测速度。一阶段检测器分为基于候选框的和无候选框的。候选框是在神经网络初始化时预先生成的不同尺寸和比例的物体候选框。一阶段基于候选框的检测器直接预测候选框的偏移和类别。经典网络有基于 SSD 的检测器，如 DSSD[17]、FSSD[18]。一阶段基于无候选框的直接预测物体的坐标和类别，比如 Cornernet[19] 预测了几个关键点，然后将这些点分成对，即对象左上角和右下角的坐标。但是由于其缺乏候选框机制而导致预测不稳定，这就需要使用更大的输入图像输入和更复杂的模型来训练。同时，其还需要额外的操作优化算法，否则网络会产生大量背景或错误的关键点，从而降低检测精度。

一、两阶段目标检测器

两阶段目标检测器因其有着一个用于区域建议的预处理步骤，因此也被称为基于区域 (Region-based) 的方法，R-CNN 系列模型是这一类型的代表。

R-CNN[4] 将检测划分为两阶段，第一阶段是基于图片训练提出一系列可能包含物体的区域 (Region Proposal)，原论文中使用的是 Selective Search 算法；第二阶段是在提出的这些建议区域上运行分类网络 (使用当时效果最好的 AlexNet[1])，回归得到每个区域内物体的类别。

文章 Fast R-CNN[5] 指出 R-CNN 耗时的原因之一是其卷积神经网络是在每一个提议区域上单独进行的，并没有共享计算，导致消耗了大量的计算资源。Fast R-CNN 在基础网络把图片运行完成后，才传入 R-CNN 子网络。这样共享了大部分计算，因此其执行效率可以高一些。

Faster R-CNN[5] 在两阶段的目标检测器中有着重要的地位，其将原来的 Selective Search 算法替换为 RPN 网络，通过这样的处理可以使得检测任务以很快的速度完成。它是在 Fast R-CNN 基础上加了 RPN 网络，去除冗余运算、共享卷积计算的特性使得 RPN 网络的整体计算量非常小，故 Faster R-CNN 能够以较高的速度运行，并且在精度方面也能够达到最佳。Faster R-CNN 的突出贡献是提出 Regional Proposal Networks，替代之前的算法。

RPN 网络将区域提议这一任务转化为二分类问题，即其中是否包含物体。它第一步是在一个滑动窗口上生成不同大小和宽高比的候选框，取定 IoU 的阈值，按 Ground Truth 标定这些候选框的正负。即传入 RPN 网络的样本数据被整理为候选框的 (坐标) 和每个候选框是否有物体 (0,1 二分类)。RPN 网络对每个样输出

一个置信度和四个坐标值信息。置信度表示这个候选框内部包含有有物体中心点的概率，四个坐标值 (中心点 x，y 与宽 w 高 h) 用于定位物体的位置。RPN 网络训练的损失函数为是否包含物体的二分类值与坐标差值的损失的加权统一。

Faster R-CNN 的突出之处在于其用了 RPN 网络来代替传统的区域提议的方法。使用不同大小和宽高比的候选框的机制也在后面提出的一些模型中被采用 (YOLO[7] 等)。Faster R-CNN 确定了"RPN+RCNN" 的模式，奠定了两阶段目标检测的基本结构。

二、一阶段目标检测器

一阶段目标检测器运行和检测速度非常快，虽然在准确性上比两阶段目标检测器要差一些，但是随着近几年对于相应算法的优化与发展，两种检测器在识别准确性上的差别正在逐渐缩小；因此一阶段目标检测器凭借着其速度优势获得了目标检测中的重要地位。由于一阶段模型直接从图片获得预测的结果，没有中间的区域提议过程，因此也被称为 Region-Free 的方法。目前主要的一阶段目标检测算法分为 YOLO[7] 系列与 SSD[8] 系列。

1. YOLO

如图2.6所示，首先，YOLO[7] 将一张输入图片分成 $S \times S$ 格。如果 Ground Truth 框的中心落在某格上，则该格负责预测这个目标。每格会预测 n 个候选框以及 n 个置信度。该置信度得分表征出的是该格内包含物体的概率有多大以及这个检测出来的预测框与实际的框之间 IOU 的大小。因此，如果这个格子上面没有物体，则该位置处的得分为 0；如果该格上包含有物体的中心，那么其得分应该等于 IOU。所以对于每个候选框而言，都会预测 5 个值，分别是置信度与中心点的横纵坐标，候选框的宽高。

对于划分的每个格子而言，除了预测 n 个候选框的信息以及其置信度的大小外，它同时还会预测 c 个类别的条件概率。VOC2007、2012 数据集中包含了 20 个类，所以它就会预测 n=21 个类别 (多的一个类是背景) 的概率。其中这个概率是在确定候选框中存在实际目标的情况下，该目标为第 i 类的条件概率。每个候选框会共享这样的概率，与 n 个候选框无关。不难看出，每格中物体的具体类别的置信度不但反映出这个类别出现在这里出现的概率，而且同时也反映出了该预测框与实际框 (*Ground Truth*) 之间的相似度有多大，即检测结果是否符合标准。

如图2.7所示，如果单纯从网络结构考虑，YOLO 和普通的用卷积神经网络

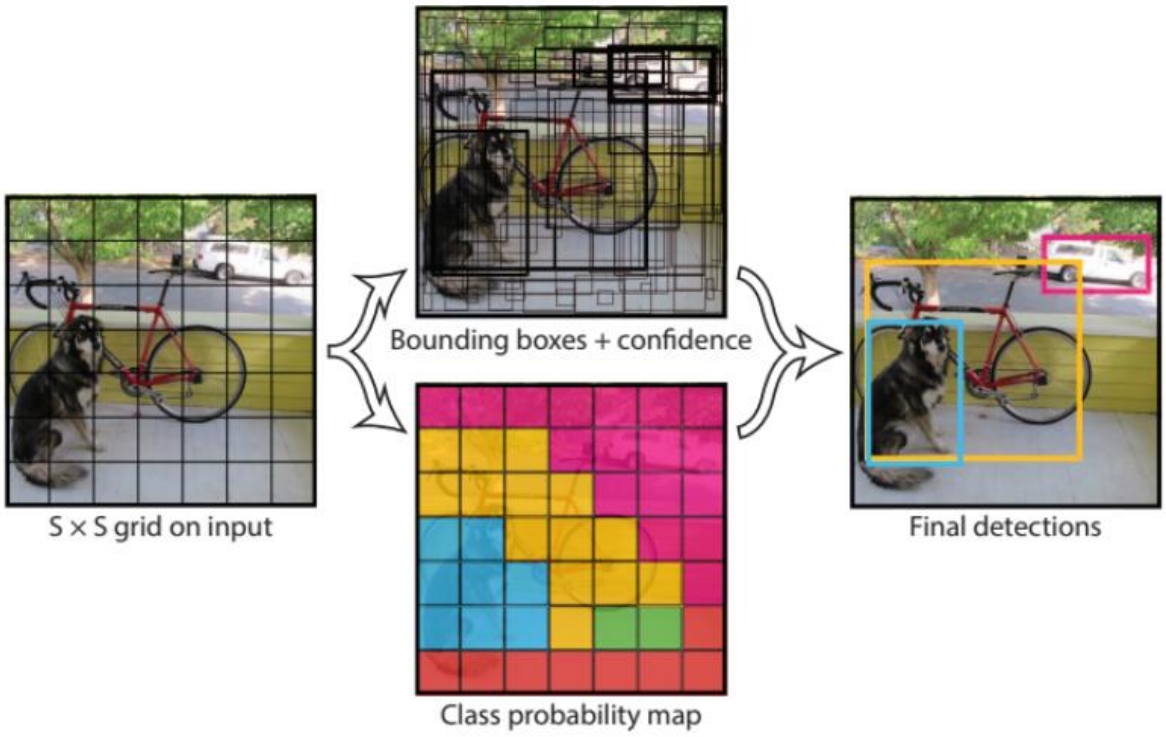


图 2.6　YOLO 预测

进行分类的网络基本没有什么区别，因为去掉候选框处理的这个步骤后，YOLO 的结构非常简单，就是单纯的卷积、池化最后加了两层全连接层。它们最大的差异是 YOLO 最后输出的地方采用的是线性函数作为激活函数，因为它不仅仅是预测某类的概率，还需要预测 Bounding Box 的位置。所以整体上来说，YOLO 的整个结构就是通过神经网络的变换把输入的图片映射成一个输出的向量。

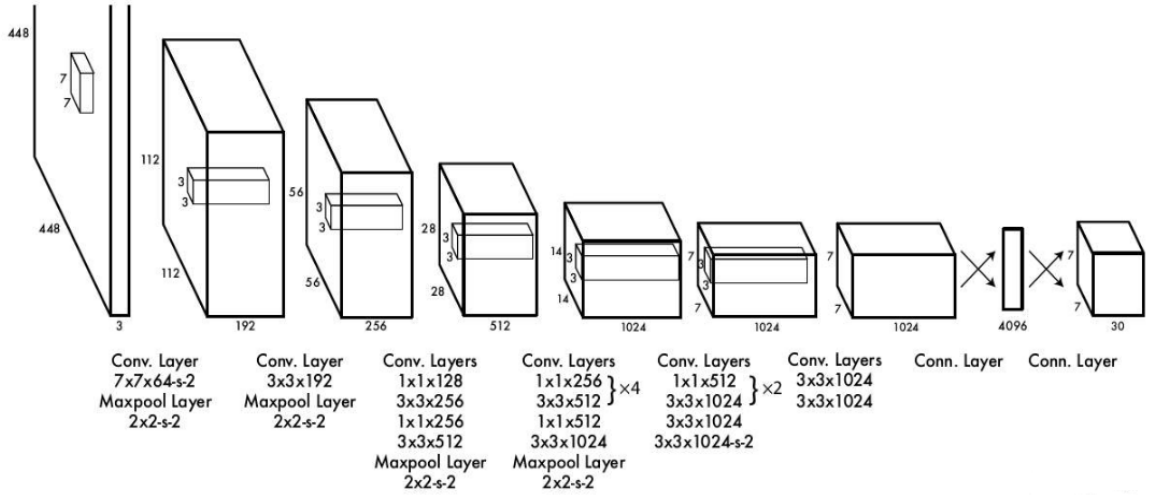


图 2.7　YOLO 网络结构

2. SSD

SSD[8] 的网络结构如图2.8所示，其使用了 (38 × 38),(19 × 19),(10 × 10),(5 × 5),(3 × 3),(1 × 1)6 个不同尺度的特征图来预测分类。这 6 个特征层会分别经过 3 × 3 的卷积后转化为包含置信度和坐标偏差的通道。由于 SSD 使用了更多的特征层，因此与 YOLO[7] 相比，它更适合检测拥有多种不同尺度目标的图像。

同时，在这 6 种不同尺寸的金字塔结构输出的特征图中，SSD 还使用了不同大小、不同宽高比的候选框。每层特征图使用不同于其他层的大小的候选框，并且根据具体的特征图尺度设置 4 种或者 6 种的宽高比来框出检测物体。之后的很多基于 SSD 的改进拓展网络都采取这种候选框方式。

与之前的网络相比，SSD 并没有使用所有的负样本，而是对这些匹配上背景的样本处理得到的由置信度损失按照降序排列，把其中损失较大的样本记作难例 (hard negative)，当作该模型需要重点学习的对象。把其中损失结果最大的前 N 个样本当作负样本 (其中正样本与负样本的比例控制需要在 1:3 左右)。而对于那些没有被选上的样本，SSD 把其标签设置为-1，即它以后不再参与训练。

论文说明数据增强处理能够明显提升网络的性能。即采用数据增强的步骤，能够彰显样本之间的多样性与差异性，从而极大地提升了模型的泛化能力。SSD 的高准确率同时也得益于空洞卷积操作。通过空洞卷积，网络能在较少的参数情况下有着较大的感受野，即能使网络能感知到更多的东西。

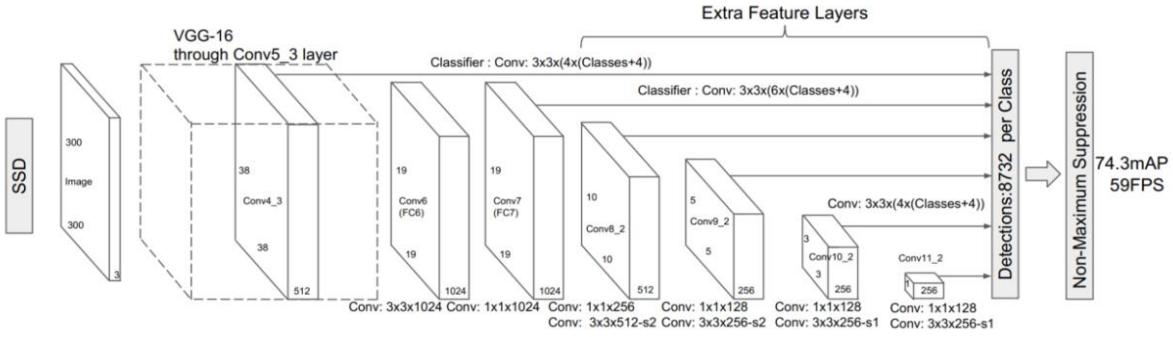


图 2.8　SSD 网络结构

SSD 奠定了一阶段目标检测的重要模式，其能够在比两阶段目标检测器速度快了近一个量级的情况下维持了较高的准确性。后续大部分的一阶段目标检测器基本是基于对 SSD 的改进展开研究的。

第三节　注意力机制

注意力模型最初被用于自然语言处理，现在也成为深度学习领域的一个重要模型。注意力机制人工智能领域已成为神经网络结构的重要组成部分，并在自然语言处理、语音识别和计算机视觉等领域有着广泛的应用。

注意力机制仿照了人观察物体的方式。一般而言，当人在看一张图片的时候，除了从整体把握一幅图片的特点之外，通常会对图片的某个局部位置感兴趣，进而注意力集中在相关区域上。例如人像照中人脸的位置，宣传图片中商品的具体位置等。在翻译一段话的时候，人们通常是从句子的整体内容入手，在阅读句子的过程中，除了需要关注其中出现的词语本身的含义，还要考虑与词语有前后关系的信息以及该词语的上下文信息。在自然语言处理学科中，如果需要对情感进行分类，那么在其中的某个句子中，一定会涉及到一些能够传达出情感的词语。处在该句子里面的其他不相关词语，则是情感词语的上下文。它们并不是对语言表达没有用，而是其对于语义表达的能力远没有那些能够直接表达情感的关键词大。因此，注意力机制是把注意力集中放在重要的信息上，而忽略其他次要的因素。其中重要程度的判断取决于应用场景。

注意力机制的处理结果通常都是以概率张量或者概率图的形式表示。从原理上来说，主要分为空间注意力模型，通道注意力模型，空间和通道混合注意力模型三种。

空间注意力模型：图像中所有的区域对于检测任务的贡献不相同，与任务相关的区域是检测的重点，是需要突出关心的。空间注意力模型就是寻找图像、网络中最重要的部分进行处理。由于我们在大部分情况下所感兴趣的内容只是输入图像中的一个小区域，因此空间注意力的本质就是完成对目标的定位。比如 Google DeepMind 提出的 STN 网络 (Spatial Transformer Network)[20]。它通过学习输入的局部重要信息，分析完成适合任务的预处理操作，这用到的是一种基于空间的注意力模型。

通道注意力机制：卷积神经网络的输入是一个二维图像，其中的一个维度是图像的尺度，即宽 w 高 h，另一个维度就是通道数。因此基于通道的注意力机制也是很常用的。通道注意力机制的本质，在于构建了各个特征之间的重要性关系，对不同的任务可以根据输入的区别进行特征分配，处理起来简单有效。例如 SENet(Sequeeze and Excitation Net)[21] 是 2017 届 ImageNet 分类比赛的冠军网络，其本质上是一个基于通道的注意力模型，它通过联系各个通道的相关程度，并对

不同的任务增强或者抑制不同的通道。

空间和通道注意力机制的融合：CBAM(Convolutional Block Attention Module)[22] 是空间和通道注意力机制融合的代表性网络，其结构如图2.9所示，可以看出一个 CBAM 模块包括两个子模块，其分别进行通道上和空间上的注意力计算，见图2.10。

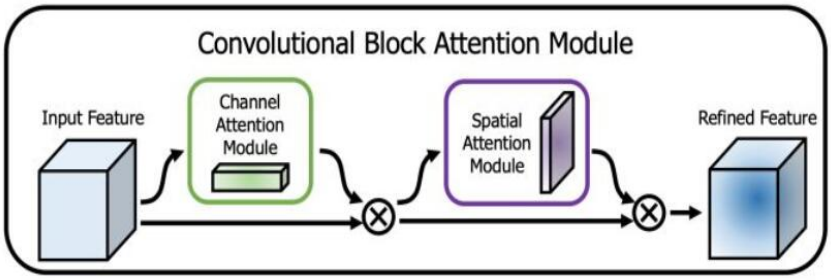


图 2.9　CBAM 结构

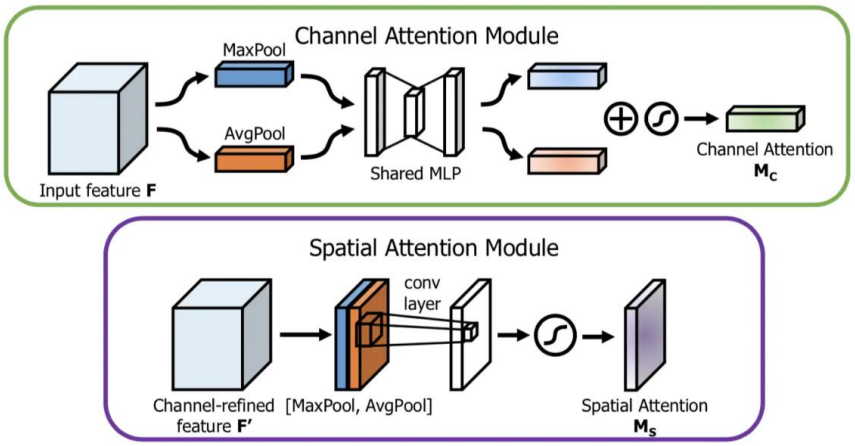


图 2.10　CBAM 空间、通道注意力模型

实验证明，在小模型上添加 CBAM 模块也能带来较稳定的性能提升，而且这样处理只会增加一点少量的计算，并不会在执行效率上产生显著影响。这个实验结论还说明，模型越大模型的表达能力会更强，性能上也会高一些。虽然小模型在性能上要低一些，但是添加 CBAM 后同样能看到结果的稳定性能提升。因此这个方法处理效果十分可观。

相比于基准的模型，在添加了 CBAM 之后，在 GRAD-CAM[23] 的可视化下，该模型能够实现更加关注目标本身。其执行效果如图2.11所示。

对于输入图像，注意力图经过运算突出显示了不同尺度大小的有用区域，如图2.12为 PASCAL VOC2007 数据集中注意力的可视化热力图。注意力图是图片

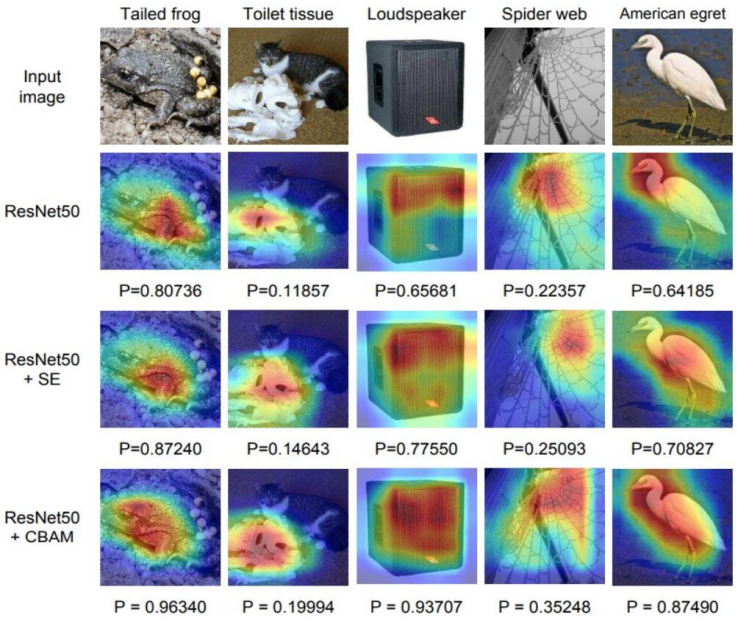


图 2.11　添加 CBAM 的注意力热力图

上的每个位置空间特征的加权之和，因此，与特征相关的区域会被突显，不相关区域 (如背景) 会被抑制。这样做有助于模型关注实际目标，从而提高了检测效率与准确性。

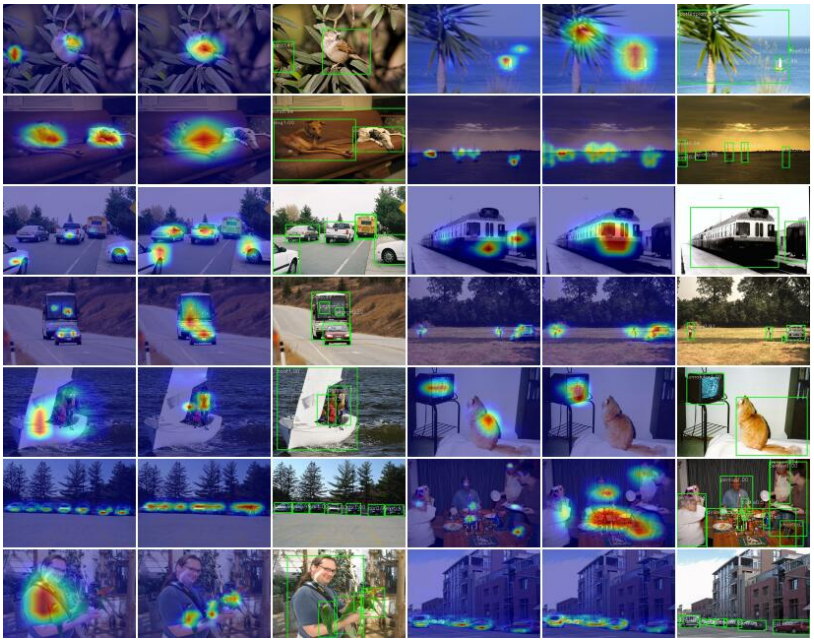


图 2.12　VOC2007 数据集注意力热力图

第三章　ASSD 目标检测器

第一节　ASSD 简介

ASSD[9] 是一阶段基于候选框的注意力目标检测网络，其是以 SSD[8] 为原型并进行改进的。SSD 虽然能够对多尺度特征映射进行检测，还能够有效地处理各种类别大小，但是，其小尺度金字塔结构层缺乏语义信息，因此小目标识别的效果不好。解决这一问题的一种方法是建立更多的卷积层，对涉及到的小目标特征进行进一步的细化，或者将语义较多的从大尺度层传入到小尺度层。考虑到速度是一阶段目标探测器的主要优势，故需要用以较小的额外计算成本提高 SSD 的精度。为了达到这样的目的，ASSD 方法构建了一个小型注意单元网络，将 SSD 输出的特征图放入运算，以提高检测结果的准确性。ASSD 将注意力单元放在特征映射和预测模块之间。ASSD 选择 SSD 作为一级探测器，它在简单性、速度和准确性之间提供了最佳的权衡，它保留了 SSD 的原始结构，并且在小尺度层引入了融合机制来加入语义信息。这种设计保留了原始 SSD 的优点，同时更有效地学习对象特性。在 PASCAL VOC 和 COCO 等数据集上，ASSD 在精度和效率方面的表现的比之前所有的一阶段目标检测器更好。鉴于 ASSD 强大的性能，我们决定在实验中选取 ASSD 作为检测器。

第二节　网络整体结构

ASSD 的整体结构如图3.1所示：

ASSD 总共进行九层卷积操作。其中第、四、五次的卷积通过语义融合输出特征图，四到九的卷积依次输出特征图，然后将这七份特征图都放入注意力单元进行运算，最后进行回归分类，得出预测结果。

第三节　候选框

如图3.2所示，首先将输入图片进行填充至宽高相等，然后 resize 至 513×513 分辨率作为输入，之后不同的金字塔模块对图像分割处理的方式不同。

对于小目标的检测集中于宽高划分小的层中 (如 38×38，19×19)，大目标的检测位于宽高划分大的金字塔模块中 (如 1×1，3×3)。

如图3.3所示，在固定划分的情况下，对于特征图上的一个具体正方形会生

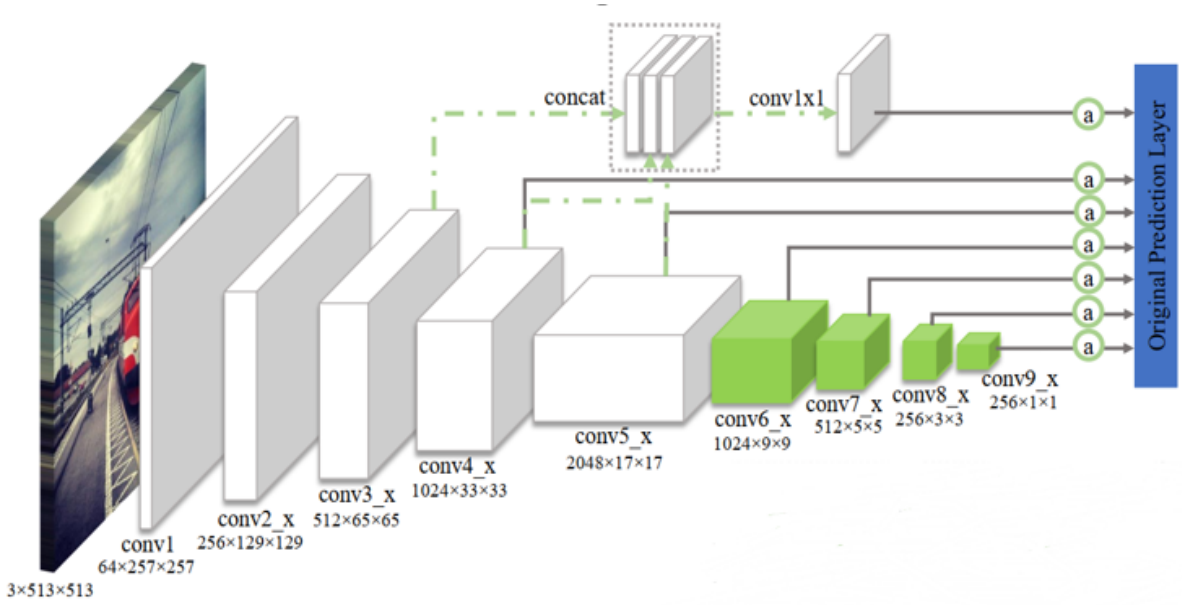


图 3.1　ASSD 整体结构示意图

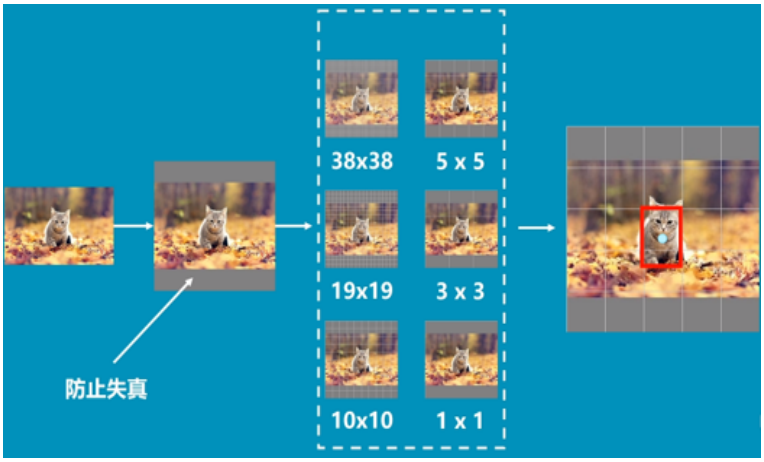


图 3.2　不同层对图片的划分

成若干种不同尺寸、宽高比的候选框，网络会对所有的候选框进行检测，然后进行分类回归得到置信度，之后与 Ground Truth 比较得出 loss 反向传播更新权值。

如图3.4所示，以第一层为例：该层将输入图片划分成 38×38 的方格，每一个方格都会生成 4 个 Anchor，于是该层总共有 $38 \times 38 \times 4 = 5776$ 个 Anchor

ASSD 采用与 SSD 相同的 Anchor Box 生成方法。

具体来说，使用长宽比 $a_r = 1, 2, 1/_2$ 作为特征映射 conv3,8,9 上的 Anchor Box，使用 $a_r = 1, 2, 1/_2, 3, 1/_3$ 作为特征映射 conv4-7 上的 Anchor Box。每个 Box 都有一个最小尺度 s_{min} 和一个最大尺度 s_{max}。Anchor Box 的标准化宽度和高度

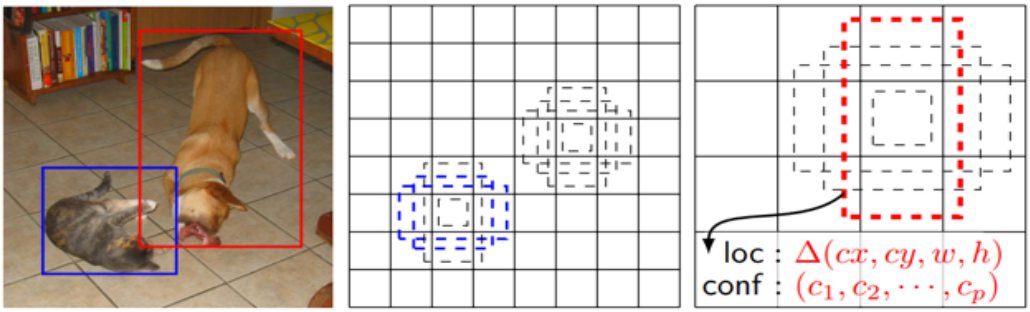


图 3.3　SSD 中的 Anchor 机制

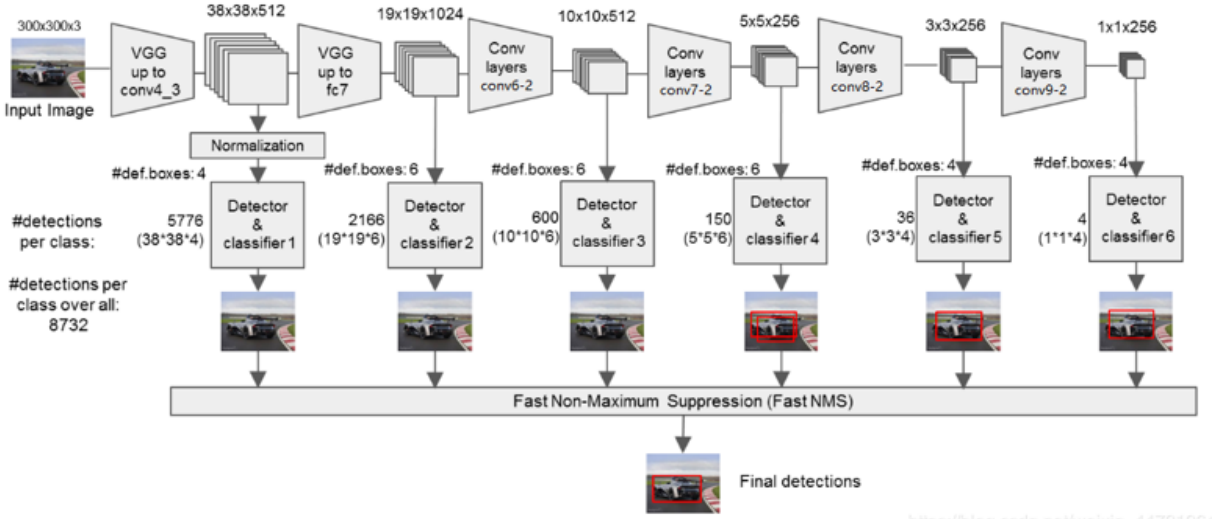


图 3.4　每层的 Anchor

按 $w = s\sqrt{(a_r)}$ 和 $h = s/\sqrt{a_r}$ 计算，其中

$$s = \begin{cases} \sqrt{s_{min} \times s_{max}}, a_r = 1 \\ s_{min}, otherwise \end{cases} \tag{3.1}$$

其使用了 7 个不同尺寸的特征图来进行预测，分别是 (65×65)，(33×33)，(17×17)，(9×9)，(5×5)，(3×3)，(1×1)。第一、六、七层对每个位置使用 4 种候选框进行预测，其余层使用 6 种。

故对于七层金字塔结构的 ASSD，对于每张图片其总共检测 $(65 \times 65) \times 4 + (33 \times 33) \times 6 + (17 \times 17) \times 6 + (9 \times 9) \times 6 + (5 \times 5) \times 6 + (3 \times 3) \times 4 + (1 \times 1) \times 4 = 25844$ 个候选框。

第四节　融合机制

ASSD 受 FSSD 的启发，它将 layer4 和 layer5 的提取到的特征图 (即上下文信息) 融合到 layer3 中，以增添其语义。实验中，单独的融合操作并不能显著提高检测精度。相反，从实验结果来看，它甚至降低了精度一点与更多的计算成本。这主要是因为这三层结构具有着不同的感受野和不同的提取特征的能力；此外，级联和 1×1 的卷子操作可能会抵消三层之间的相对重要性，并抑制了原始 layer3 层中的关键特征。实验表明，在融合后放置注意力单元时，效果会有明显的改善。深层语义的引入帮助了注意单元，让它可以发现存在于原始 layer3 层中的有用信息。如果只使用注意单元时，其效果与融合和注意机制模型相比性能较差。如图3.5所示，这说明特征融合机制和注意力单元是相辅相成的，也就是说要同时使用这两个机制对模型进行训练。

Method	Backbone	Time (s)	mAP
SSD513	ResNet101	0.1417	79.75
SSD513+fusion	ResNet101	0.1466	79.57
SSD513+att	ResNet101	0.1593	82.13
SSD513+fusion+att	ResNet101	0.1648	**82.95**

图 3.5　特征融合与注意力机制的表现

语义融合的过程可以表示为：

$$x^3 = W^3 Concat x^3, x^4, x^5 + b^3 \tag{3.2}$$

其中 $x^s \in R^{c^s \times N^s}$ 是 s 层的特征图,$W^3 \in R^{C^3 \times C'}$,$b^3 \in R^{C^3}$。

Concat 操作通过双线性插值对第四层和第五层进行上采样，以使其大小与第三层的大小对齐，从而能够输出三层间的综合特征，如图3.6所示。

第五节　注意力单元

注意力单元如图3.7所示，其具体计算方式如下：

假设 $x^s \in R^{C^s \times N^s}$ 是 s 层的特征图，进行如下操作：

$$q(x^s) = W_q^{sT} x^s \in R^{C' \times N^S} \tag{3.3}$$

$$k(x^s) = W_k^{sT} x^s \in R^{C' \times N^S} \tag{3.4}$$

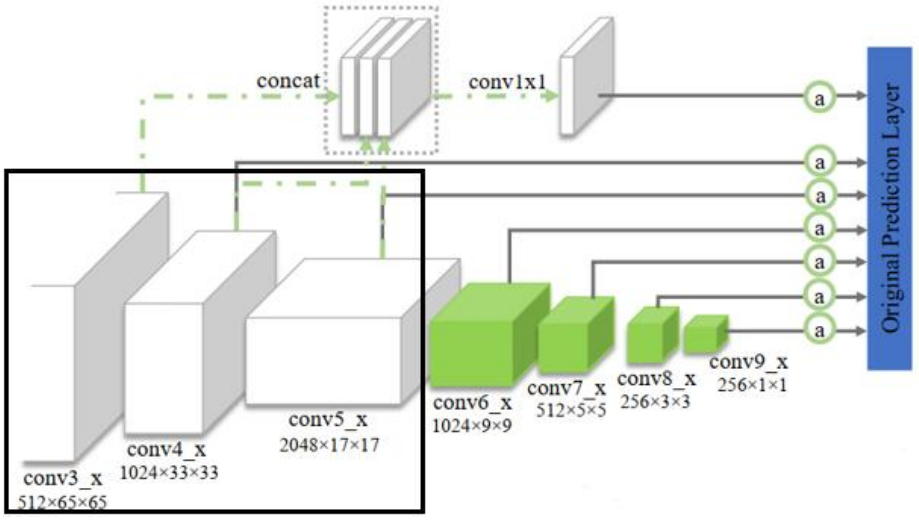


图 3.6　融合

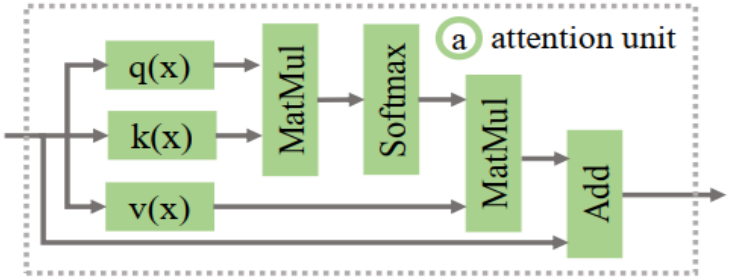


图 3.7　注意力单元示意图

$$v(x^s) = W_v^{sT} x^s \in R^{C^s \times N^S} \tag{3.5}$$

其中

$$W_q^s, W_k^s \in R^{C^s \times C'}, W_v^s \in R^{C^s \times C^s} \tag{3.6}$$

设 a^s 为注意力得分矩阵，其产生如下：

$$a^s = q(x^s)^T k(x^s) \in R^{N^s \times N^s} \tag{3.7}$$

将 a^s 每一行通过 $softmax$ 操作规范化：

$$\bar{a}_{ij}^s = \frac{exp(a_{ij}^s)}{\Sigma_j^{N^S} exp(a_{ij}^s)}, i, j = 1, 2, \cdots, N^S \tag{3.8}$$

其中 $\bar{a}_i^s$ 表示查询特征图的第 i 个位置时的像素关系，称之为“注意力图”。将输入特性 x^s 转换为 q 和 k 的原因是为了减小矩阵运算的开销、降低计算成本。$q(x^s)$ 和 $k(x^s)$ 的矩阵计算特征相似度，并创建一个 $N \times N$ 注意力图，以便显示

特征之间的关系。接下来，我们应用 $v(x^s)$ 和注意图 a^s 之间的乘法操作。使用这种运算，我们通过计算一个更新后的的特征图，作为每个位置上的单个特征的加权和。最后，我们将之前的矩阵乘法结果添加回输入特征映射 x^s 中：

$$x^{s'} = x^s + (\bar{a}^s v(x^s)^T)^T \tag{3.9}$$

注意力图 $\bar{a}^s$ 将特征图在所有位置的大范围关系联系起来，因此学习了特征图的全局上下文。它突出了特征图的相关部分，并以细化的信息指导检测。

第六节　损失函数

训练过程中的 $loss$ 为置信度的损失函数 $loss\_conf$ 与预测边框的损失函数 $loss\_locs$ 之和：

$$loss = loss_l + loss_c \tag{3.10}$$

其中，$loss\_conf$ 为 Pytorch 中计算交叉熵的 $F.cross\_entropy$，计算方式如下：

$$loss(x, class) = -log(\frac{exp(x[class])}{\Sigma_j exp(x[j])}) \tag{3.11}$$

$loss\_locs$ 使用 $SmoothL1$ 作为损失函数，计算方法如下：

$$SmoothL1(x) = \begin{cases} 0.5x^2, & if\,|x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \tag{3.12}$$

其中，$s = f(x_i) - y_i$ 为真实值和预测值的差值。

第四章　实验过程与分析

第一节　实验过程

一、实验环境

本实验使用 Pycharm 完成。

操作系统：Win10。

硬件：处理器为 i7-7700(主频 4.2GHz)，显卡为 NVIDIA TITAN X（Pascal 架构显存 12G），内存容量为 32G（速度 2400MHz）。

软件：代码运行在 Pycharm 上，使用 Pytorch 框架。

二、数据集

本实验数据集为在 1-3 米分辨率下的遥感图像，总共 4615 张图片。每张图片如图4.1所示：

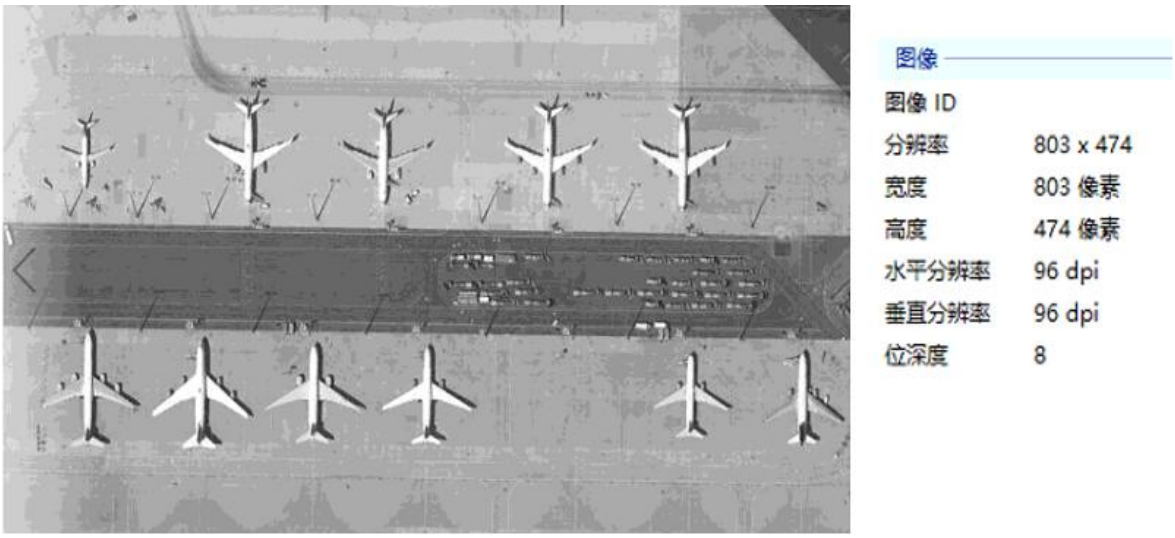


图 4.1　遥感图像数据集样例

每张图片的含有关于飞机、油库、船具体坐标的标注 (作为 $GroundTruth$), 如图4.2所示：

三、评价指标

本实验的评价指标包括每个类的召回率 $Reacall$、平均准确率 AP 以及三个类的 $MeanAP$。

```
<object>
    <name>ship</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
  - <bndbox>
        <xmin>196</xmin>
        <ymin>193</ymin>
        <xmax>322</xmax>
        <ymax>242</ymax>
    </bndbox>
</object>
```

图 4.2　人工标注示例

1. 混淆矩阵

我们通常使用混淆矩阵来评估一个二分类问题。混淆矩阵会把样本实例分为正类 (*positive*) 与负类 (*negative*)。在训练测试开始前我们已经有了 *GroundTruth* 的信息，然后根据使用训练完成的模型来对测试集进行处理，输出每张图片上的检测结果。对于图像目标检测问题来说，它对正、负类的评价标准是这个图片上该位置处物体的类别信息是否与 *GroundTruth* 相同，并且在类别检测正确后判断 IOU 是否达到阈值，如果大于等于阈值就认为该检测结果是正类。因此，对测试结果评估这个二分类问题而言，检测结果会出现以下四种情况：

真正类 ($True\ Positive, TP$)，测试输出的结果为正类的正样本。

假正类 ($False\ Positive, FP$)，测试输出的结果为正类的负样本。

假负类 ($False\ Negative, FN$)，测试输出的结果为负类的正样本。

真负类 ($True\ Negative, TN$)，测试输出的结果为负类的负样本。

如图4.3所示。

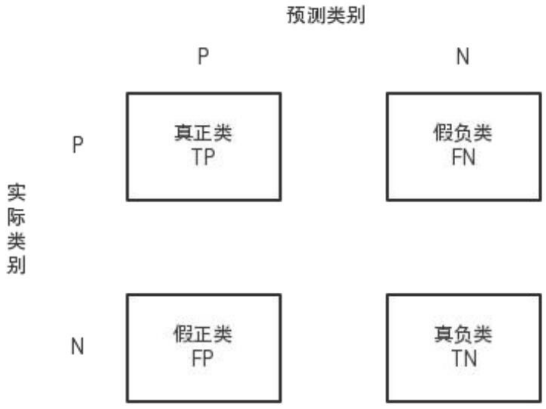


图 4.3　混淆矩阵

2. 准确率与召回率

准确率 (*Precision*) 是所有测试输出为正的样本 (正真类与假正类) 在总样本数中的占比，用 P 表示，其计算方式如下：

$$P = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

召回率 (*Recall*) 是真正类在所有正类 (正真类与假负类) 中的占比，在公式中用 R 表示，计算方式如下：

$$R = \frac{TP}{TP + FN} \tag{4.2}$$

准确率与召回率的定义可以由图4.4直观感受。

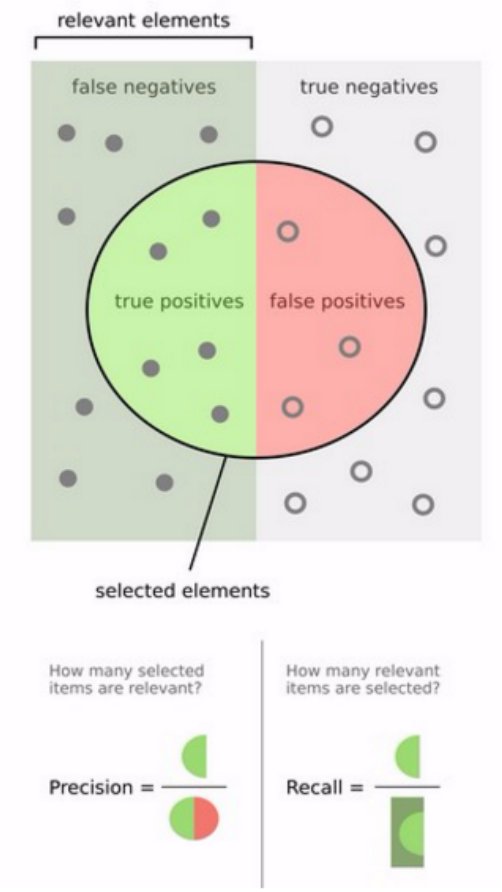


图 4.4　准确率、召回率的定义

3. AP 与 MAP

上面讲述了 *Precision* 和 *Recall* 的含义，而我们需要一个同时考虑准确率与召回率的指标。

AP(*Average Precision*)：平均准确率，在不同 *recall* 下的最高 *precision* 的

准确率 (一般会对各类别分别计算各自的 AP)。

$mAP(mean\ AP)$：平均准确率的均值，各类别的 AP 的均值。

试想，如果在一个 $BoundingBox$ 里，检测器识别出来某类 T 的 $score$ 最高，可是结果也只有 0.1，那么它很可能还是负样本。所以我们需要一个阈值，如果识别出了 T 类而且分数大于这个阈值才说明它是正样本，否则它是负样本。阈值影响 $Precision$ 和 $Recall$ 的方式如下所述。在 T 类的例子中，如果阈值太高，$Prediction$ 非常严格，所以我们认为是 T 类的基本都是 T 类，$Precision$ 就高了；但也因为筛选太严格，我们也放过了一些 $score$ 比较低的 T 类，所以 $Recall$ 就低了；如果阈值太低，什么都会被当成 T 类，$Precision$ 就会很低，而 $Recall$ 会很高。这样我们就明确了阈值确实对 T 类的 $Precision$ 和 $Recall$ 产生影响和变化的趋势，也就说明，$Precision$ 不是一个绝对的东西，而是相对阈值而改变的东西，$Recall$ 同理。那么单个用 $Precision$ 来作为标准判断，就不合适。需要综合考虑 $Precision$ 与 $Recall$ 之间的关系，用一组固定值表述不够全面，因为我们根据不同的阈值，可以取到不同（也可能相同）的 $Precision-Recall$ 值。这样想的话对于每个阈值，我们都有相应的 $(Precision,Recall)$ 对，也就有了 $Precision$ 和 $Recall$ 之间的曲线关系。这样一条“$P-R$ 曲线”，它衡量着两个有价值的判断标准，即 $Precision$ 和 $Recall$ 的关系，将这两个指标一起动态考虑，就有了 T 这个类的 $AveragePrecision$，即曲线下的面积，他可以充分的表示在这个模型中，$Precision$ 和 $Recall$ 的总体优劣。最后，我们计算每个类的 $AveragePrecision$，就得到了 $Mean\ Average\ Precision$。

故 AP 衡量的是学出来的模型在具体某个类别上检测结果的好坏，而 mAP 衡量的是学习的模型在所有类别上的结果。

第二节　实验分析

一、训练细节

本实验检测飞机、油库、船三类；总类数还要加上背景，故 $num\_class=4$

$labelmap=('airplane',' ship',' storage\_tank')$

$VOC\_CLASSES=('\_\_background\_\_',' airplane',' ship',' storage\_tank')$

训练集使用总图片数的 80%，3694 张；测试集使用总图片数的 20%，921 张。

二、检测结果

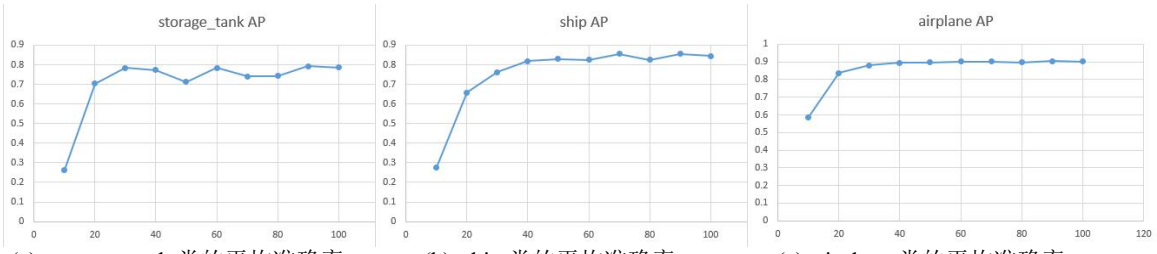
本实验总共训练 110 个 *Epoch*，每 10 个 *Epoch* 进行一次 *Loss*、*MAP*、每类的 *AP* 输出 (由于到第 100 个 *Epoch* 后出现 *Loss* 升高、*MAP* 降低的过拟合现象，故展现前 100 个 *Epoch* 的结果)，如表4.1、表4.2、图4.5所示。

表 4.1　训练 Epoch 与 MAP 的关系

Epoch	10	20	30	40	50	60	70	80	90	100
MAP	0.3759	0.7331	0.8085	0.8288	0.8127	0.8371	0.8325	0.8216	0.8446	0.8507

表 4.2　训练 Epoch 与 Loss 的关系

Epoch	10	20	30	40	50	60	70	80	90	100
Loss	0.4290	0.3726	0.3439	0.3116	0.2996	0.2859	0.2713	0.2632	0.2602	0.2586



(a) storage_tank 类的平均准确率　(b) ship 类的平均准确率　(c) airplane 类的平均准确率

图 4.5　各类 AP 值与 Epoch 的关系

训练完 100 个 *Epoch* 后各类的 *Recall*、*AP* 值如表4.3所示，每个类的 *Recall* 都很高，说明绝大部分的目标都被检测出来了。同时各类的 *AP* 也很高，说明在综合准确率召回率的情况下该网络检测的效果非常好，验证了 ASSD 算法的有效性。

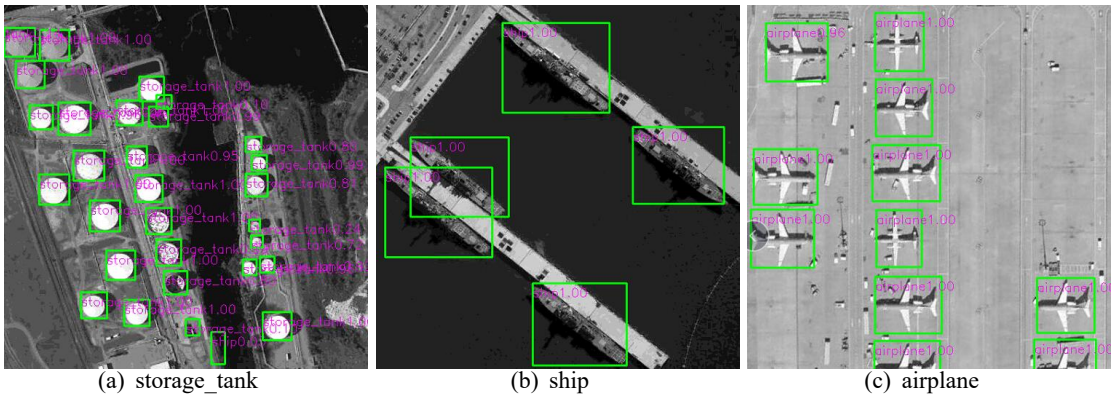
表 4.3　各类 Recall 与 AP 值

information	airplane	ship	stroage_tank
Recall	0.9640	0.9387	0.8741
AP	0.9029	0.8453	0.7857

从训练过程中的输出可以看出 *Loss* 在不断降低 *MAP* 在不断升高，故选择第 100 个 *Epoch* 的结果进行后面的可视化处理。

三、可视化结果

训练过程中记录下该模型中的各项权重，执行完 100 个 *Epoch* 后调用可视化程序对验证集中的图片进行查看，其效果如图4.6所示：



(a) storage_tank (b) ship (c) airplane

图 4.6　各类可视化检测结果

其中各类的识别与其轮廓的标注都很符合 *Ground Truth*，可见检测效果还是很不错的。即 ASSD 通过使用语义融合机制与注意力机制在遥感图像数据集上实现了较好的表现效果。

第五章　总结和展望

第一节　全文总结

本文研究一阶段基于候选框、特征融合机制以及注意力机制的目标检测器ASSD的整体结构与实现细节，考虑到原论文是基于VOC2007与2012数据集进行的验证，本文采用遥感图像数据集对ASSD的性能进行了测试，验证了其算法的有效性。

本文首先讨论了目标检测算法的背景和意义，同时介绍了国内外的研究现状。在相关技术介绍部分主要分析了了本实验中利用到的各项主要概念、技术与原理；包括卷积神经网络、Resnet网络、目标检测器的分类及特点、注意力机制以及遥感图像的介绍。之后介绍了其网络结构、候选框机制、特征融合机制与注意力单元，并分析了ASSD的整体结构与它的优势所在。然后在服务器上配置好实验环境并对遥感图像1-3米分辨率的数据集进行测试，根据目标检测算法常用的评估标准进行处理，最后得出实验结果并分析。

第二节　未来展望

考虑到遥感图像在军事领域上的重要作用，对于其上的目标检测需要有高效的性能。一阶段目标检测器的优势在于其分析速度快，虽然在检测结果的准确率方面比两阶段目标检测器效果差，但是满足了速度的需求。本实验中遥感图像数据集的图像个数较少，识别的物体类别数也不多，随着遥感技术的不断发展，未来遥感图像相关的数据集的数量与质量会更加丰富，这样对模型的训练会有很大的帮助。同时，将来可以在一些诸如金字塔结构的优化(调整各层的尺度)、主干网络的调整、候选框的选择(增加一些新的宽高比甚至增加倾斜的候选框)、学习率的调整等方面对ASSD进行改进，使之更加适应贴合遥感图像数据集。相信在深度学习与并行处理器飞速发展的当下，目标检测领域未来必然会有新的突破。

参 考 文 献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25(2).

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[3] WANG X, HAN T X, YAN S. An hog-lbp human detector with partial occlusion handling[C]//IEEE International Conference on Computer Vision. 2009.

[4] GIRSHICK R D T, Donahue J. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. CVPR. IEEE, 2014, 2014.

[5] GIRSHICK R. Fast r-cnn[C]//2015 IEEE International Conference on Computer Vision (ICCV). 2016.

[6] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[J]. 2015.

[8] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[J]. 2016.

[9] YI J, WU P, METAXAS D N. Assd: Attentive single shot multibox detector[J]. Computer Vision and Image Understanding, 2019.

[10] Y L, L B. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998.

[11] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [J]. 2013.

[12] KAREN SIMONYAN A Z. Very deep convolutional networks for large-scale image recognition[J]. 2014.

[13] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[J]. 2014.

[14] SZEGEDY C, IOFFE S, VANHOUCKE V. Inception-v4, inception-resnet and the impact of residual connections on learning[J]. 2016.

[15] GEORGIA H K G. Mask r-cnn[J]. 2018.

[16] ZHANG Z, QIAO S, XIE C, et al. Single-shot object detection with enriched semantics[J]. 2017.

[17] FU C Y, LIU W, RANGA A, et al. Dssd : Deconvolutional single shot detector [J]. 2017.

[18] LI Z, ZHOU F. Fssd: Feature fusion single shot multibox detector[J]. 2017.

[19] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2018.

[20] PARK D, CHUN S Y. Classification based grasp detection using spatial transformer network[J]. 2018.

[21] LI G, ZHANG C, LEI R, et al. Hyperspectral remote sensing image classification using three-dimensional-squeeze-and-excitation-densenet (3d-se-densenet) [J]. Remote Sensing Letters.

[22] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module [J]. 2018.

[23] SELVARAJU R R, DAS A, VEDANTAM R, et al. Grad-cam: Why did you say that?[J]. 2016.