计算机视觉基础

中国石油大学（华东）
青岛软件学院、计算机科学与技术学院
宫文娟

计算机视觉基础

- 第一章 概论
- 第二章 基础知识
- 第三章 图像分类
- 第四章 图像语义分割
- 第五章 目标检测
- 第六章 识别
- 第七章 目标跟踪
- 第八章 多目视觉
- 第九章 视觉问答

第五章 目标检测

- 5.1 概述
- 5.2基于经典手工特征的目标检测算法
- 5.3 深度学习时代目标检测
- 5.4 行人检测

5.1 概 述

目标检测是计算机视觉的经典问题之一，主要任务是标定图像中目标的位置，并给出目标的类别。

从传统的人工设计特征加浅层分类器的框架，到基于深度学习的端到端检测框架，目标检测技术包含了很多不同类型的方法。

5.1 概 述

《图像语义分割》包含低层的像素级别的处理方法也包含高层的语义级别的处理方法。

《目标检测》基本都是高层的语义级别的处理方法。

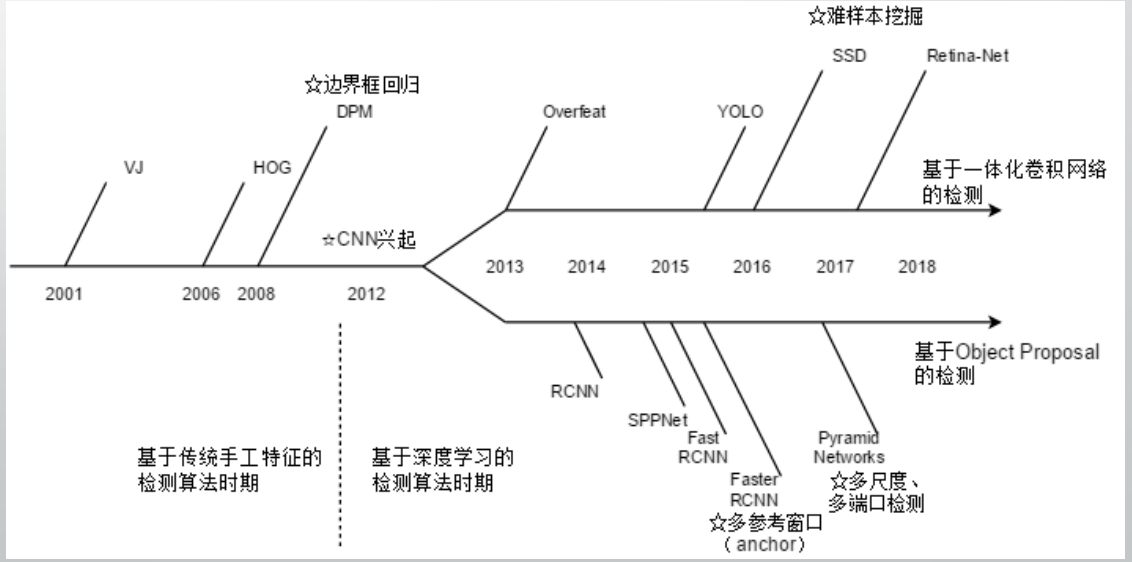
5.2 基于经典手工特征的目标检测算法

目标检测发展：

2013之前，自然图像的目标检测算法基本上都采用基于传统手工特征。

2013年至今，目标检测主要采用基于深度学习的方法。

从技术发展上来讲，目标检测的发展则分别经历了“边界框(Bounding Box)回归”、“深度神经网络兴起”、“多参考窗口（Multi-References，又称Anchors）”以及“难样本挖掘与聚焦”几个里程碑式的技术进步。



5.2 基于经典手工特征的目标检测算法

- 早期的目标检测算法大多是基于手工特征构建。由于在深度学习诞生之前缺乏有效的图像特征表达方法，人们不得不尽其所能设计更加多元化的检测算法，以弥补手工特征表达能力上的缺陷。
- 由于计算资源的缺乏，人们不得不同时寻找更加精巧的计算方法对模型进行加速。

多数基于经典手工特征的目标检测算法都遵循两阶段流程:

1. 使用滑动窗口寻找目标，
2. 使用模板匹配或者其改进方法对滑窗选中的区域进行判断。

研究人员的主要工作集中在三个方面:

- 检测窗口选择
- 特征设计
- 分类器设计

5.2 基于经典手工特征的目标检测算法

基于经典手工特征的目标检测算法的方法类别：

5.2.1 滑动窗口与模板匹配检测法

5.2.2 选择性搜索

5.2.3 Viola-Jones检测器

5.2.4 可变形部件模型

5.2 基于经典手工特征的目标检测算法

5.2.1 滑动窗口与模板匹配检测法

- 模板匹配是一种最原始、最基本的模式识别方法，用来研究某一特定对象的图案位于图像的位置，进而识别对象。
- 是图像处理中最基本、最常用的方法。

□ 模板匹配具有自身的局限性，它只能进行平行移动，若原图像中的匹配目标发生旋转或大小变化，该算法就会失效。

- 模板就是一副已知的小图像，即目标检测中的目标
- 模板匹配就是在一副大图像中搜寻目标，已知该图中有要找的目标，且该目标同模板有相同的尺寸、方向和图像元素
- 通过一定的算法可以在图中找到目标，确定其坐标位置
  - ➢ 通常使用滑动窗口算法

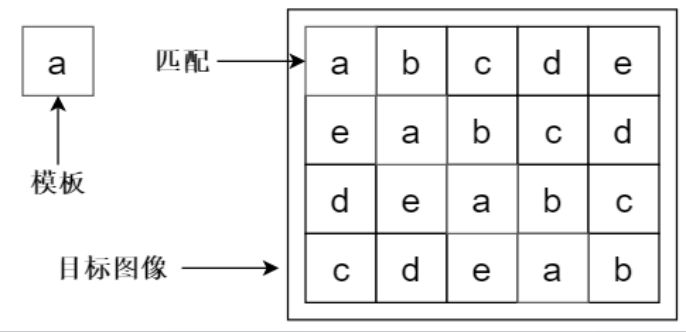


图5-2 模板匹配示意图

5.2.2 选择性搜索

5.2.2 选择性搜索

- 选择性搜索综合了蛮力搜索和分割的方法。
- 与传统的单一策略相比，选择性搜索提供了多种策略
- 与蛮力搜索相比，大幅度降低搜索空间，可以较好地与识别算法相结合。
- 选择性搜索用于目标的候选区域检测算法，计算速度快，具有很高的召回率，考虑颜色，纹理，大小和形状兼容计算相似区域的分层分组。
- 优势：
  - 多尺度
  - 多元化
  - 快速计算

主要包含两类方法：

1. 分级分组算法
2. 多元化策略

5.2.2 选择性搜索

5.2.2.1 分级分组算法

- 分级分组算法认为图像中区域特征比像素更具代表性
- 分级分组算法使用有效的基于图的图像分割[2]方法产生图像初始区域，使用贪心算法对区域进行迭代分组:

1. 使用有效的基于图的图像分割方法得到初始分割区域R={r1,r2,…,rn}。
2. 初始化相似度集合S=∅。
3. 计算两两相邻区域之间的相似度，将其添加到相似度集合S中。
4. 从集合S中找出相似度最大的两个区域ri和rj，将其合并成为一个区域rt，从集合中删去原先与ri和rj相邻区域之间计算的相似度，计算rt与其相邻区域（与ri或rj相邻的区域）的相似度，将其结果加入到相似度集合S中。同时将新区域rt添加到区域集合R中。
5. 获取每个区域的包围盒L，输出物体位置的可能结果L。

- 选择性搜索算法[1]在不同尺寸、颜色形成多层次的候选区域列表，如图5-3所示：

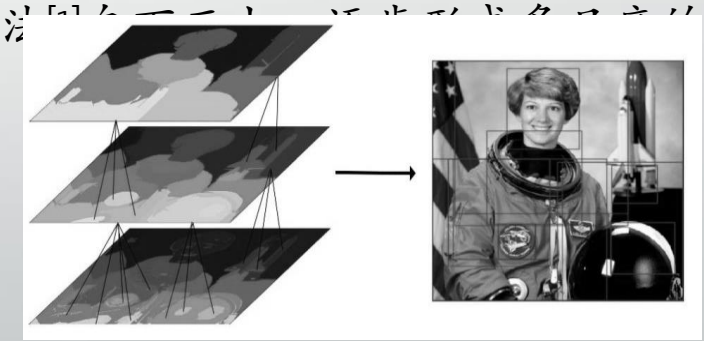


图5-3 选择性搜索[1]方法示意图

[1]. Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. Netherlands: Kluwer Academic Publishers, International journal of computer vision, 2013, 104(2): 154-171.
[2]. Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. Netherlands: Kluwer Academic Publishers, International journal of computer vision, 2004, 59(2): 167-181.

5.2.2 选择性搜索

5.2.2.2 多元化策略

- 采用了多样性的策略，使得抽样多样化，主要有三个不同方面：
  1. 色彩空间多样化：使用不同的色彩空间，以提取不同的不变属性；
  2. 互补的相似度度量：采用不同的相似性度量（例如颜色相似度、纹理相似度、尺寸相似度、填充相似度，最后计算最终相似度衡量）；
  3. 改变起始区域：采用不同算法选取起始区域，以达到最佳分隔效果。

➢ 有效的基于图的图像分割方法[2]采用这些策略取得了较好效果。

[2]. Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. Netherlands: Kluwer Academic Publishers, International journal of computer vision, 2004, 59(2): 167-181.

5.2.3 Viola-Jones检测器

Viola-Jones（VJ）检测器

- VJ检测器采用了最传统也是最保守的目标检测方法——滑动窗口检测。

□ 这类方法计算开销巨大

- VJ人脸检测之所以器能够在有限的计算资源下实现实时检测，其中有三个关键要素：

1. 多尺度Haar特征的快速计算，
2. 有效的特征选择算法
3. 高效的多阶段处理策略。

5.2.3 Viola-Jones检测器

Viola-Jones（VJ）检测器

- 方法流程：

1. 多尺度Haar特征快速计算

- Haar特征是指在窗口的某个位置取一个矩形的小块，然后将这个矩形小块划分为黑色和白色两部分，并分别对两部分所覆盖的像素点的灰度值求和，最后用白色部分像素点灰度值的和减去黑色部分像素点灰度值的和，得到一个Haar特征的值。

➢Haar特征反映了局部区域之间的相对明暗关系，能够为人脸和非人脸的区分提供有效的信息，例如眼睛区域比周围的皮肤区域要暗，通过Haar特征就可以将这一特点表示出来。

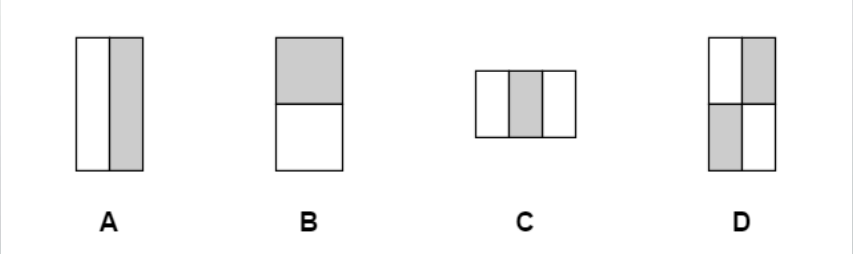


图5-4 Haar特征示意图

5.2.3 Viola–Jones检测器

Viola-Jones（VJ）检测器

- 方法流程：
  1. 多尺度Haar特征快速计算
  2. 特征选择
     - VJ人脸检测器采用了AdaBoost方法，希望降低计算代价，所以只用简单的分类器，同时也希望分类准确度高，于是把多个简单的分类器组合起来，聚弱为强，将多个弱分类器组合成一个强分类器。

5.2.3 Viola-Jones检测器

Viola-Jones（VJ）检测器

- 方法流程：

  1. 多尺度Haar特征快速计算
  2. 特征选择
  3. 多阶段处理
     - 该方法采用了级联结构。整个检测器由多级Adaboost决策器组成，每一级决策器又由若干个弱分类决策桩组成。级联结构的核心思想是将较少的计算资源分配在背景窗口，而将较多的计算资源分配在目标窗口。如果某一级决策器将当前窗口判定为背景，则无需后续决策就可继续开始下一个窗口的判断。

5.2.4 可变形部件模型

可变形部件模型（DPM）

- 可变形部件模型的主要思想：
  - 为将传统目标检测算法中对目标整体的检测问题拆分并转化为对模型各个部件的检测问题，然后将各个部件的检测结果进行聚合得到最终的检测结果，即“从整体到部分，再从部分到整体”。

➢ 例如：

- 对汽车目标的检测问题可以在可变形部件模型的思想下分解为分别对车窗、车轮、车身等部件的检测问题。
- 对行人的检测问题可以类似地被分解为对人头、四肢、躯干等部件的检测问题。

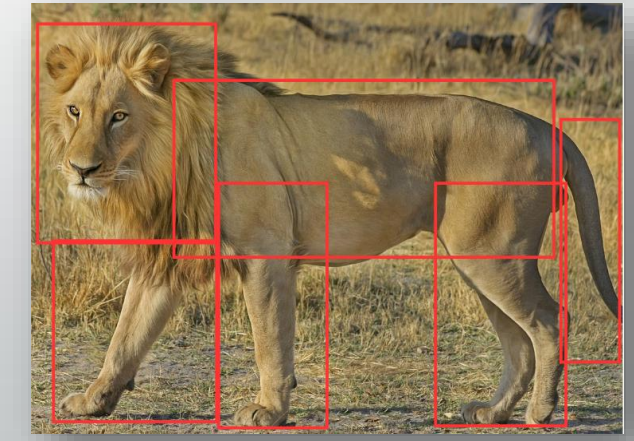


图5-5可变形部件模型部件检测示意图

5.2.4 可变形部件模型

可变形部件模型（DPM）

1. 模型结构

- 整个可变形部件模型检测器由基滤波器和一系列部件滤波器构成。这项工作由Felzenszwalb等人在07年提出，被称为星型模型。后来Girshick在星型模型的基础上进一步将其拓展为混合模型，用于解决真实世界中三维物体不同视角下的检测问题。

5.2.4 可变形部件模型

可变形部件模型（DPM）

1. 模型结构
2. 模型优化

- 由于可变形部件模型在训练过程中并未要求详细标注出各个部件的位置，所以采用了一种弱监督学习的策略。由于部件滤波器可以视为模型中的隐含变量，故Girshick进一步将其转化为隐含变量结构支持向量机的优化问题，并结合难样本挖掘和随机梯度优化策略对该问题进行求解。

5.2.4 可变形部件模型

可变形部件模型（DPM）

1. 模型结构
2. 模型优化
3. 模型加速

- Girshick曾将可变形部件模型中的线性支持向量机分类器"编译"为一系列的级联决策桩分类器，在不牺牲精度的前提下，将可变形部件模型加速了10倍。值得一提的是，这种加速策略本质上是借鉴了VJ检测器快速检测人脸技术的思路。

5.2.4 可变形部件模型

可变形部件模型（DPM）

1. 模型结构
2. 模型优化
3. 模型加速
4. 算法后处理

- 可变形部件模型算法采用边界框回归和上下文信息集成进一步提升检测准确率。
  1. 边界框回归的主要作用是将检测得到的基滤波器以及部件滤波器所对应的边界框进行整合，并利用线性最小二乘回归来得到最终精确的边界框坐标。
  2. 上下文信息集成的作用是利用全局信息对检测结果进行重新调整。本质上，上下文信息反映了各个类别的目标在图像中的联合先验概率密度分布，即哪些类别的目标可能同时出现，哪些类别的目标则不太可能同时出现。

5.2.4 可变形部件模型

可变形部件模型（DPM）

虽然近几年基于深度学习的检测模型从精度上已远远超越了可变形部件模型，但可变形部件模型DPM中的很多思想直到今天依然重要：

- ✓ 混合模型
- ✓ 难样本挖掘
- ✓ 边界框回归
- ✓ 上下文信息的利用等。

■ 时至今日，这些方法还都深深影响着目标检测领域的发展。

5.3 深度学习目标检测

比较流行的基于深度学习的目标检测算法可以分为两类：

1. 基于候选区域的R-CNN系算法（R-CNN、Fast R-CNN、Faster R-CNN等），它们是两阶段的算法，需要先通过算法产生目标候选框，也就是目标位置，然后再对候选框做分类与回归。
2. Yolo，SSD这类单阶段算法，其仅仅使用一个卷积神经网络CNN直接预测不同目标的类别与位置。

◆ 第一类算法的准确度高一些，但是速度慢；

◆ 第二类算法的速度快，但是准确性要低一些。

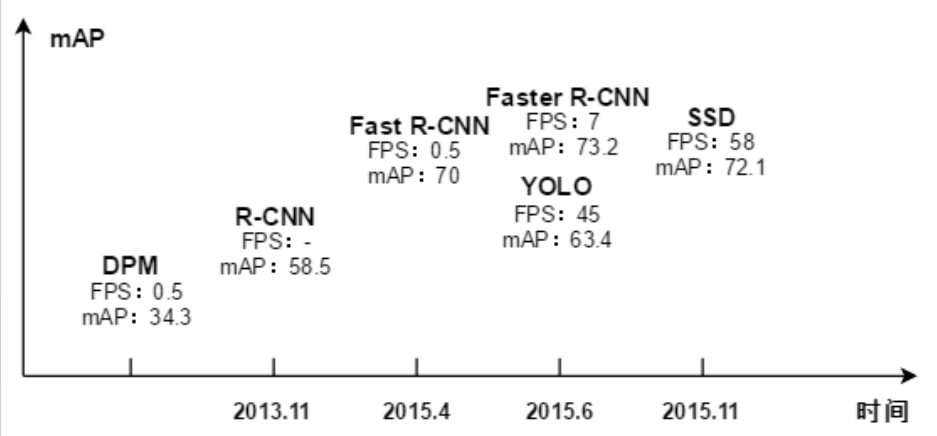


图5-6 深度学习时代目标检测算法发展概览

5.3.1 两阶段检测方法

5.3.1.1 R-CNN

- R-CNN类算法基本流程相似，一般分为两个步骤：
  1. 生成目标候选区域
  2. 使用神经网络对候选区域做分类和回归
- 目标检测有两个主要任务：物体分类和定位。

■ R-CNN [1]借鉴了滑动窗口思想，采用对区域进行识别的方案包含以下步骤：

1. 输入一张图片，通过指定算法从图片中提取2000个类别独立的候选区域（可能目标区域）。
2. 对于每个候选区域利用卷积神经网络获取一个特征向量。
3. 对于每个区域相应的特征向量，利用支持向量机进行分类，并通过一个边界框回归，调整目标边界框的大小。

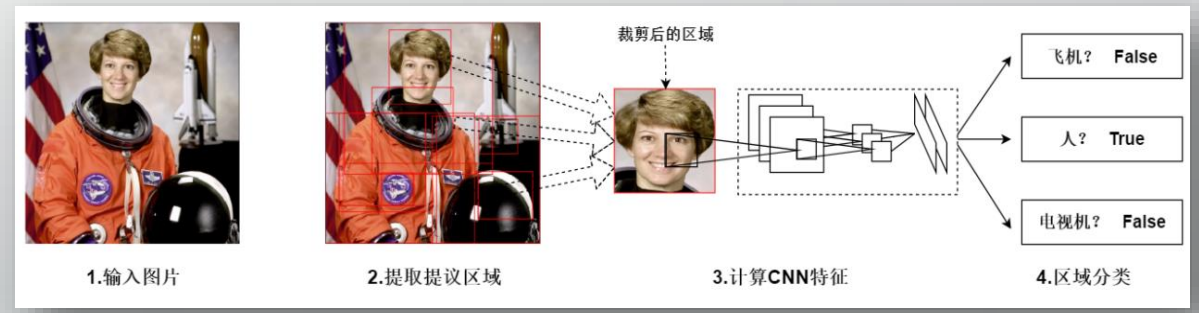


图5-7 R-CNN流程示意图

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. Piscataway, NJ: IEEE, Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

5.3.1 两阶段检测方法

5.3.1.2 Fast R-CNN

- R-CNN存在一些问题：
  - 选择性搜索算法效率低，对一张图像做候选区域提议，需要花费2s；
  - 串行式CNN前向传播效率低，对于每一个候选框，都需经过一个AlexNet提取特征，为所有的候选框提取特征大约花费47s；
  - 三个模块（CNN特征提取、支持向量机分类和边框修正）是分别训练的，并且在训练的时候，对于存储空间的消耗很大。
- 针对R-CNN的缺陷，Fast R-CNN[2]进行了改进（如图5-9）
- Fast R-CNN的解决方案包含以下步骤

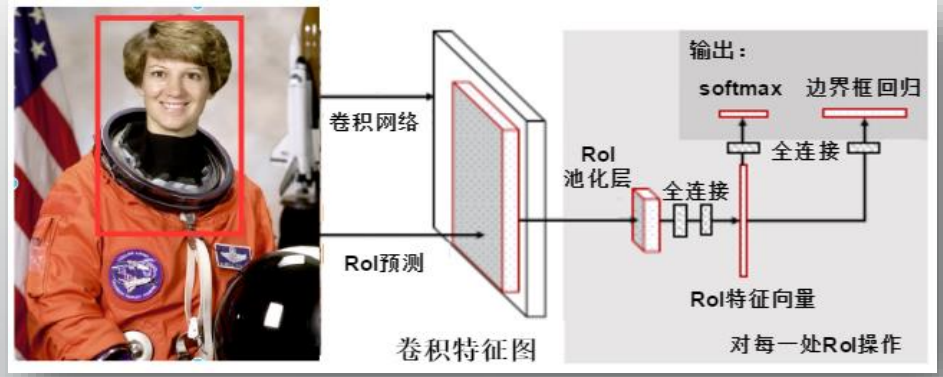


图5-7 R-CNN流程示意图

[2] Girshick R. Fast r-cnn[C]. Piscataway, NJ: IEEE, Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

5.3.1 两阶段检测方法

5.3.1.2 Fast R-CNN

- Fast R-CNN的解决方案包含以下步骤：

1. 采用选择性搜索提取2000个候选框，即感兴趣区域(ROI)。
2. 使用一个卷积神经网络对全图进行特征提取。
3. 使用一个感兴趣区域池化层在全图特征上摘取每一个感兴趣区域对应的特征。
4. 分别经过21和84维的全连接层（并列的，前者是分类输出，后者是回归输出），Fast R-CNN通过CNN直接获取整张图像的特征图，再使用感兴趣区域池化层在特征图上获取对应每个候选框的特征，避免了R-CNN中的对每个候选框串行进行卷积。

5.3.1 两阶段检测方法

5.3.1.2 Fast R-CNN

- 感兴趣区域池化层主要做了两件事：
  1. 为每个感兴趣区域选取对应的特征，
  2. 为了满足全连接层的输入需求，将每个感兴趣区域对应特征的维度转化成某个定值。
- 感兴趣区域池化层的包含以下步骤：
  1. 候选区域划分为目标H×W大小的分块。
  2. 对每一个分块做最大池化。
  3. 将所有输出值组合起来便形成固定大小为H×W的特征图。

5.3.1 两阶段检测方法

5.3.1.2 Fast R-CNN

- Fast R-CNN的贡献主要有：
  - 取代R-CNN的串行特征提取方式，直接采用一个CNN对全图提取特征；
  - 加入了多任务损失函数，除了选择性搜索Selective Search，其他部分都可以合在一起训练。

Fast R-CNN也有缺点，比如耗时的选择性搜索Selective Search依旧存在。

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

- Faster R-CNN[3]提出了区域提议网络（RPN）取代选择性搜索，直接通过一个区域提议网络生成待检测区域。
- 在生成感兴趣区域区域的时候，时间从2s缩减到了10ms。

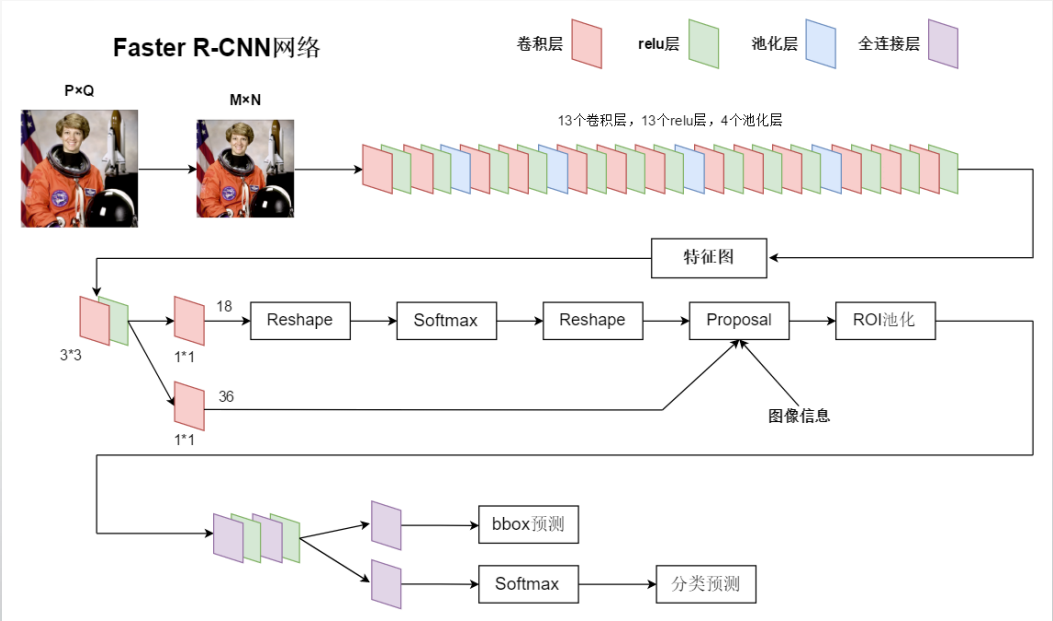


图5-10 Faster R-CNN结构图

[3]. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Cambridge, Massachusetts: MIT Press, Advances in neural information processing systems. 2015: 91-99

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

- Faster R-CNN由共享卷积层、区域提议网络、感兴趣区域池化以及分类和回归四部分组成。

1. 首先使用共享卷积层为全图提取特征图。
2. 将得到的特征图送入区域提议网络，区域提议网络生成待检测框(指定感兴趣区域的位置)，并对感兴趣区域的边界框进行第一次修正。
3. 感兴趣区域池化层根据区域提议网络的输出在特征图上面选取每个感兴趣区域对应的特征，并将维度置为定值。
4. 使用全连接层对框进行分类，并且进行目标边界框的第二次修正。

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

- 区域提议网络的具体结构
  - 区域提议RPN网络分为2条支线：
    1. 上面一条支线通过Softmax来分类锚点，获得前景和背景（检测目标是前景），
    2. 下面一条支线用于计算锚点的边框偏移量，以获得精确的候选区域。
    3. 最后的候选区域提议层则负责综合前景锚点和偏移量获取候选区域，同时剔除太小和超出边界的候选区域。整个网络在候选区域提议层即完成了目标定位的功能。

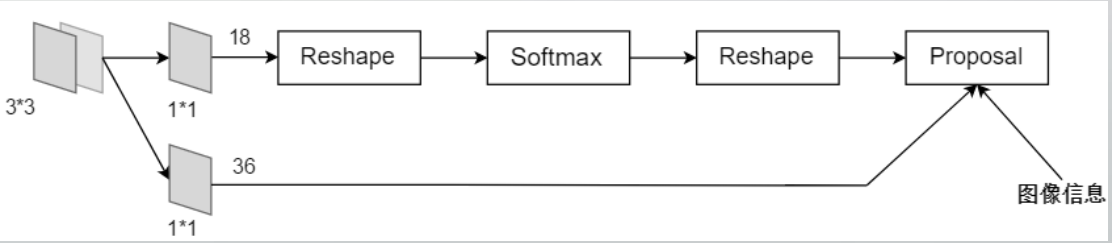


图5-11 区域提议网络的结构

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

1. 锚框

- 区域提议网络依靠一个在共享特征图上滑动的窗口，为每个位置生成9种预先设置好长宽比与面积的目标框，即锚框。这9种初始锚框Anchor包含三种面积(128×128，256×256，512×512)，每种面积又包含三种长宽比(1:1，1:2，2:1)，如图5-12。
- 由于共享特征图的大小约为40×60，所以区域提议网络RPN生成的初始锚框Anchor的总数约为20000个(40×60×9)。

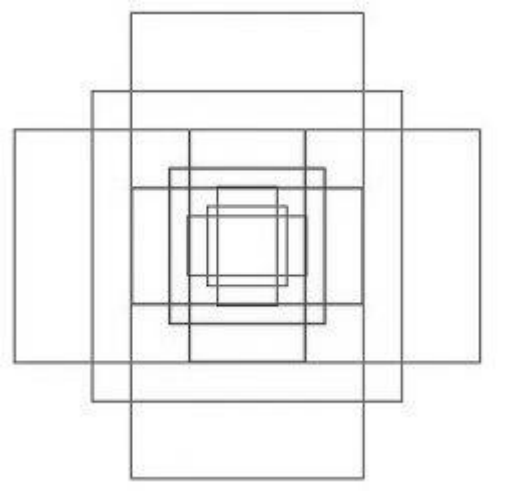


图5-12 锚框示意图

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

1. 锚框
2. 判断前景或背景

- 对于所有的锚框，首先需要判断锚框是否为前景。区域提议网络使用Softmax损失直接训练，在训练的时候排除掉了超越图像边界的锚框Anchor。

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

1. 锚框
2. 判断前景或背景
3. 边框修正

- 如图5-14，较大的框表示的是飞机的实际框标签，内部的框表示的其中一个候选区域(前景锚框)，即被分类器识别为飞机的区域，但是由于预测区域定位不准确，这张图相当于没有正确检测出飞机，所以希望采用一种方法对内部的预测框进行微调，使得候选区域和实际框更加接近。

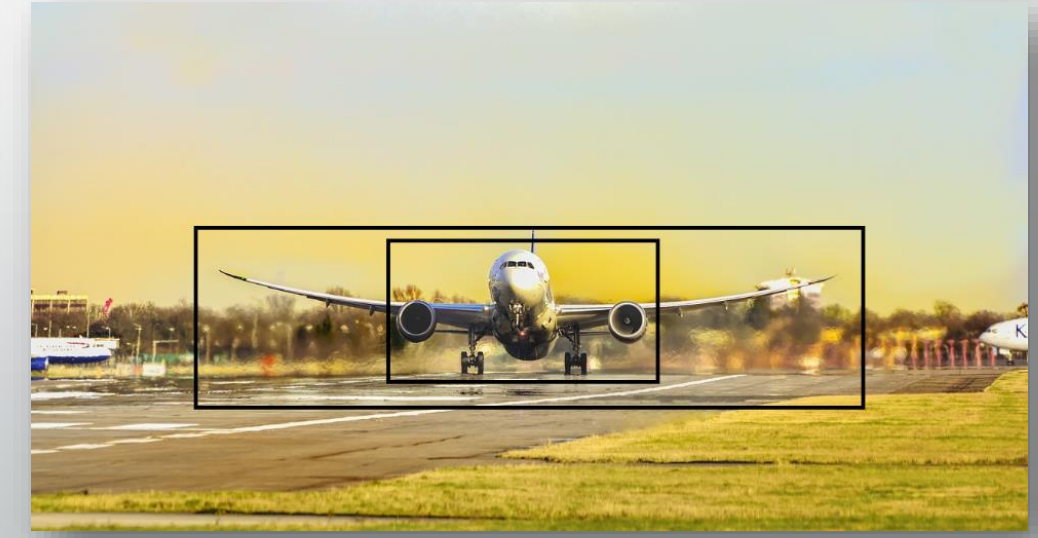


图5-14 候选框与实际框标签

5.3.1 两阶段检测方法

5.3.1.3 Faster R-CNN

1. 锚框
2. 判断前景或背景
3. 边框修正

- 对于窗口一般使用四维向量 (x, y, w, h) 表示，分别表示窗口的中心点坐标和宽高。如图5-14，中心点位于左下角的框A代表原始的正锚框，中心点位于右上角的框G代表目标的实际框标签，目标是寻找一种关系，使得原始输入的锚框A经过映射得到一个跟真实窗口G更接近的回归窗口G'。

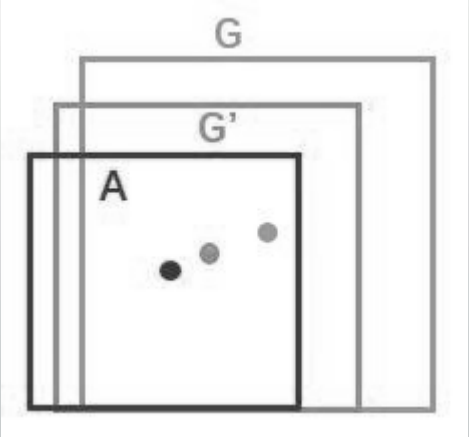


图5-15 锚框与实际框标签映射图

5.3.1 两阶段检测方法

5.3.1.4 Mask R-CNN

- Mask R-CNN是一个非常灵活的框架，可以增加不同的分支用来完成不同的任务，包括目标分类、目标检测、语义分割、实例分割、人体姿势识别等多种任务。
- Mask R-CNN[4]可以分解为如图5-16所示的3个模块：Faster-RCNN、感兴趣区域对齐和掩模。

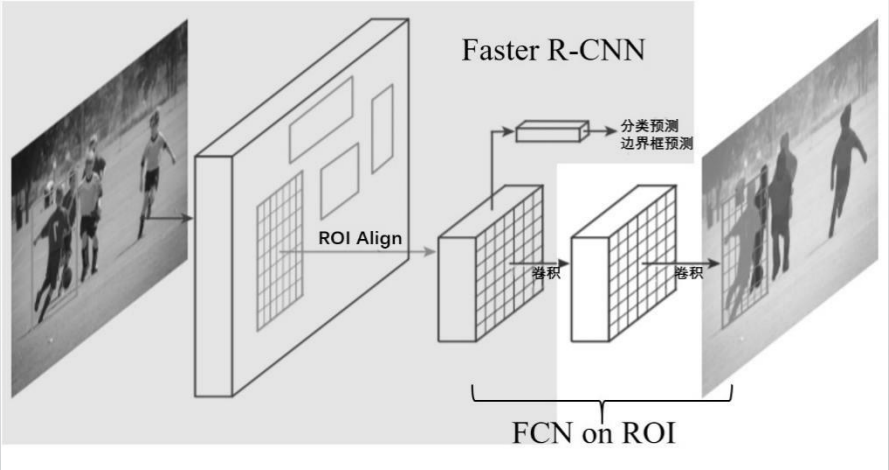


图5-16 Mask R-CNN流程图

[4]. He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Piscataway, NJ: IEEE, Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

5.3.1 两阶段检测方法

5.3.1.4 Mask R-CNN

- 方法流程：

1. 输入一幅待处理的图片，进行对应的预处理操作，或者直接输入预处理后的图片。
2. 将其输入到一个预训练好的神经网络中（ResNeXt等）获得对应的特征图。
3. 对这个特征图中的每一点设定预定个数的感兴趣区域，从而获得多个候选感兴趣区域。
4. 将这些候选的感兴趣区域送入区域提议网络进行二值分类（前景或背景）和边界框回归，过滤掉一部分候选的感兴趣区域。
5. 对这些剩下的感兴趣区域进行感兴趣区域对齐操作（即先将原图和特征图的像素对应起来，然后将特征图和固定的特征区域对应起来）。
6. 对这些感兴趣区域进行分类（N类别分类）、边界框回归和掩模生成（在每一个感兴趣区域里面进行FCN操作）。

5.3.1 两阶段检测方法

5.3.1.4 Mask R-CNN

- Mask R-CNN使用感兴趣区域对齐取代了Faster R-CNN中的感兴趣区域池化。
- 感兴趣区域池化和感兴趣区域对齐最大的区别是：
  - 前者使用了两次量化操作，后者并没有采用量化操作，而是使用了线性插值算法，如图5-17和5-18所示。

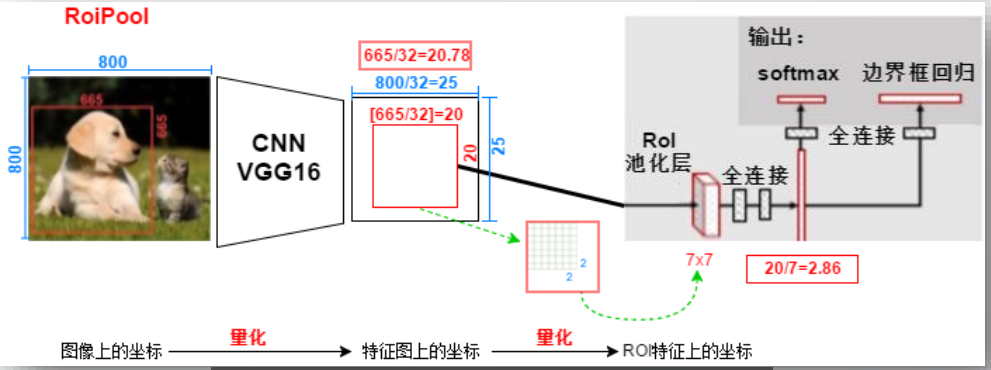


图5-17感兴趣区域池化的计算流程

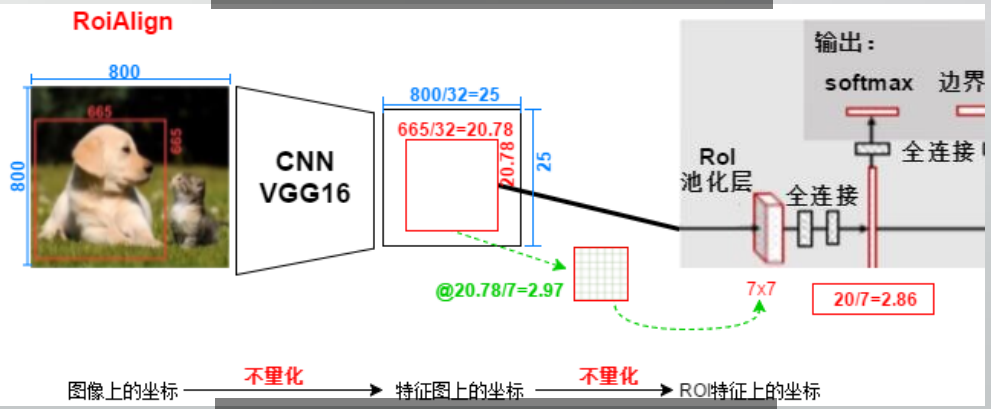


图5-18 感兴趣区域对齐的计算流程

5.3.1 两阶段检测方法

5.3.1.4 Mask R-CNN

- 在Mask R-CNN中的感兴趣区域对齐之后有一个头部分，主要作用是将感兴趣区域对齐的输出维度扩大，这样在预测掩模时会更加精确，称为掩模分支Mask Branch。
  - 在掩模分支Mask Branch的训练环节，作者没有采用全卷积神经网络式的Softmax损失，反而是输出了K个掩模预测图(为每一个类都输出一张)，并采用平均二值交叉熵损失训练，当然在训练掩模分支的时候，输出的K个特征图中，也只是对应真实标签类别的那一个特征图对掩模损失有贡献。

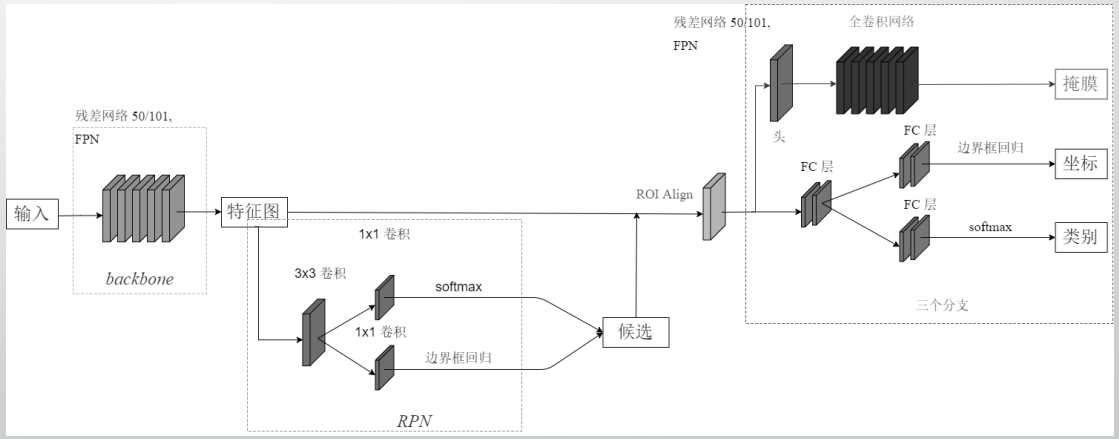


图5-20 掩模分支计算流程

5.3.2 单阶段检测方法

单阶段检测方法

- 单阶段的检测方法流程可以大致认定为将一个网络分成卷积层、目标检测层和筛选层三部分的一种目标检测方法流派。
- 接下来我们将学习Yolo和SSD。

5.3.2 单阶段检测方法

5.3.2.1 Yolo

- Yolo[5]创造性地提出了单阶段检测方法，也就是将物体分类和物体定位在一个步骤中完成。
- Yolo直接在输出层回归边界框的位置和边界框所属类别，从而实现单阶段检测方法。
- Yolo可实现45帧每秒的运算速度，完全能满足实时性要求（达到24帧每秒，人眼认为是连续的）
- 网络结构如图5-21所示。主要分为三个部分：卷积层，目标检测层，非极大值抑制筛选层。

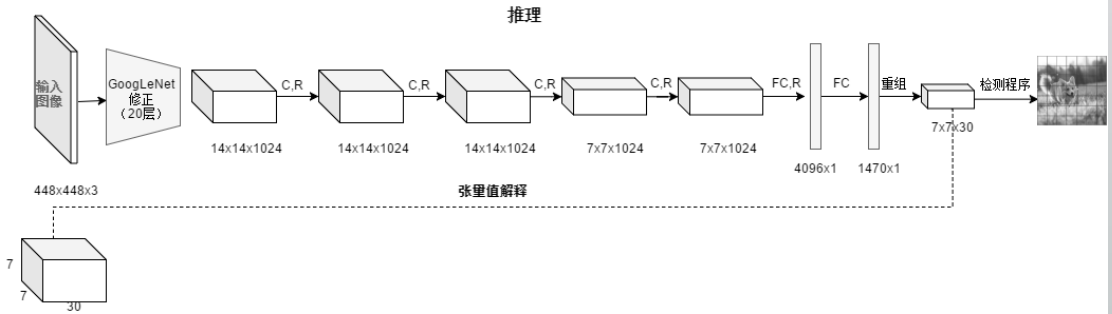


图5-21 Yolo算法流程图

[5]. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Piscataway, NJ: IEEE, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

5.3.2 单阶段检测方法

5.3.2.1 Yolo

1. 卷积层

- 采用Google inception V1网络，是这个模型的第一个阶段，共20层。但作者对inception V1进行了改造，没有使用inception module结构，而是用一个1x1的卷积，并联一个3x3的卷积替代。
- 这一层主要是进行特征提取，从而提高模型泛化能力。

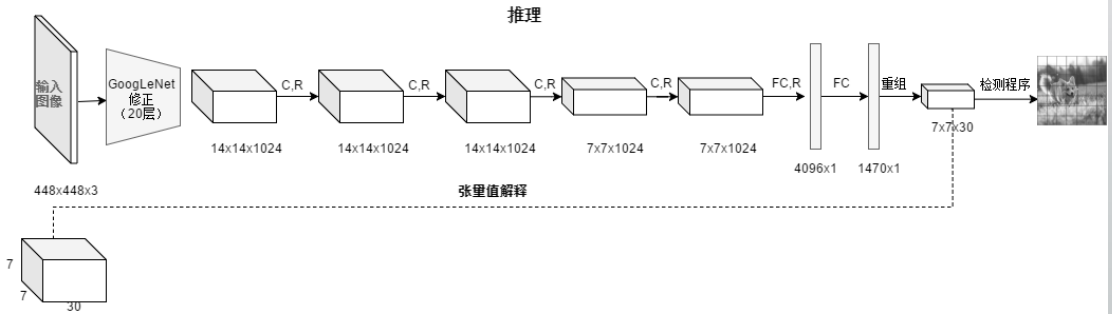


图5-21 Yolo算法流程图

5.3.2 单阶段检测方法

5.3.2.1 Yolo

1. 卷积层
2. 目标检测层

- 先经过4个卷积层和2个全连接层，最后生成7x7x30的输出。先经过4个卷积层的目的是为了提高模型泛化能力。Yolo将一副448x448的原图分割成了7x7个网格，然后每个单元格负责检测那些中心点落在该格子内的目标，

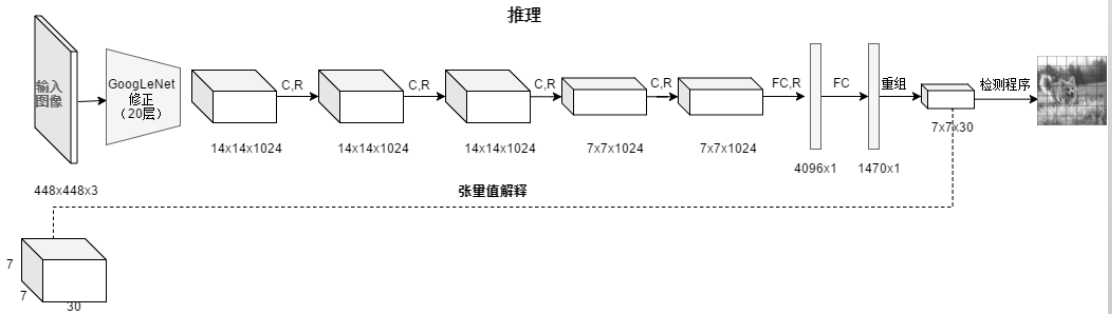


图5-21 Yolo算法流程图

5.3.2 单阶段检测方法

5.3.2.1 Yolo

1. 卷积层
2. 目标检测层

- 如图5-22所示，可以看到狗这个目标的中心落在中间靠右的一个单元格内，那么该单元格负责预测这个狗。每个单元格会预测B个边界框以及边界框的置信度。
- 所谓置信度其实包含两个方面，一是这个边界框含有目标的可能性大小，二是这个边界框的准确度。

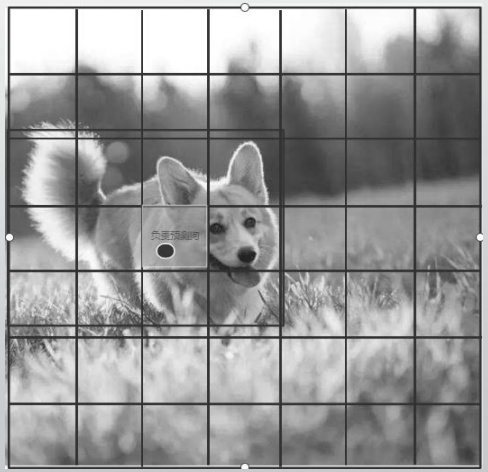


图5-22 Yolo目标检测示意图

5.3.2 单阶段检测方法

5.3.2.1 Yolo

1. 卷积层
2. 目标检测层
3. 非极大值抑制筛选层

- 非极大值抑制筛选层是为了在多个结果中（多个边界框）筛选出最合适的几个，这个方法和Faster R-CNN 中基本相同。都是先过滤掉置信度低于阈值的边界框，对剩下的边界框进行非极大值抑制，去除掉重叠度比较高的边界框，这样就得到了最终的最合适的几个边界框及其类别。

5.3.2 单阶段检测方法

5.3.2.1 Yolo

- Yolo算法开创了单阶段检测的先河，它将物体分类和物体检测网络合二为一，都在全连接层完成。
- 故它大大降低了目标检测的耗时，提高了实时性。

它的缺点也十分明显：

1. 每个网格只对应两个边界框，当物体的长宽比不常见（即训练数据集覆盖不到时），效果很差；
2. 原始图片只划分为7x7的网格，当两个物体靠的很近时，效果很差；
3. 最终每个网格只对应一个类别，容易出现漏检，即物体没有被识别到的情况；
4. 对于图片中比较小的物体，效果很差。

5.3.2 单阶段检测方法

5.3.2.2 SSD

- Faster R-CNN准确率mAP较高，召回率Recall较低，但速度较慢。
- Yolo速度快，但准确率和漏检率不尽人意。
- SSD[6]综合了他们的优缺点
- SSD的网络结构如图5-23所示：

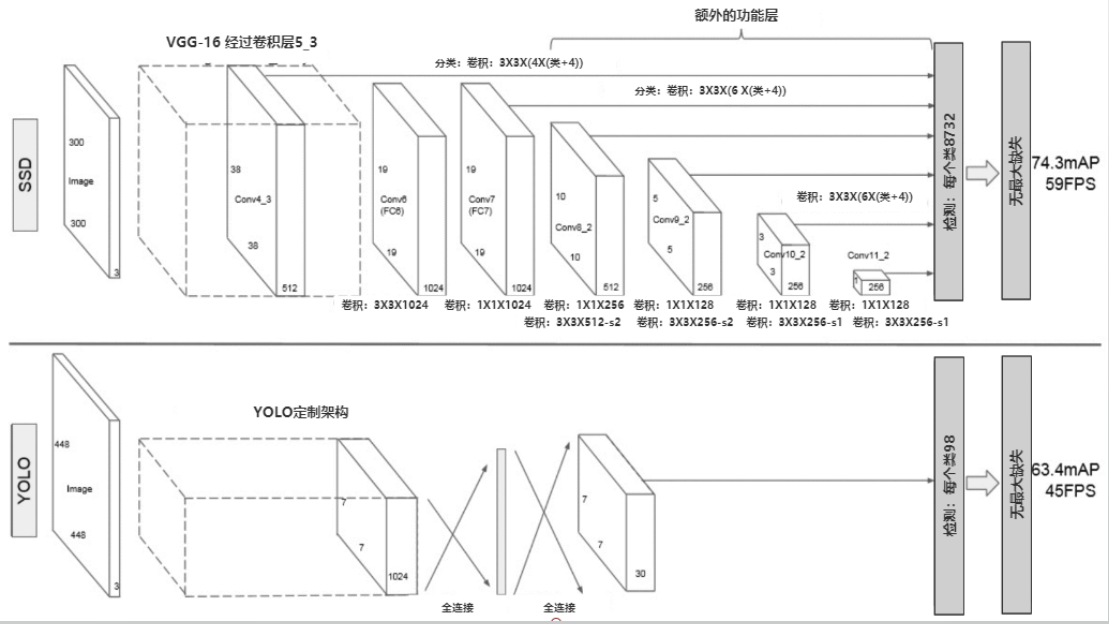


图5-23 SSD网络结构示意图

[6]. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. German: Springer, European conference on computer vision. Springer, Cham, 2016: 21-37.

5.3.2 单阶段检测方法

5.3.2.2 SSD

- 与Yolo类似，SSD也分为三部分：卷积层，目标检测层和非极大值抑制NMS筛选层。

1. 卷积层
   - SSD采用了VGG16的基础网络，先用一个CNN网络来提取特征，然后再进行后续的目标定位和目标分类识别。
2. 目标检测层
   - 目标检测层由五个卷积层和一个平均池化层组成。去掉了最后的全连接层。SSD认为目标检测中的物体，只与周围信息相关，它的感受野不是全局的，故没必要也不应该做全连接。

5.3.2 单阶段检测方法

5.3.2.2 SSD

1. 多尺寸特征图上进行目标检测

- 每一个卷积层，都会输出不同大小感受野的特征图。在这些不同尺度的特征图上，进行目标位置和类别的训练和预测，从而达到多尺度检测的目的，可以克服Yolo对于宽高比不常见的物体，识别准确率较低的问题。

- Yolo中，只在最后一个卷积层上做目标位置和类别的训练和预测。这是SSD相对于Yolo能提高准确率的一个关键所在。

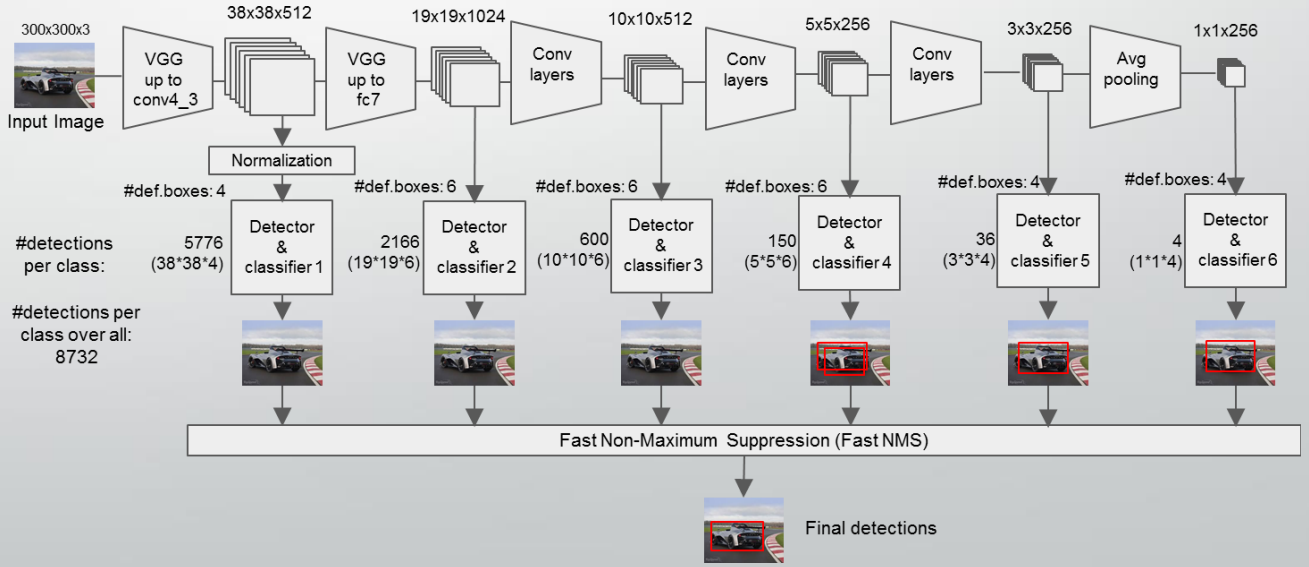


图5-24 SSD的多尺度检测示意图

5.3.2 单阶段检测方法

5.3.2.2 SSD

1. 多尺寸特征图上进行目标检测

2. 设置先验框

- 在Yolo中，每个单元预测多个边界框，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。
- SSD也提出了锚框的概念。卷积输出的特征图，每个点对应为原图的一个区域的中心点。以这个点为中心，构造出6个宽高比例不同，大小不同的锚框，SSD中称为先验框。每个锚框对应4个位置参数(x,y,w,h)和21个类别概率（voc训练集为20分类问题，再加上锚框是否为背景，共21分类）。

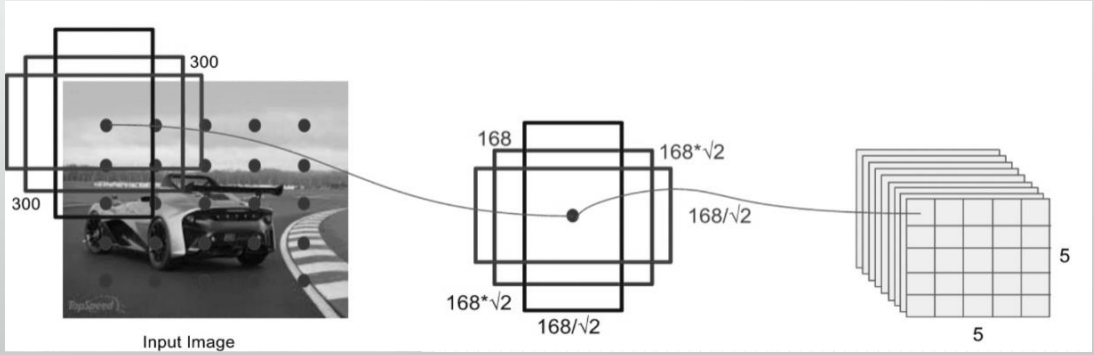


图5-25 SSD的先验框示意图

5.3.2 单阶段检测方法

5.3.2.2 SSD

1. 多尺寸特征图上进行目标检测
2. 设置先验框
3. 筛选层

- 与Yolo的筛选层基本一致，同样先过滤掉类别概率低于阈值的先验框，再采用非极大值抑制，筛掉重叠度较高的。只不过SSD综合了各个不同特征图上的目标检测输出先验框。