《人工智能导论》探究报告
目标检测

姓名：郭江彤　　学号:2113485　　R-CNN
姓名：蒋薇　　学号:2110957　　Fast R-CNN
姓名：任鸿宇　　学号:2113630　　Faster R-CNN

作业正文

1 问题描述

1.1 文字描述

人工智能目标检测问题是指在图像或视频中识别出特定的目标并标注出其位置和大小的问题。具体来说，该问题需要解决以下几个子问题：

1、目标分类：确定要检测的目标属于哪一类别，比如人、车、动物等。

2、目标定位：确定目标在图像或视频中的位置，通常使用矩形框来标注出目标的位置和大小。

3、目标识别：对于同一类别的不同目标，能够识别出它们之间的差异，比如不同的人、不同的车等。

人工智能目标检测问题的解决可以通过深度学习等技术实现。常见的方法包括基于卷积神经网络（CNN）的目标检测算法，如 RCNN、Fast R-CNN、Faster R-CNN、YOLO、SSD 等。这些算法通过对图像或视频进行卷积和池化等操作，从而实现对目标的分类、定位和识别。

1.2 公式化描述

给定一张图像和一组预定义的物体类别，目标检测问题的目标是在图像中检测这些物体的位置和大小，并将其标记为相应的类别。更形式化地说，目标检测问题可以定义为：给定一个输入图像 I 和一组类别 C=c1,c2,⋯,cn，找到图像中所有属于 C 中某个类别的物体的边界框 B=b1,b2,⋯,bm，其中每个边界框 bi 表示物体的位置和大小，并将其标记为相应的类别 cj C。因此，目标检测问题可以看作是一种多类别分类和边界框回归的组合任务。

2 核心内容

2.1 R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

2.1.1 背景和主要贡献

文提出了一种简单且可伸缩的目标检测算法,相比之前最好的算法,在 2010VOC 数据集上 mAP 有了很大的提升。本文的方法组合了两个关键点：1. 把卷积神经网络应用到候选区域的定位和分隔；2. 当带标签的数据稀少的时候，先用预训练作为辅助任务，然后用特定领域的数据进行微调，也会得到性能的提升。因为本文组合了候选区域和 CNN 特征, 所以该方法成为 R-CNN。本文也比较了 R-CNN 和基于滑窗的检测算法 OverFeat, 发现在 ILSVRC2013 数据集上，R-CNN 优于 OverFeat 算法。文章涉及的方法与主要贡献：在此之前，我们使用的是 overfeat 模型进行目标检测，这是一种暴力穷举的方法，从左到右、从上到下滑动窗口，利用分类识别目标。这种方法会消耗大量的计算力量，并且由于窗口大小问题可能会造成效果不准确。但是提供了一种解决目标检测问题的思路。此文章采用的是候选区域方法（region proposal method）, 创建目标检测的区域改变了图像领域实现物体检测的模型思路，R-CNN 是以深度神经网络为基础的物体检测的模型。CNN 在当时以优异的性能令世人瞩目，以 R-CNN 为基点，后续的 SPPNet、Fast R-CNN、Faster R-CNN 模型都是照着这个物体检测思路。

2.1.2 研究方法

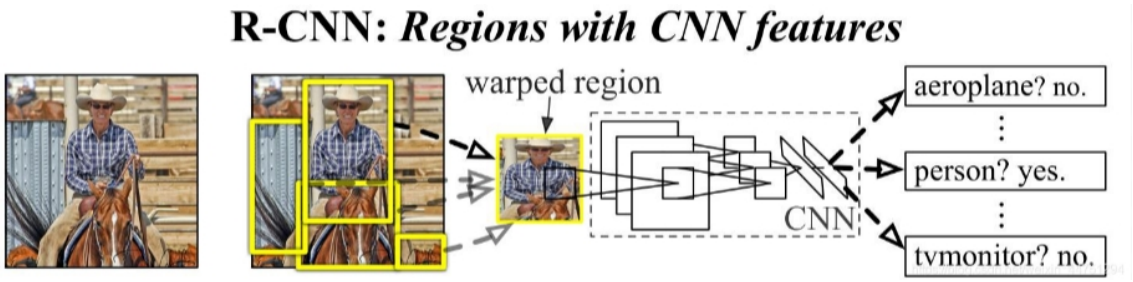


图 1: 示意图

对于给定的输入图像，使用选择性搜索 (selective search) 的区域建议方法提取出大约 2000 个候选区域，即首先过滤掉那些大概率不包含物体的区域，通过这一阶段将原始图像中需要处理的区域大大减少; 选择性搜索是指首先将每个像素作为一组。然后，计算每一组的纹理，并将两个最接近的组结合起来。但是为了避免单个区域吞噬其他区域，我们首先对较小的组进行分组。我们继续合并区域，直到所有区域都结合在一起。对每个候选区域，使用深度网络提取特征，该步可以分为两小步：首先需要对第一步中抽取得到的候选区域，经过一个叫做”warp” 的过程，这个 warp 实际就是一个缩放的过程，因为第一步我们提取出的候选区域大小不一，但是后续接入的深度网络的输入是固定的，因此这些区域的大小需要适配 CNN 网络固定大小的输入；然后将第一小步中 warp 之后的候选区域接入到卷积

神经网络，抽取一个相对低维的特征；假设一张图片的 2000 个侯选区域，那么提取出来的就是 2000 x 4096 这样的特征向量（R-CNN 当中默认 CNN 层输出 4096 特征向量）。R-CNN 选用 SVM 进行二分类。假设检测 20 个类别，那么会提供 20 个不同类别的 SVM 分类器，每个分类器都会对 2000 个候选区域的特征向量分别判断一次，这样得出 [2000, 20] 的得分矩阵。SVM 的权值矩阵中的每一列代表了一类的权值；而相乘后所得到的矩阵的每一行，代表了一个候选框的 20 个类别的概率，2000 表示一共有 2000 个候选框。然后分别对上述 2000×20 维矩阵中每一列即每一类进行非极大值抑制剔除重叠建议框，得到该列即该类中得分最高的一些建议框。对 NMS 处理后剩余的建议框进一步筛选。具体做法是，保留与真实标准的边界框有相交的，并且 iou 要大于某一个阈值，不满足就要将其删除掉。接着分别用 20 个回归器对上述 20 个类别中剩余的建议框进行回归操作，最终得到每个类别的修正后的得分最高的 bounding box。这是针对卷积神经网络输出的特征向量进行预测的。利用每一个边界框 4096 维的特征信息来进行预测的。通过回归分类器之后，会得到四个参数，分别得到目标建议框中心的 x 偏移量与 y 偏移量，以及边界框的高度缩放因子与宽度缩放因子。一共四个值，通过这四个值来对建议框进行调整，得到一个红色的建议框。

2.2 **规定论文题目** 2-Fast R-CNN

论文题目:Girshick, Ross. **"**Fast r-cnn.**"** Proceedings of the IEEE international conference on computer vision. 2015.

2.2.1 **背景**

复杂性的产生是因为检测需要目标的精确定位，这就导致两个主要的难点。首先，必须处理大量候选目标位置 (通常称为 "提案")。第二，这些候选框仅提供粗略定位，其必须被精细化以实现精确定位。这些问题的解决方案经常会影响速度，准确性或简单性。

2.2.2 R-CNN、SPP **缺点**

R-CNN 是多阶段模型，需要使用 Selective Search 选择 proposals，然后对每个 proposals 进行 resize，卷积，全连接。可以发现缺点就是每个 proposal 都需卷积，而且尺寸固定。SPPNet 原理就是利用 ROI 池化层将 CNN 的输入从固定尺寸改为任意尺寸，通过最大池化层，可以将任意宽度的、高度的卷积特征转换成固定长度的向量，原始图像中的候选框，实际也可以对应到卷积特征中相同位置的框。利用 SPP 层可以将不同形状的特征对应到相同长度的向量特征。与 R-CNN 相比

比，SPPNet 具有更快的速度。但是 SPPNet 难以（效率低）对 SPP 层下面所有的卷积层进行后向传播，Fast R-CNN 了借鉴 SPPnet 的空间金字塔池化层思想，解决了 SPPNet 对卷积层进行后向传播效率太低的缺点。

2.2.3 主要内容

这篇文章中，简化了基于卷积网络的目标检测器的训练过程，提出了一个单阶段训练笔法，联合学习候选框分类和修正他们的空间位置。

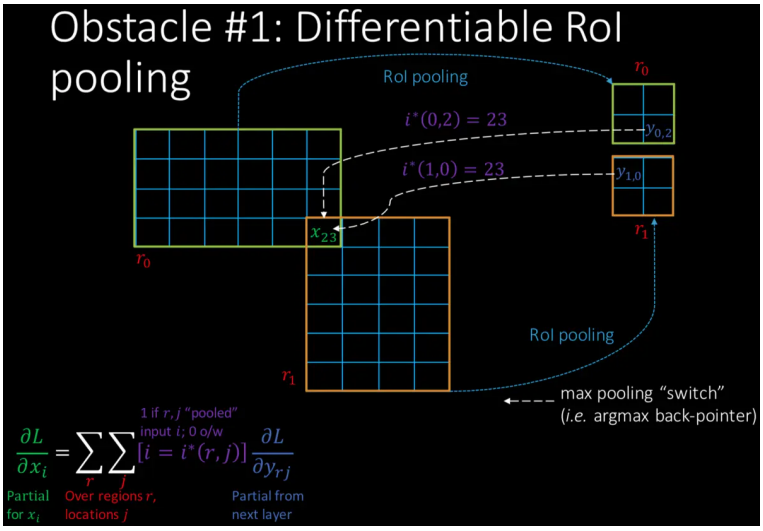
fast R-CNN **架构与训练**

Fast R-CNN 架构输入图像和多个感兴趣区域 (Rol) 被输入到全卷积网络中，每个 Rol 被池化到固定大小的特征图中，然后通过全连接层 (FC) 映射到特征向量。网络对于每个 Rol 具有两个输出向量:Soflmax 概率和每类检测框回归停移量。该架构是使用多任务丢失惴到端训练的。

Fast R-CNN 网络将整个图像和一组候选框作为输入。网络首先使用几个卷积层 (conv) 和最大池化层来外理整个图像，以产生卷积特征图，然后，对于每个候选框，Rol 池化层从特征图中提取固定长度的特征向量，每个特征向量被送入一系列全连接 (fc) 屏中，其最终分支成两个同级输出层: 一个输出 KK 个类别加上 1 个背暑类别的 Softmax 概率估计，另一个为 KK 个类别的每一个类别输出四个实数值。每组 4 个值表示 KK 个类别的一个类别的检测框位置的修正。

Rol 池化层

Rol 池化层使用最大池化将任何有效的 Rol 内的特征转换成具有 HxWHxW(例如，7x77x7) 的固定空间范围的小特征图，其中 HH 和 WW 是层的超参数，独立于任何特定的 Rol。在本文中，Rol 是卷积特征图中的一个矩形窗口。每个 Rol 由指定其左上角 (r.c)(r.c) 及其高度和宽度 (h.w)(h.w) 的四元组 (r.c.h.w)(r,c,h,w) 定义。Rol 最大池化通过将大小为 hxwh*w 的 Rol 窗口分割成 H*WH*W 个网格，子窗口大小约为 h/HxwWh/HxwW，然后对每个子窗口执行最大池化，并将输出合并到相应的输出网格单元中。同标准的最大池化一样，池化操作独立应用于每个特征图通道。Rol 层只是 SPPnets 5 中使用的空间金字塔池层的特殊情况，其只有一个金字塔层。



从预处理网络初始化

论文中实验了三个预训练的 ImageNet9 网络，每个网络有五个最大池化层和五到十三个卷积层 (网络详细信息，请参见实验配置)。当预训练网络初始化 fast R-CNN 网络时，其经历三个变换。首先，最后的最大池化层由 Rol 池层代替，其将 H 和 W 设置为与网络的第一个全连接层兼容的配置 (例如，对于 VGG16，H=W=7H=W=7)。然后，网络的最后一格全连接层和 Softmax(其被训练用于 1000 类 ImageNet 分类) 被替换为前面描述的两个同级层 (全连接层和 K+1K+1 个类别的 Softmax 以及类别特定的检测框回归)。最后，网络被修改为采用两个数据输入: 图像的列表和这些图像中的 Rol 的列表。

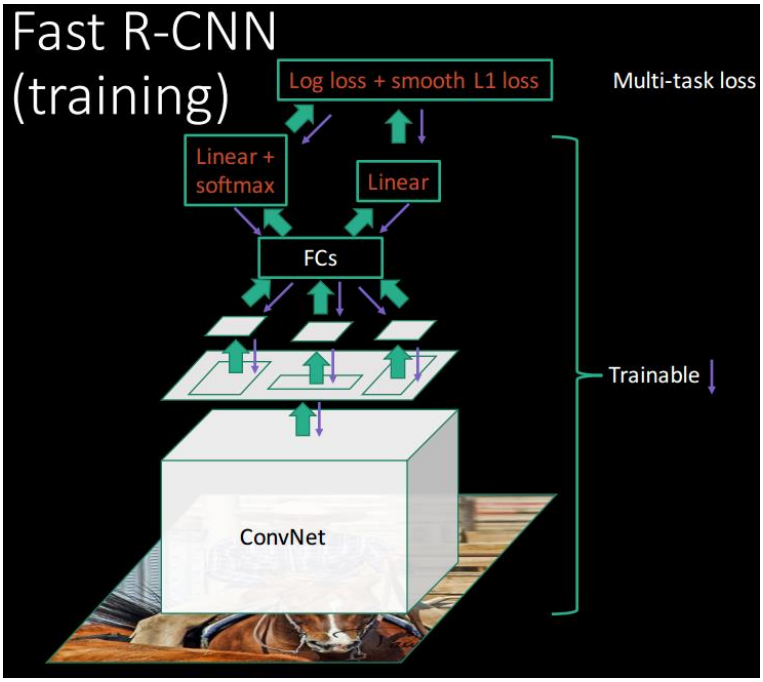
微调

出了一种更有效的训练方法，利用训练期间的特征共享。在 Fast RCNN 网络训练中，随机梯度下降 (SGD) 的小批量是被分层采样的，首先采样 NN 个冬像，然后从每个图像采样 R/NR/N 个 Rol。关键的是，来自同一图像的 Rol 在向前和向后传播中共享计算和内存。减小 NN，就减少了小批量的计算。例如，当 N=2N=2 和 R=128R=128 时，得到的训练方案比从 128 幅不同的图采样一个 Rol(即 R- CNN 和 SPPnet 的策略) 快 64 倍。

2.2.4 创新

ROI Pooling layer

先对原图进行卷积，得到卷积层，在将 Selective Search 选择的 proposals 对应到卷积层，由于 proposal 尺寸不一样，需要进行 RoI pooling。



Multi-task loss

roposals 给出 bbox 是粗糙的，还需要细化，R-CNN 是直接再训练一个模型专门用于检测 bbox，Fast R-CNN 是把这个任务加入到类别检测里。

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v)$$

其中，
真实类标为u
类标为u的bbox区域：$v = (v_x, v_y, v_w, v_h)$
预测的结果为：$p = (p_0, \ldots, p_K)$，包括背景有K+1类
预测的bbox位置：$t^u = (t^u_x, t^u_y, t^u_w, t^u_h)$

2.2.5 结论

Fast R-CNN，一个对 R-CNN 和 SPPnet 干净，快速的更新，一种快速的基于区域的卷积网络方法，稀疏目标候选区域似乎提高了检测器的质量，过去探索这个问题在时间上过于昂贵，但 Fast R-CNN 使其变得可能。

2.3 规定论文题目 3

2.3.1 论文概述

“Towards Real-Time Object Detection with Region Proposal Networks” 是由 Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun 在 2016 年发表的一篇论文。该论文提出了一种称为 Faster R-CNN 的目标检测框架，该框架使用了一种称

为 Region Proposal Network 的新型神经网络模块，可以在较短的时间内实现高准确度的目标检测。Faster R-CNN 的框架包括两个主要组件：共享卷积层和 RPN。共享卷积层用于提取图像特征，而 RPN 用于生成目标候选区域。RPN 是一种全卷积神经网络，它可以在图像中生成多个不同大小和宽高比的候选区域，并计算每个候选区域是否包含目标。这些候选区域随后被送入 Fast R-CNN 检测器中，以进行目标分类和位置回归。RPN 的设计基于锚点（anchor）的概念。

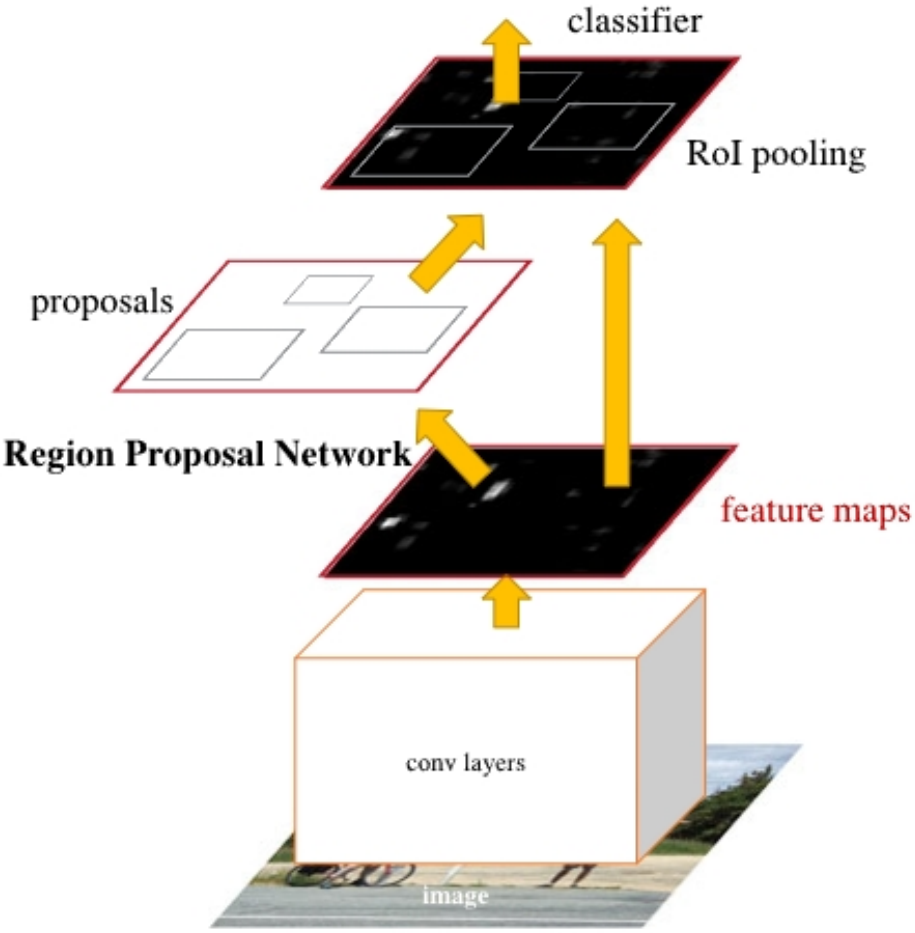


Fig. 2. Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

锚点是预定义的一组框，它们具有不同的大小和宽高比，可以覆盖图像中的各种目标。RPN 使用锚点来生成候选区域，以及预测每个候选区域是否包含目标，并对目标边界框进行回归。Faster R-CNN 框架的主要优点是可以实现准确率和速度之间的平衡。与以往的目标检测方法相比，它不需要在图像中进行滑动窗口搜索，因此可以大大减少计算量。此外，RPN 可以共享卷积层的特征图，从而使整个框架更加高效。

2.3.2 关键贡献

该论文的主要贡献是提出了一种新的目标检测框架，即 Faster R-CNN，该框架通过使用 RPN 来生成目标候选区域，可以实现高准确度和较短的处理时间。该框架的设计具有以下几个关键特点：基于锚点的候选区域生成：RPN 使用锚点来生成候选区域，这些锚点具有不同的大小和宽高比，可以覆盖图像中的各种目标。

共享卷积层：RPN 和 Fast R-CNN 检测器共享卷积层的特征图，从而可以减少计算量并提高效率。

端到端的训练：整个 Faster R-CNN 框架可以端到端地进行训练，从而可以更好地优化整个系统的性能。

2.3.3 实验结果

该论文在多个公共数据集上对 Faster R-CNN 进行了实验，包括 PASCAL VOC 2007、2010 和 2012 以及 MS COCO。实验结果表明，Faster R-CNN 在准确率和速度之间实现了很好的平衡，并且在各个数据集上都取得了最先进的成果。

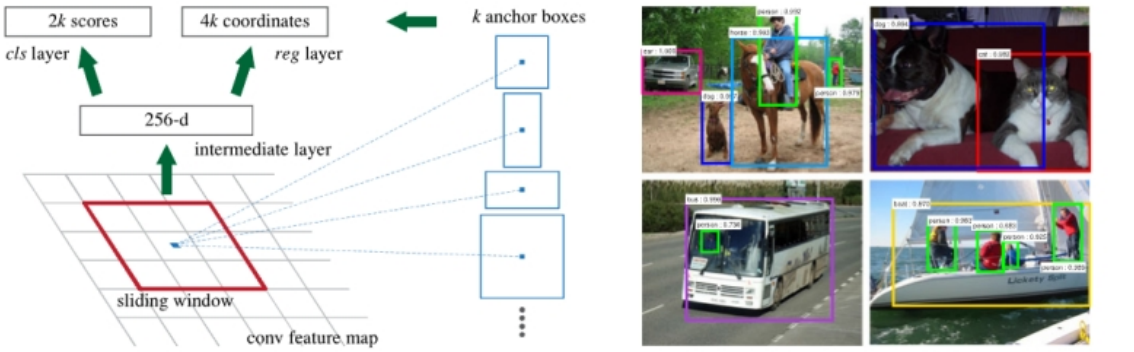


Fig. 3. **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.

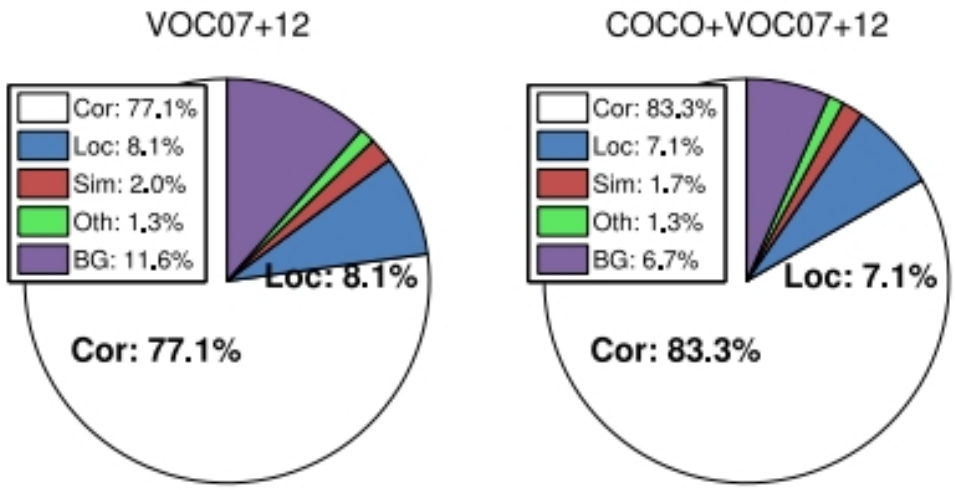


Fig. 7. Error analyses on models trained with and without MS COCO data. The test set is PASCAL VOC 2007 test. Distribution of top-ranked Cor (correct), Loc (false due to poor localization), Sim (confusion with a similar category), Oth (confusion with a dissimlar category), BG (fired on background) is shown, which is generated by the published diagnosis code of [40].

2.3.4 **总结**

"Towards Real-Time Object Detection with Region Proposal Networks" 论文提出了一种新的目标检测框架，即 Faster R-CNN，该框架使用 RPN 来生成目标候选区域，并可以在较短的时间内实现高准确度的目标检测。该框架的设计具有很多优点，包括基于锚点的候选区域生成、共享卷积层和端到端的训练。实验结果表明，Faster R-CNN 在各个数据集上都取得了最先进的成果，成为目标检测领域的经典算法之一。

2.4 **目标检测** – YOLOv1

论文：《You Only Look Once: Unified, Real-Time Object Detection》

论文地址 点击此处 **代码地址** 点击此处

2.4.1 R-CNN **系列不足**

R-CNN 系列算法（R-CNN、SPPNet、Fast R-CNN、Faster R-CNN）均是采用 two-stage 的方法（1. 提取 region proposal 2. 分类 + 边框回归），主要是对 region

proposal 进行识别定位。虽然这类方法检测精度很高，但由于需要一个单独的网络进行提取 region proposal，因此在速度上无法突破瓶颈。

2.4.2 YOLOv1 **创新点**

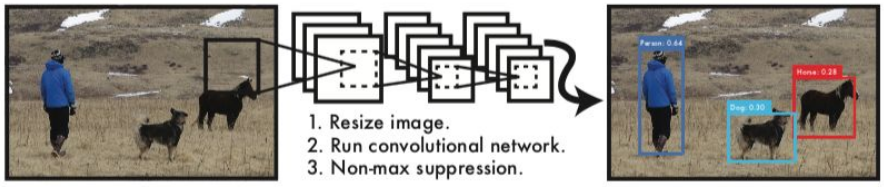
将 detection 视为回归问题，仅使用一个 neural network 同时预测 bounding box 的位置和类别，因此速度很快。

由于不需提取 region proposal，而是直接在整幅图像进行检测，因此 YOLOv1 可以联系上下文信息和特征，减少将背景检测为物体的错误。

YOLOv1 学习到的是目标的泛化表示（generalizable representations），泛化能力非常强，更容易应用于新的领域或输入。

2.4.3 **设计思路**

由于不需提取 region proposal，则 YOLOv1 的检测流程为：



Resize image：将输入图片 resize 到 448x448。Run ConvNet：使用 CNN 提取特征，FC 层输出分类和回归结果。Non-max Suppression：非极大值抑制筛选出最终的结果。没有 region proposal，那定位包含目标的区域并固定输出 YOLOv1 的做法是：CNN 网络将 resize 后的图像分割成 S x S（7 x 7）的单元格，若目标的中心点落在某一单元格，则该单元格负责检测该目标，输出该目标的类别和边框坐标。

2.4.4 **网格结构**

YOLOv1 的网络结构借鉴了 GooLeNet 设计，共包含 24 个卷积层，2 个全链接层（前 20 层中用 1×1 reduction layers 紧跟 3×3 convolutional layers 取代 GooLeNet 的 inception modules）。

1. 输入：448 x 448 x 3，由于网络的最后需要接入两个全连接层，全连接层需要固定尺寸的输入，故需要将输入 resize。

2.Conv + FC：主要使用 1x1 卷积来做 channle reduction，然后紧跟 3x3 卷积。对于卷积层和全连接层，采用 Leaky ReLU 激活函数:，但是最后一层采用线性激活函数。

3. 输出：最后一个 FC 层得到一个 1470 x 1 的输出，将这个输出 reshap 一下，得到 7 x 7 x 30 的一个 tensor，即最终每个单元格都有一个 30 维的输出，代表预测结果。

输入图像被划分为 7 x 7 的单元格（grid），输出 tensor 中的 7 x 7 对应着输入图像的 7 x 7 个单元格，每个单元格对应输出 30 维的向量。如上图所示，输入图像左上角的网格对应到输出张量中左上角的向量。要注意的是，并不是说仅仅网格内的信息被映射到一个 30 维向量。经过神经网络对输入图像信息的提取和变换，网格周边的信息也会被识别和整理，最后编码到那个 30 维向量中。

总结输入图片被分成 7 x 7 个单元格，每个单元格预测输出 2 个 bounding box，每个 bounding box 包含 5 个值（4 个坐标 +1 置信度），另外每个单元格预测 20 个类别，所以最终预测输出 7 x 7 (4 + 1 + 20) = 7 x 7 x 30 的 tensor。

2.4.5 **训练**

首先在 ImageNet 上对网络中的前 20 层进行预训练，之后再在这 20 层后连上 4 层卷积和 2 层全连接层进行训练。所以，前 20 层是用预训练网络初始化，最后的这 6 层是随机初始化的并在训练过程中更新权重。此外，因为 detection 需要更多图片细节的信息，所以在训练时，统一将输入图片的 size 从 224 * 244 调整为 448 * 448。对于 loss 函数，是通过 ground truth 和输出之间的 sum-squared error 进行计算的，所以相当于把分类问题也当成回归问题来计算 loss。

2.4.6 **测试**

将一张图输入到网络中，然后得到一个 7*7*30 的预测结果。然后将计算结果中的每个单元格预测的类别信息和每个 bbox 的置信度信息相乘即可得到每个 bbox 的 class-specific confidence score。

根据同样的方法可以计算得到 7 x 7 x 2 = 98 个 bbox 的 confidence score，然后根据 confidence score 对预测得到的 98 个 bbox 进行非极大值抑制，得到最终的检测结果。

2.4.7 YOLOv1 缺点

因为 YOLO 中每个 cell 只预测两个 bbox 和一个类别，这就限制了能预测重叠或邻近物体的数量，比如说两个物体的中心点都落在这个 cell 中，但是这个 cell 只能预测一个类别。

此外，不像 Faster R-CNN 一样预测 offset，YOLO 是直接预测 bbox 的位置的，这就增加了训练的难度。

YOLO 是根据训练数据来预测 bbox 的，但是当测试数据中的物体出现了训练数据中的物体没有的长宽比时，YOLO 的泛化能力低。

同时经过多次下采样，使得最终得到的 feature 的分辨率比较低，就是得到 coarse feature，这可能会影响到物体的定位。

损失函数的设计存在缺陷，使得物体的定位误差有点儿大，尤其在不同尺寸大小的物体的处理上还有待加强。

3 思考与理解

3.1 联系与区别

联系：

都是基于卷积神经网络（CNN）的目标检测算法，可以提高检测的准确率和速度。

都是两阶段检测算法，即先生成候选框，再对候选框进行分类和回归，从而得到最终的检测结果。

区别：

R-CNN 是一种基于区域的检测算法，先生成一系列候选框，然后对每个候选框进行卷积特征提取和分类，最后进行回归来修正候选框的位置。而 YOLOv1 是一种全卷积的检测算法，将整张图片分成多个网格，同时预测每个网格中是否有物体以及物体的位置。

YOLOv1 相比 R-CNN 具有更快的检测速度，因为它是一次性对整张图片进行检测，而 R-CNN 需要对每个候选框进行处理，速度相对较慢。

3、R-CNN 在精度上相对较高，因为它对每个候选框进行了深入的卷积特征

提取和分类，可以更准确地判断是否为目标物体。而 YOLOv1 虽然速度快，但是在小目标检测和物体定位上相对较差。

3.2 尚未解决的问题

多目标检测：目前大部分算法都只能检测到单个目标，而在实际应用中，需要检测到多个目标，因此需要解决多目标检测的问题。

小目标检测：在一些应用场景中，目标往往很小，比如红外图像中的人脸识别，因此需要解决小目标检测的问题。

复杂场景下的检测：在一些复杂场景下，目标会被遮挡、变形、模糊等，因此需要解决复杂场景下的检测问题。

实时检测：在一些应用场景中，需要实时检测目标，比如自动驾驶中的行人检测，因此需要解决实时检测问题。

低光照条件下的检测：在一些低光照条件下，目标往往难以被检测到，因此需要解决低光照条件下的检测问题。

3.3 未来研究趋势

深度学习算法的进一步发展：随着深度学习在目标检测领域的应用越来越广泛，未来将进一步加强对深度学习算法的优化，提高检测算法的准确度和鲁棒性。

大规模数据集的构建和应用：大规模数据集对于目标检测算法的训练和测试至关重要，未来研究将继续关注如何构建更多、更丰富、更真实的数据集，并应用于实际的场景中。

多模态融合的研究：随着多种传感器和数据源的普及，未来将进一步探索如何将多模态数据融合到目标检测算法中，提高检测算法的准确度和鲁棒性。

目标跟踪和目标识别的技术融合：目标检测、跟踪和识别是目标感知的核心技术，未来将进一步探索如何将这些技术融合到一起，提高目标感知的全面性和准确度。

应用场景的拓展：未来将继续关注目标检测技术在各种应用场景中的应用，如智能交通、安防监控、医疗诊断等，为人类社会提供更多的智能化服务。

参考文献:

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik.Rich feature hierarchies for accurate object detection and semantic segmentation-v5.

[2] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[3] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[4] Joseph Redmon,Santosh Divvala,Ross Girshick,Ali Farhadi.You Only Look Once: Unified, Real-Time Object Detection on Computer Vision and Pattern Recognition.