物体检测中的特征构建与模型优化

文 / 孔涛

摘　要　本文针对物体检测中的环境变化多样、物体尺度变化不一、搜索空间巨大等挑战性问题，围绕特征构建、模型优化和应用等方面进行研究。针对物体检测中的多尺度特征融合问题，提出针对物体检测的神经网络特征融合方法 HyperNet；进一步提出了逆向连接的特征金字塔物体检测方法，将不同尺度的物体分配不同层次的特征，该方法大大减少了多尺度物体检测的难度；提出了无需候选窗的物体检测模型 FoveaBox，摒弃了传统依赖候选窗扫描的过程。本文提出的系列方法已经在检测、分割、姿态估计等方面成功得到拓展。

关键词　跨层连接；无需候选窗；物体检测

0 引言

物体检测是一种使计算机能在图像中自动发现感兴趣的物体，并判断物体的类别、位置的技术。具体来讲，物体检测解决的问题是物体是什么，以及在什么位置这两个基本问题。传统的物体检测方法以滑动窗口法为代表，给定一幅输入图像，算法首先对图像进行扫描得到一系列的子窗口；然后在这些子窗口内部提取相应的特征，针对这些特征进行计算和识别，得到子窗口的识别结果。最新的物体检测模型取代了基于传统人工设计特征的候选框生成算法。在 Faster R-CNN 中，研究者提出了锚点（Anchor）的概念，利用与物体检测模型共享的特征直接在卷积神经网络的特征层上生成候选框。自此之后，基于卷积神经网络的物体检测方法可以大致分为两种，一种是基于候选区域的两阶段及多阶段的方法；第二种是单阶段方法，这种方法直接基于锚点进行预测，判断其所属的类别并进一步调整锚点的位置。

从提取图像特征的演进角度，物体检测的发展大致经历了三个阶段，第一个阶段是基本的图像元素作为图像描述的阶段；第二个阶段是特征描述子的阶段；第三个阶段是将卷积神经网络作为特征和框架的阶段。物体检测的难度主要体现在物体形状和尺度的多样性、物体自身的多样性、搜索空间问题和正负样本不均衡问题这四个方面。

本文围绕通用物体检测的特征构建、搜索空间划分、模型框架及应用等基本问题，提出了一系列的解决方案。下面分别进行介绍。

1 基于跨层连接的物体检测模型 HyperNet

自 2014 年 Ross 等提出基于目标区域候选的 R-CNN 物体检测模型之后，该领域得到了快速发展。然而，在本章提出的跨层连接的融合技术出现之前，主流的物体检测模型还是依赖单一的卷积特征层进行预测。利用单一的特征层虽然也能得到不错的物体检测效果，但对多尺度（尤其

是小目标）的检测能力有限。利用多尺度、多层次的特征能有效提高模型对多尺度物体的检测能力。

HyperNet 遵从了两阶段法，首先生成候选区域建议，之后进行物体检测的流程。图 1 是详细的模型示意图。基于跨层连接的物体检测模型由以下几个部分组成：给定一张图像，首先利用卷积神经网络模型计算该图像在不同网络层次的特征；然后将这些不同层次、不同语义程度的特征融合起来；之后利用一个轻量级的目标候选生成网络去预测少量的候选框；最后对这些候选框进行分类和微调，输出最终的物体检测结果。

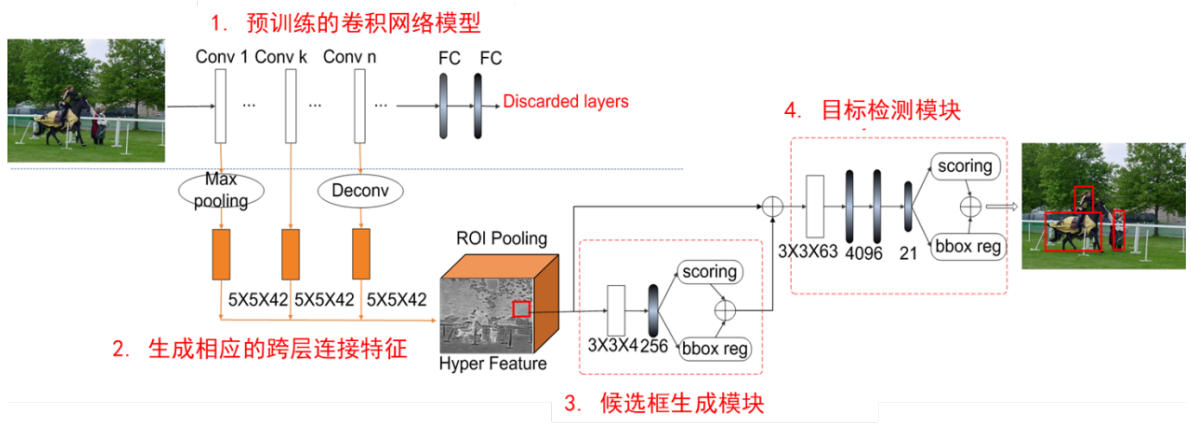

图 1 HyperNet 示意图

• Hyper Feature 生成：给定一幅二维图像，首先利用卷积神经网络计算整幅图像在不同层次的特征。为了将不同层次的特征结合起来，需要将不同层次的特征进行归一化。不同层之间的特征数值可能存在尺度不一的问题，为此，我们利用局部响应归一化技术来归一化这些特征，最后将不同层特征拼接成一个统一的特征，称之为 Hyper Feature。

• 候选区域建议模块：候选框生成模块包含一个感兴趣区域提取层（ROI-Pooling）、一个卷积层和一个全连接层，后边便是并行的前背景分类和候选框回归。对于每张图片，神经网络首先在图像上生成大约 3 万个具有不同大小和长宽比的候选窗口。经过非极大抑制之后，我们选择前 200 个候选建议框进入后边的物体检测模块。

• 物体检测模块：首先将感兴趣区域提取层得到的 Hyper 特征进行压缩，之后接两个全连接层，最后进行网络的输出。目标检测模块包含两个并行的输出——对每个区域建议框，需要进行不同类别的分类；针对每个类别输出进行候选框位置的进一步微调。

小结：本章提出了跨层连接的概念。考虑到卷积神经网络的高层语义与低层物体细节，并通过简单、有效的融合方式形成全新的特征，在此基础之上进行物体的检测工作。该方法可以同时生成一张图片中感兴趣物体的区域候选框，以及最终的物体检测结果。HyperNet 是在基于深度卷积神经网络物体检测模型中，对多层融合的首次成功的尝试。该方法试图从特征的角度上去解决多尺度的问题，但也存在一定的局限性，虽然考虑了多尺度、多层次的特征，最终仍然在单一尺度的特征上检测所有尺度的物体。

更多详细内容参见文章：

• Tao Kong, Anbang Yao, Yurong Chen, Fuchun Sun. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection.

CVPR 2016.

2 基于逆向连接的特征金字塔物体检测

上一章提出了基于跨层连接的物体检测模型 HyperNet，该模型利用多层特征融合的思想有效提升了物体检测模型对多尺度目标定位的能力，本章进一步提出了基于逆向连接的特征金字塔物体检测框架 RON。该框架结合了跨层连接的思想，同时利用卷积神经网络的多层输出最终的检测结果，可以进一步解决多尺度目标的定位问题。

本章提出的基于逆向连接的特征金字塔模型的主要思想是利用卷积神经网络不同深度、不同分辨率的特征层检测不同尺度的物体，通过将不同尺度的物体分配到不同的特征层，一方面可以大大降低每一个层次检测目标的难度，另一方面将多层检测的结果进行结合可以有效提高检测的精度。如图 2 所示，给定一幅输入图像，首先计算该图像在卷积神经网络的特征；然后利用逆向连接构造不同尺度、不同层次的特征；之后利用不同层次的特征检测不同尺度的物体。

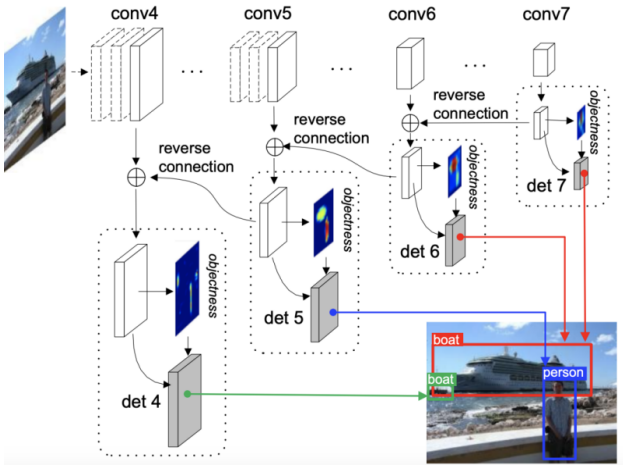

图 2 基于逆向连接（Reverse Connection）的多尺度物体检测模型

• 参考框设计：在生成多尺度特征的基础之上，在特征图的每个位置均匀预定义一系列的候选框，候选框的大小与物体的感受野相对应。在 RON 的结构中，不同尺度物体的检测被分配到了相应的特征层上。

• 感兴趣区域先验响应：提出了感兴趣区域响应图来引导目标的搜索过程。

• 目标响应图与物体检测的结合：在训练的过程中，模型将会对每个包围框的不同类别的评分结合目标响应图进行更新。在训练的每一个批次中，神经网络模型将会同时计算感兴趣区域先验响应图和后边的物体检测模块。在反向传播的检测模块，仅有响应的部分会参与到模型的更新训练中，如图 3 所示。

• 模型测试：在测试阶段直接将物体检测得分与感兴趣区域响应相乘得到最终的物体检测结果。最终的物体检测得分将感兴趣区域响应和分类的结果进行融合。最后采用非极大抑制算法来去掉冗余的检测框。

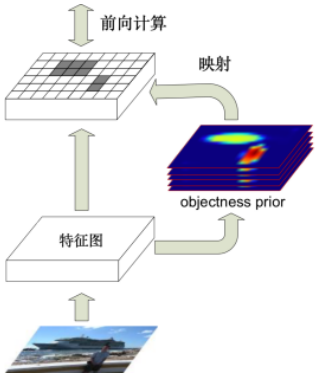

图 3 利用目标区域响应图来减少物体检测部分的搜索空间示例图

小结：本章提出了一种利用逆向连接技术构造图像的深度金字塔特征，进而利用不同层次、不同分辨率的特征图检测不同尺度物体的方法。基于逆向连接的特征金字塔构造技术可以有效地提升低层特征的语义表达能力，从而帮助小物体的检测。本方法不仅可以生成最终的物体检测结果，而且可生成候选区域建议。值得一提的是，与本方法同一时期的工作也包括来自 Facebook 人

工智能研究院的 FPN 和 Google 的 TDM，通过深度特征金字塔构造技术，可以形成对输入图像不同尺度、不同层次的特征描述来帮助视觉处理任务。

更多详细内容参见文章：

• Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Yurong Chen, Ming Lu. RON: Reverse Connection with Objectness Prior Networks for Object Detection. CVPR 2017.

3 深度特征金字塔重组技术

在理想情况下，一组成功的金字塔特征需要具备以下属性：充分利用卷积神经网络中多层特征的表达能力；对每一层次特征的表达能力都有一定的增强。基于逆向连接技术的特征金字塔构造技术在一定程度上满足了这两个条件。我们通过对 FPN 这种典型的特征金字塔构造过程进行分析（见图 4），发现其构造过程实际上等价于将不同层次的特征进行线性组合。当然，线性组合在一定程度上可以提升不同层的表达能力，但其无法捕捉更高维的依赖关系。下面将从理论上对目前基于深度卷积神经网络的特征金字塔构造过程进行分析，并给出特征金字塔重组技术的介绍。

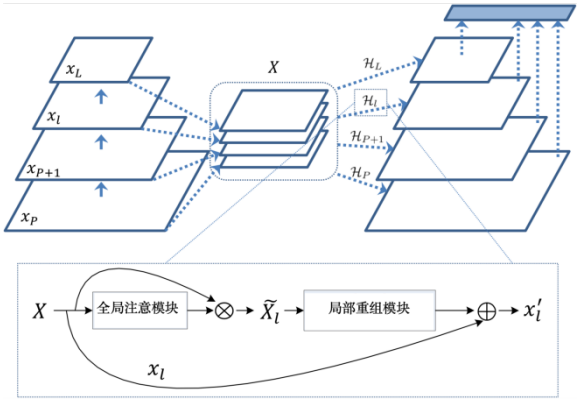

图 4 特征金字塔重组技术示意图

给定一幅图像卷积神经网络特征的层次结构 X，理想的物体检测模型重要的因素是为检测不同尺度的物体提供合适的特征。本文对第 l 层特征生成过程表示为输入层次结构的非线性变换 $x_l'=H_l(X)$。本文将这种特征变化过程 H_l 表示为全局注意和局部重组问题。两个模块都由一个轻量的神经网络组成，因此均可以直接嵌入到现有的网络结构中进行端到端训练。因为全局注意力模块和局部重组模块从不同的角度来处理特征的层次结构，因此具有一定的互补性。

• 全局注意模块：其主要作用是通过分析特征层次的全局结构，为不同尺度的物体选择更有用的特征。本文采用缩放激励结构（SE）作为基本的模块。

• 局部重组模块：局部重组模块将输入特征的层次结构映射到输出层，同时保证不同位置共享相同的参数。受残差结构启发，本文设计了一种基于残差学习的轻量结构来完成局部重组任务。

小结：在构造金字塔特征的过程中，一个重要的因素是如何充分挖掘卷积神经网络不同层次之间的互补性。我们提出了基于全局注意和局部重组的技术来解决这个问题。全局注意和局部重组模块显式地为不同尺度的物体选择其对应的特征。与直接利用卷积神经网络多层信息，以及逆向连接的技术相比，全局注意和局部重组可以有效地提升模型对多尺度物体的检测能力和泛化性能。

更多详细内容参见文章：

• Tao Kong, Fuchun Sun, Chuanqi Tan, Huaping Liu, Wenbing Huang. Deep Feature Pyramid Reconfiguration for Object Detection. ECCV 2018.

• Fuchun Sun, Tao Kong, Wenbing Huang, Chuanqi Tan, Bin Fang, Huaping Liu. Feature Pyramid Reconfiguration with Consistent Loss for Object Detection. TIP 2019.

4 无需候选框的物体检测方法

物体检测包含了识别和定位两个子任务。给

定一幅图像，物体检测模型需要一方面判断该图像中是否包含我们感兴趣的物体。如果包含，进一步返回这些物体所在的位置。为了给物体检测模型物体定位的能力，滑动窗口技术成为主流的、标准的解决方案。即使是在目前主流的基于深度学习的物体检测模型中，滑动窗口技术也扮演着重要的角色。本章提出的物体检测框架 FoveaBox 彻底摒弃了滑动窗口，直接利用卷积神经网络特征来完成物体的检测过程。

考虑人类是如何来进行物体检测这一任务的：当我们进入到一个场景中会自然而然地关注到物体的大体位置。如果想得到该物体边界，仅需对物体部分进行左右上下浏览就可以。在这个过程中，不需要任何的扫描窗口，也不需要预定义物体的大小和长宽比。受这一过程的启发，我们不自觉会问这样的问题——基于预定义窗口的方式是最优的物体检测过程吗？如果不是，是否可以设计出一种全新的、不需要候选框的物体检测框架？

FoveaBox 受到人类视觉系统感知物体的启发：人类的视网膜中有一个中心凹（Fovea）区域，该区域对物体的细节极为敏感。通过中心凹区域采集更丰富的信息更有利于检测物体的精确位置（见图 5 ）。具体来讲 FoveaBox 由一个骨干网络和两个子网络组成。骨干网络用于计算整幅图像的卷积神经网络特征，可以用现成的网络结构（比如 ResNet）。第一个子网络用于对网络输出位置的分类操作，第二个子网络对每个位置潜在的目标进行包围框预测。

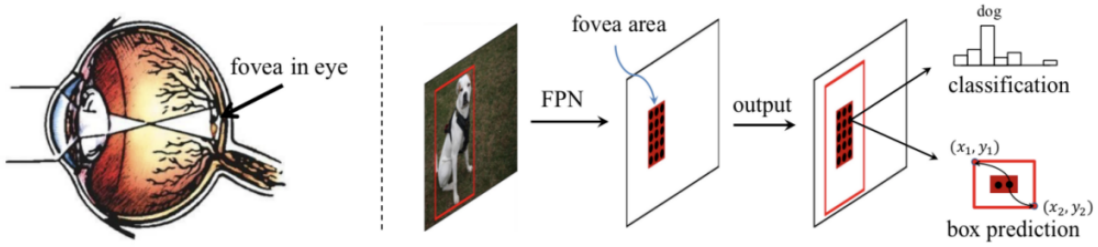

图 5 FoveaBox 物体检测模型

• 深层特征金字塔表示：采用标准的特征金字塔结构 FPN 来作为基本的网络。概括来讲，给定一幅图像，FPN 用多个逆向连接的结构来构造最终的金字塔特征。特征金字塔的每一层用来检测相应尺度的物体。

• 物体尺度的分配：将目标物体的尺度划分为几种，将不同尺度的物体投影到对应的金字塔特征层上进行预测。

• 包围框预测：为了确定准确的物体位置，需要预测每个物体的边界。对于每个正确的标注，网络学习从当前的坐标位置到真实标注的映射。当该分支被训练完成后，就可以针对每个输出位置给出相应的包围框。将包围框与检测得分结合就是该位置最终的检测结果。

图 6 给出了 FoveaBox 的部分检测结果。对于每一组的左边，是最终的包围框、类别和相应的得分；右边是 FoveaBox 在非极大抑制之前的结果，得分的高低由颜色的深浅来表示。对于每个物体的位置，大约有几个物体的中心位置点激活，而这些激活的点都可以预测出同一个物体边界。

图 6 FoveaBox 的检测结果示例

通过该图可以看出 FoveaBox 可以生成鲁棒的最终检测结果，而不依赖于候选框或者锚点技术。

小结：本章提出了无需候选框扫描、无需锚点预定义的方法 FoveaBox。该方法受到人类的视觉系统检测目标过程的启发，首先初步确定物体中心点的位置，然后基于中心点位置预测左右上下边界。FoveaBox 可以直接嵌入到目前主流的特征金字塔网络结构中，利用金字塔特征的多层检测对应尺度的物体。在标准的物体检测数据集中，FoveaBox 表现出优异的物体检测性能。值得相信，这种定位的思路可以拓展到三维物体定位、视频中的物体检测、行为分析等相关领域中。

更多详细内容请参考以下参见文章：

• Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, Jianbo Shi. FoveaBox: Beyond Anchor-based Object Detector. TIP 2020.

5 结束语

物体检测是计算机视觉领域基本的任务之一，其包含识别和定位两个重要的子任务。物体检测模型的好坏受到多种因素影响。在这些影响因素中，特征、搜索空间和框架起到了至关重要的作用。近 6 年物体检测领域飞速发展。本文对物体检测中的特征构造、搜索空间和框架方面进行了深入研究。未来物体检测一定是朝着更准确、高速的方向发展，如更完善的特征金字塔构造过程、无需候选框的物体检测和分割方法，以及向弱监督、无监督物体检测。

作者介绍

孔涛

清华大学博士毕业，师从孙富春教授；现为字节跳动人工智能实验室研究员。主要研究方向为计算机视觉。在 TIP、CVPR 等国际权威期刊会议发表论文 10 余篇，授权美国发明专利 2 项。曾获 IBM 中国奖学金、IROS 国际机器人抓取操作竞赛冠军等。