计算机视觉中的目标检测

【摘要】 本文主要介绍了目标检测的发展过程及其在国内外的前沿技术。其中图片检测主要有R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, YOLO等成果，视频中的目标检测主要有I3D，D&T，RDN等模型。图片检测已经取得了很好的发展，检测准确度已经非常高了，但是视频检测还有很多发展空间，这是因为在视频检测中，会有运动模糊，相机散焦等问题，而且相应的提出了动作检测等需求。

【关键词】 目标检测；图片检测；视频检测

Object Detection In Computer Science

【**Abstract**】 This paper mainly introduces the development process of object detection and its frontier technology. The object detection algorithm in image mainly includes R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, YOLO, etc. The object detection algorithm in video mainly includes I3D, D&T, RDN and other models. While the image object detection has contain a lots of success, high accurate and fast, video object detection is faced of lots of difficulty, such as motion blur, rare pose, camera defocus, etc.

【**Key Words**】 object detection; image object detection; video object detection

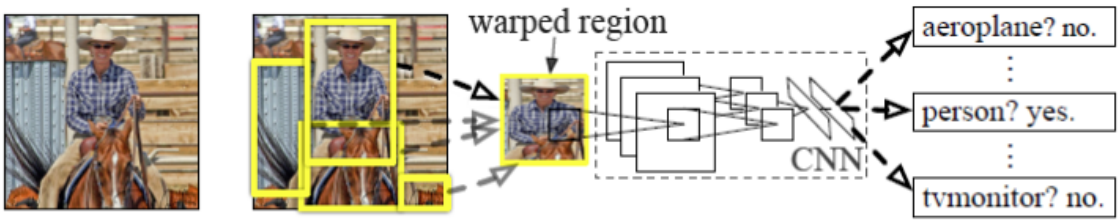
第 1 章 图片中的目标检测

1.1 R-CNN 系列

R-CNN 初次将深度神经网络引入目标检测，并且取得了超过传统机器学习算法的准确率，在后续的改进中逐步实现了高准确度，端到端，实时性，个体检测等优势。

1.1.1 R-CNN

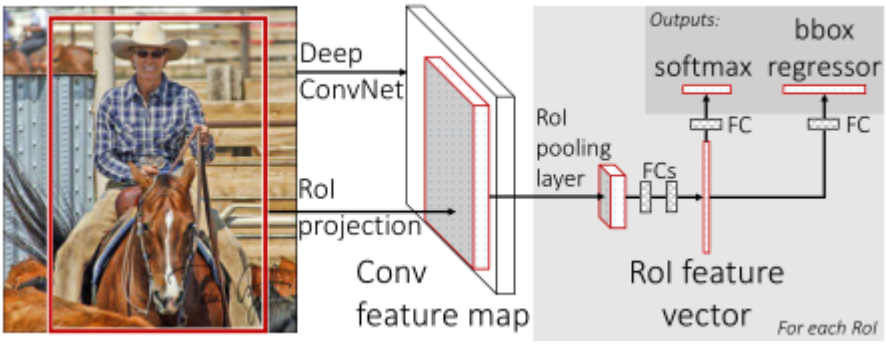
R-CNN 主要分为以下四个步骤：候选区域生成, 特征提取, 类别判断, 位置精修。不同于传统方法，R-CNN 采用 Selective Search 的方法生成候选区域，相对原来滑窗的方式效率高了不少但还是无法满足实时性。而后将图片放缩到同一尺寸送入 CNN 中提取图像的特征。CNN 采用迁移学习的方式，直接 Alexnet 的参数初始化，然后再 fine-tune 训练。因为该阶段分类效果还比较差所以调用 SVM 进行分类能将准确度提高好多，但牺牲了端到端的优势。



1.1.2 Fast-RCNN

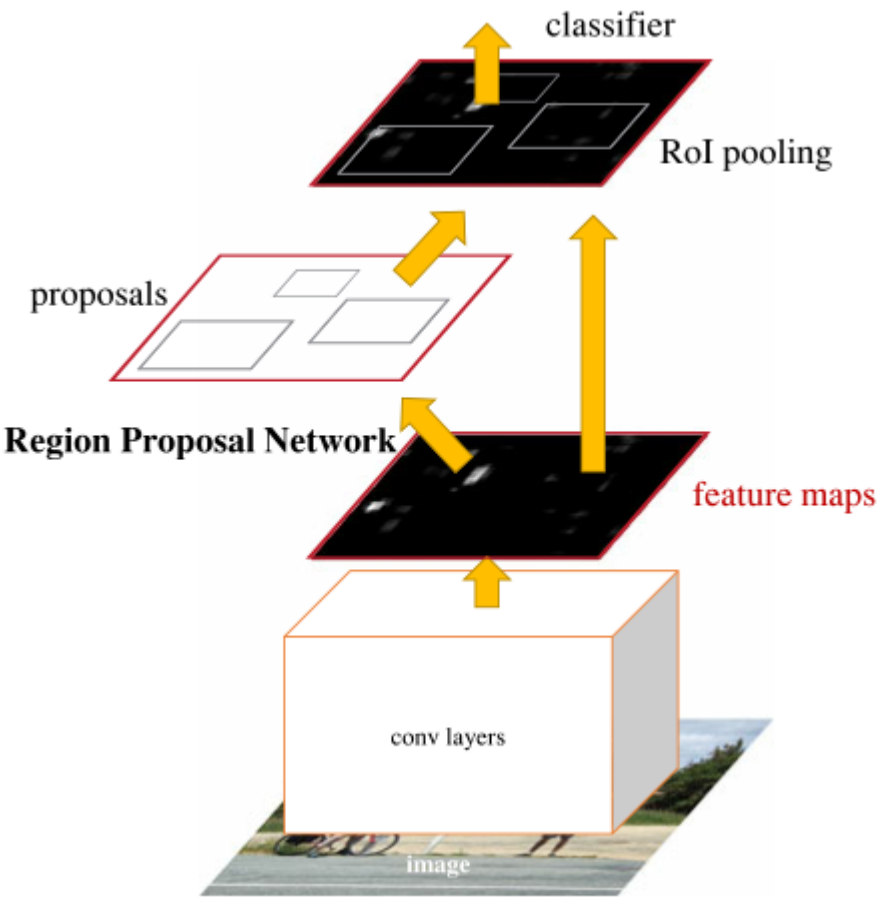
Fast-RCNN 借鉴了 SPPNet，设置了 ROI pooling 层，使得网络可以接受输入不同尺度的边框而不用标准化，与 SPPNet 的区别在于 Fast-RCNN 仅在单尺度下进行边框提取，且实验证明多尺度下的提取只能提高非常小的准确度但开销明显提升。Fast-RCNN 把矩形选框投影到 feature map 上，使得模型省去对不同的选框进行卷

积计算，因为这当中有非常多的重复计算。完成特征提取后，该模型分别进行了 softmax 求解分类和 bbox regressor 进行边框修正而不再采用 SVM，从而实现了端到端的模型，之所以不再采用 SVM 是因为在该模型中，准确率已经提高到一定高度，在这种情况下采用深度学习的方式去训练结果更好。



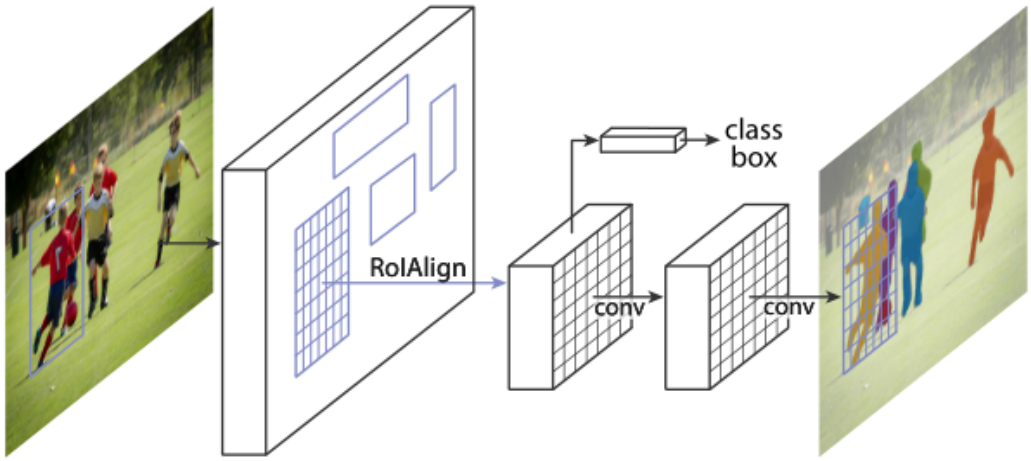
1.1.3 Faster-RCNN

Faster-RCNN 的主要贡献在于提出了 RPN 网络，将建议选框也融入到了 Fast-RCNN 之中。首先在 feature map 上滑动获得特征点，然后以该特征点映射回原图的中心点作为锚点，分别以三个不同尺度，三个不同高宽比组合共计九种方式提取选框然后对其进行非极大值抑制然后送回 rcnn 进行目标检测。作者的实验表明降低 proposal 的数量并不会降低模型的精度且在 RPN 中 reg 层的影响远小于 cls 层的影响。该模型极大的缩减了预选框生成的时间，使得目标检测可以满足实时性的需求。



1.1.4 Mask-RCNN

Mask-RCNN 在 Faster-RCNN 的架构上加入了实体分割的模块，实体分割经常涉及像素层面的边界，对位置信息非常敏感，所以原 Faster-RCNN 的 ROIPool 层对该任务来说太粗糙，遂引入了 ROIAlign，使用双线性插值使得，坐标点定位变得更加精确。该模型还有一个不同之处在于先分类，后按照类别优化 mask 层，减弱了类别之间的竞争。



1.2 YOLO 系列

YOLO 核心思想：从 R-CNN 到 Fast R-CNN 一直采用的思路是 proposal+分类（proposal 提供位置信息， 分类提供类别信息）精度已经很高，但是速度还不行。YOLO 提供了另一种更为直接的思路： 直接在输出层回归 bounding box 的位置和 bounding box 所属的类别(整张图作为网络的输入，把 Object Detection 的问题转化成一个 Regression 问题)。

1.2.1 YOLOv1

该模型首先用 CNN 提取图片的特征，然后将图片划分成 7x7 共 49 个模块，对每一模块的中心点取不同高宽比的两个预选框以他的类别信息，边框信息，及分类信息编码作为他的特征向量，然后综合各个特征的损失来训练模型。测试时滤掉得分低的特征向量，NMS 余下的向量得到最终结果。

1.2.2 YOLOv2

v2 主要是对 v1 版本的许多缺陷作了改进。第一，v1 版本中使用全连接层进行 bounding box 预测，忽略了空间位置信息，改用卷积层替代了全连接层。第二，采用了类似 Faster-RCNN 中 anchor 的思想从提取两个框改进到 9 种框。同时改为以 GoogleNet 作为基准模型。约束了位置预测的范围，使得模型更快收敛，精度也

得到提高，以及其他非常多的 tricks。

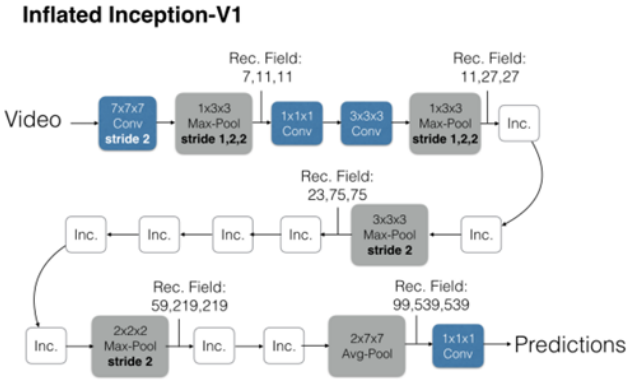
1.2.3 YOLOv3

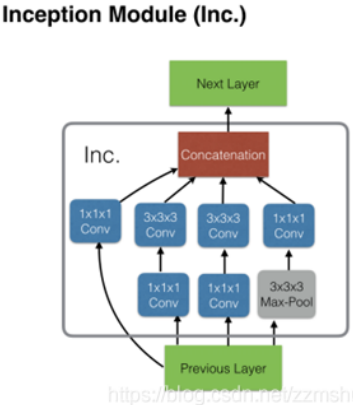
YOLOv3 采用了更深的 DarkNet 作为基准模型，以及对多尺度空间进行了选框提取，将尝试预测框数量提升了 10 多倍，将精度提高了不少。用 logistic 替代 softmax 适应了多标签对象的预测。YOLOv3 在精度提高不少的前提下，速度仍是其他模型的 3，4 倍。

第 2 章 视频中的目标检测

2.1 I3D 模型

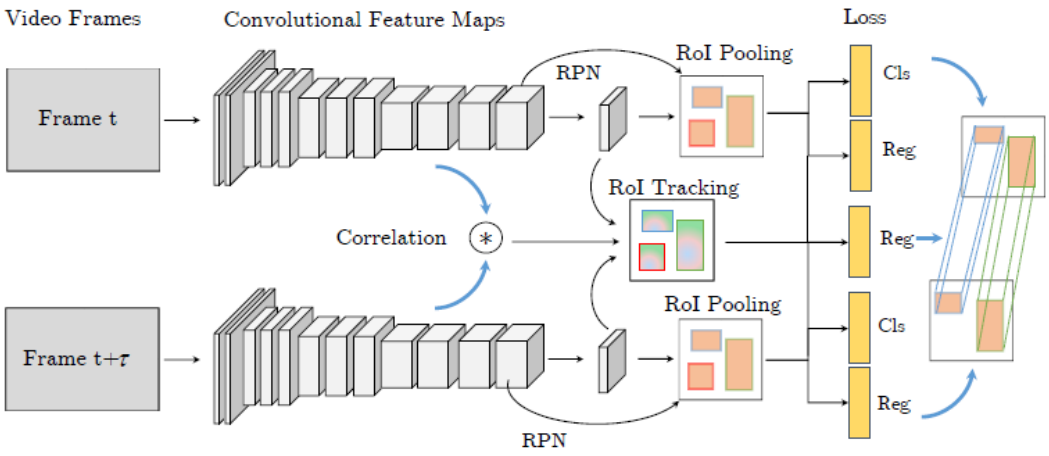
I3D（Two-Stream Inflated 3D ConvNets）模型是结合 3D-ConvNet 和 Two-Stream 模型，将 2DCNN 迁移学习到 3DCNN。其中 3DCNN 的初始化通过将 ImageNet 的每一张图片视为每一帧都是这张的图片的 boring video，把 ImageNet 的参数在时间上平均后用于初始化 3DCNN。经过实验发现，简单的将 NxN 卷积核扩充到 NxNxN 并不合适，要根据视频的帧率调整在时间方向上的维度大小。





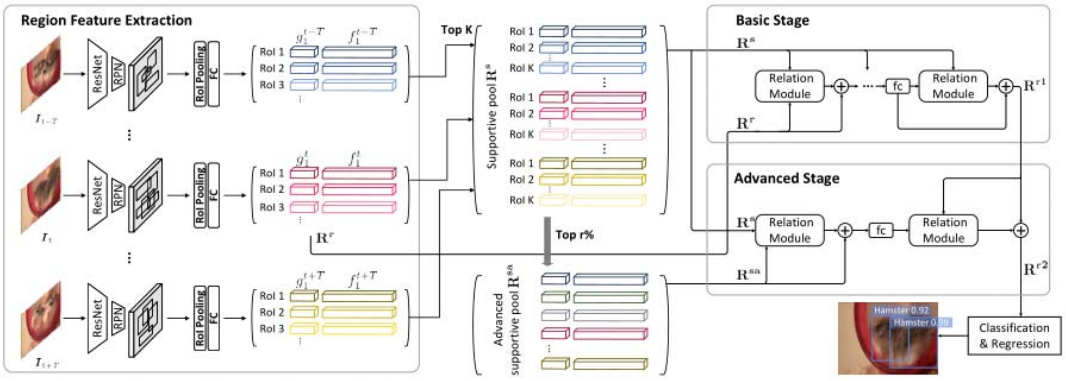
2.2 D&T（Detect and Track）

I3D 模型非常的繁重，需要大量的计算资源和时间，而 D&T 主要解决的就是在开销方面的问题。该模型提出了 Linking tracklet 的信息抽取方法，通过计算最优路径来选取最具代表性的几个帧来替代视频，极大的简化了视频体量。该模型采用 2 个 FCN 来对前后两个帧进行信息提取，各采用一个 RoIpooling 来分类和框回归，然后再利用一个 RoI-tracking 层来回归框的变化。



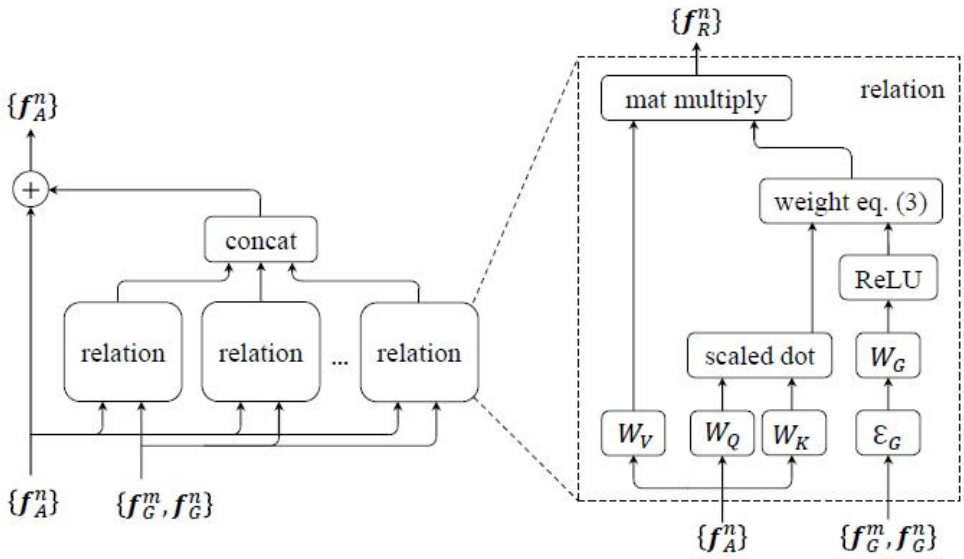
2.3 RDN

RDN 沿用 D&T 中提取帧的做法，将某一时刻的图片与前序关键帧和后续关键帧一同送入模型。RDN 分为两个部分，特征提取部分沿用 ResNet-101 模型提取特征，右半部分利用 attention 机制进行分类和框回归。由于 attention 会引入很大的计算量，所以首先分别提取三张关键帧中最重要的 K 个帧，然后利用 Relation Module 计算 R^s 和 R^r 的相关性，加强相关性强的特征，重复多次继续加强然后输出 R^{r1}。然后在将 R^{r1} 与 R^s 的 Top r%部分 R^{sa} 再次进行精馏，最终输出 R^{r2}。RDN 最主要的贡献在于把 attention 机制引入了视频中的目标检测，做到了把位置信息纳入考虑，使得位置信息也可以影响对对象的行为的判断，同时将位置相关度作为权重赋予 box linking 提高了准确率。



其中的Relation Module是基于[9]的模块再加入位置信息完成，Relation Module 是利用三个神经元 W_v，W_q，W_k 分别代表 value，query，key，首先对自身做一个 attention，然后与参考特征和 W_g 的点积进行点积输出一个权重给 W_v。通过这一方

式可以合理的分配权重给前后文对该点的影响。



2.4 Seq-Bbox matching

该方法主要解决部分帧在运动的过程中引起的动态模糊而导致的帧无法识别或是识别错误。第一，该方法提出将当前帧与前后时刻距离最近的对象框关联在一起，一起分类，平均得分，以纠正个别帧的错误。第二，对于没有检测出对象的帧，该方法设置了一个可合并间隔，若是两个集合的距离小于这个间隔那么就合并他们，并通过双线性插值获得中间没有检测出来的对象，将前后两帧的得分的平均值赋予他们。

结论

在目标检测的任务中，我们首先关注的是如何准确的返回目标的边框，在此基础上我们又降低了开销，赋予了实时性，开发了更多的功能，如个体检测，动作识别等。在该领域中，我们经常用到线性插值的方法来提升准确度，用卷积的方法获得更高层次的语义减少计算量，虽然两者有的时候不可兼得，但是目前方法仍不够完善，可以通过不断优化来同时提高两者性能。以上所有工作都离不开深度学习的应用与优化，深度学习的不可解释性使得模型很容易建立有许多方向可以优化，但相应的也缺乏确切的理论指导和严谨的证明推导,很多几年前好用的结论及方法在今天就已经被淘汰。在模型的建立中，迁移学习是非常重要的方法，多数模型都是采用先前比较完善且已完成训练的模型，再加入合适的特征提取模型，以及优化模型的复杂度，减少时间开销的方式下提出来的。现阶段的深度学习是由数据驱动的，或许可以朝着多种数据信息互相结合的方向，像是视频信息可以和配音，字幕，弹幕等信息结合，或者是把小说翻拍的电影与小说相结合，既能用视觉理解文字，又能用文字理解视觉，以多种感觉相结合的方式提高深度学习的泛化性能。

参考文献

[1] Shaoqing Ren, Kaiming He, Ross Girshick. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE,2017.

[2] Girshick, Fast R-CNN, ICCV, 2015.

[3] Christoph Feichtenhofer，Axel Pinz，Andrew Zisserman. Detect to Track and Track to Detect, ICCV,2017.

[4] J.G. Teng, Q.G. Xiao, T. Yu. Three-dimensional finite element analysis of reinforced concrete columns with FRP and/or steel confinement[J], Engineering Structures, 2015, 1(97): 15-28.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollar. Mask R-CNN, ICCV, 2017.

[6] João Carreira, Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR, 2017.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell. Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE, 2014.

[8] Jiajun Deng, Yingwei Pan, Ting Yao. Relation Distillation Networks for Video Object Detection, ICCV, 2019.

[9] H. Hu, J. Gu, Z. Zhang, J. Dai and Y. Wei, "Relation Networks for Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 3588-3597, doi: 10.1109/CVPR.2018.00378.