Visual Object Detection

视觉对象检测

智能系统实验室
清华大学基础工业训练中心

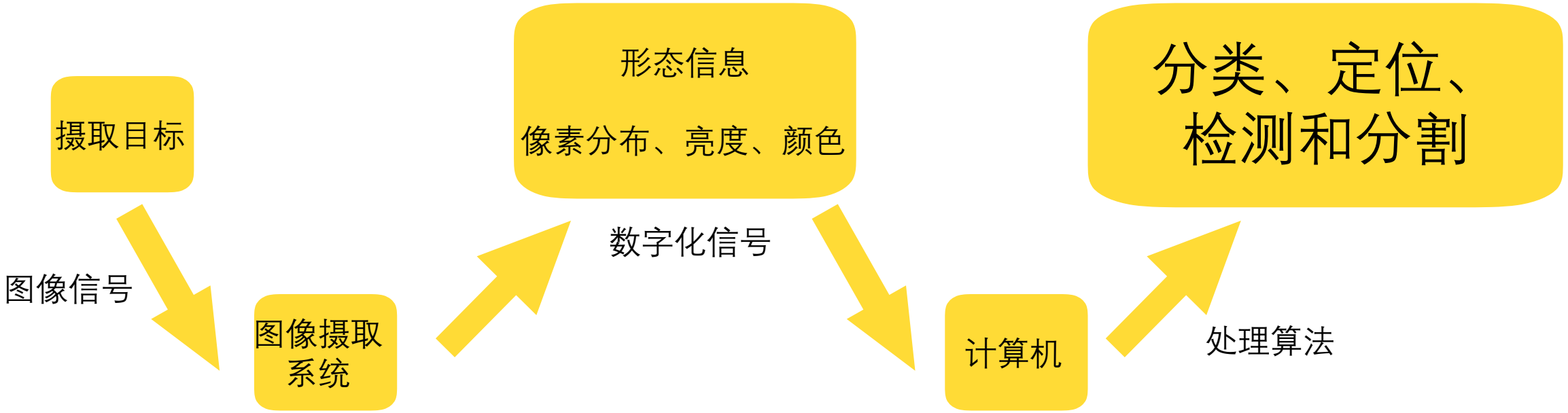
目录

- 计算机视觉的任务
- 计算机视觉的识别指标
- 视觉对象检测的方法
- 图像语义分割的方法

计算机视觉的任务

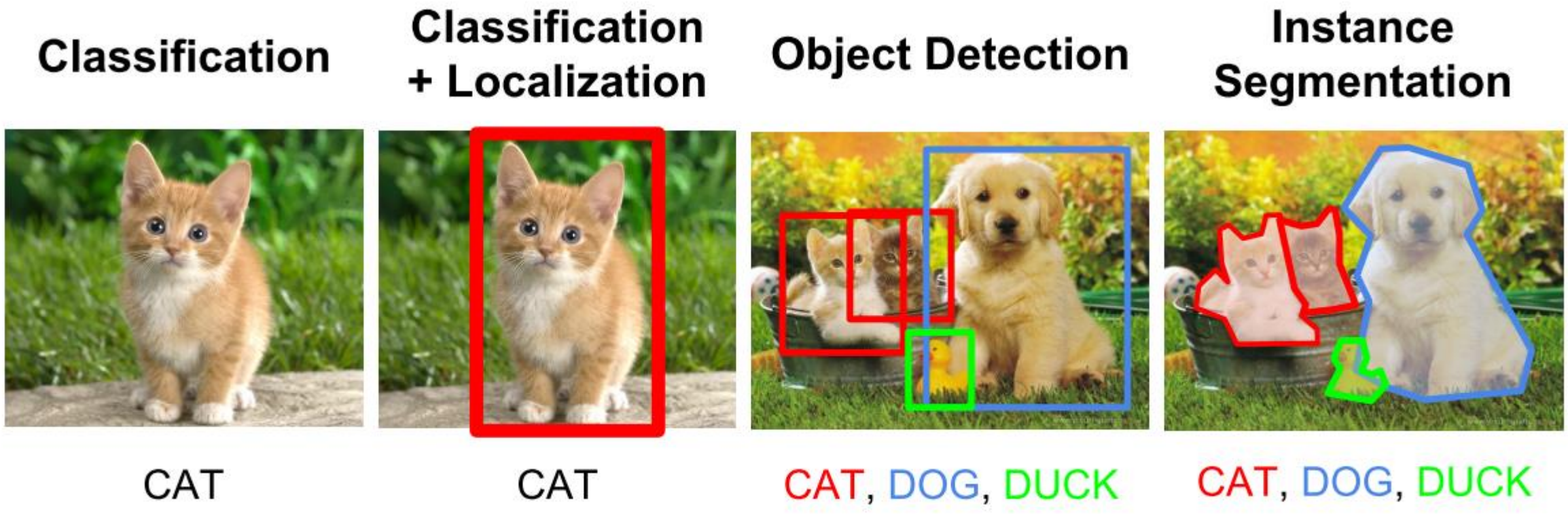
计算机视觉

- 计算机视觉就是用计算机代替人眼来做测量和判断（简单说来）。
- 计算机视觉是人工智能快速发展的一个分支。
- 计算机视觉的主要任务包括：分类、定位、检测和分割



分类、定位、检测、分割

Visual Object, 对象，又称物体，目标等



计算机视觉识别指标

OD Index

识别的指标

- 精确率（precision）是针对预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。预测（分类）为正有两种可能：
  - 一种是把**正类预测为正类(TP)**，
  - 另一种是把**负类预测为正类(FP)**
- 召回率（recall）是针对原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。预测（分类）为负有两种可能：
  - 一种是把**原来的负类预测成负类(TN)**，
  - 另一种是把**原来的正类预测为负类(FN)**
- 准确率(accuracy) 是指对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。（也就是损失函数是0-1损失时测试数据集上的准确率）
- 精确率(precision) = TP/(TP+FP)
- 召回率(recall) = TP/(TP+FN)
- 准确率(accuracy) = (TP+TN)/(TP+FN+FP+TN) = = 预测对的/所有

举例说明

- 例子:
  - 假设我们手上100张样本图片，有70个正样本（猫图片），30个负样本（狗图片），
  - 计算机视觉的任务要找出所有的正样本（猫图片），
  - 识别系统查找出50个（猫图片），其中只有40个是真正的正样本（猫图片）。
- 计算识别指标:
  - TP: 将正类预测为正类数 40
  - FN: 将正类预测为负类数 30
  - FP: 将负类预测为正类数 10
  - TN: 将负类预测为负类数 20
- 精确率(precision) = TP/(TP+FP) = 80%
- 召回率(recall) = TP/(TP+FN) = 4/7
- 准确率(accuracy) = 预测对的/所有 = (TP+TN)/(TP+FN+FP+TN) = 60%

对象检测的识别精确率指标

- 常用的识别精确率指标：
  - 平均精确率均值mAP
  - PR曲线的覆盖率AUC：P为精确率，R为召回率

平均精确率均值mAP（识别准确率指标之一）

- 平均精确率均值mAP（Mean Average Precision）是对象检测研究中常用数据集VOC 2007所采用的评价指标，被该领域的研究者们广泛使用
- VOC 2007对于mAP的数学定义如下，其中*p*和*r*分别表示模型在取不同的阈值参数时的精确率（Precision）和召回率（Recall）

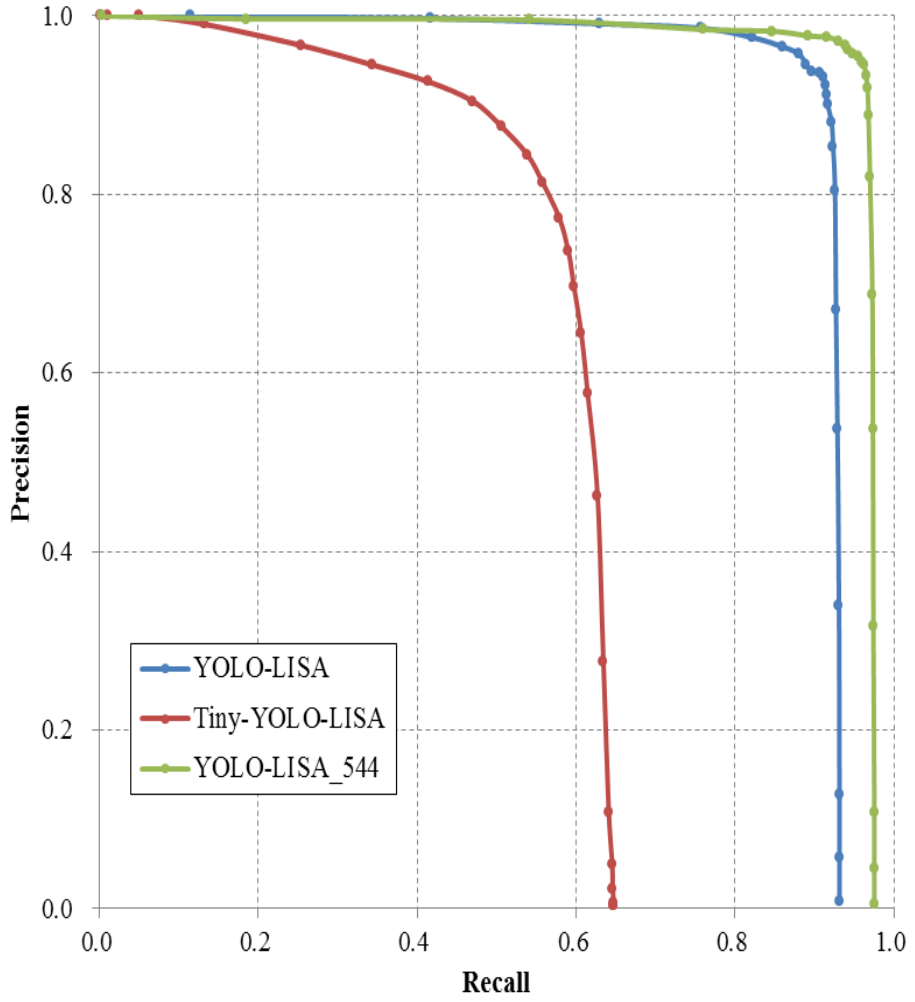
$$\mathrm{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \cdots, 1\}} \max_{\tilde{r}:\, \tilde{r} \geq r} p(\tilde{r})$$

$$\mathrm{mAP} = \frac{1}{\#\mathrm{classes}} \sum_{c \in \mathrm{classes}} \mathrm{AP}(c)$$

  - mAP指标度量模型在不同情况下的平均精确率，是对精确率和召回率之间平衡取舍问题的一种有效处理方式。
  - mAP越高，说明模型的检测准确性越好。

PR曲线的AUC指标（识别准确率指标之二）

- AUC=Area under the PR Curve
- 2015年VIVA（Vision for Intelligent Vehicles and Applications）交通标志检测比赛。
- VIVA主办方采用了PR曲线（Precision–Recall Curve）的面积覆盖率AUC（Area under Curve）作为对象检测的识别准确性的评价指标。
- 面积覆盖率（AUC）越高，则对象检测的识别准确性越好。



最佳工作状态

- 针对具体应用场景，对精确率和召回率之间进行一个平衡取舍，从而选择合适的阈值参数，使对象检测器处于最佳的工作状态。

- F_1的数学含义其实就是精确率P和召回率R的调和平均数，综合考虑了二者的影响。

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 \cdot P \cdot R}{P + R}$$

视觉对象检测的算法

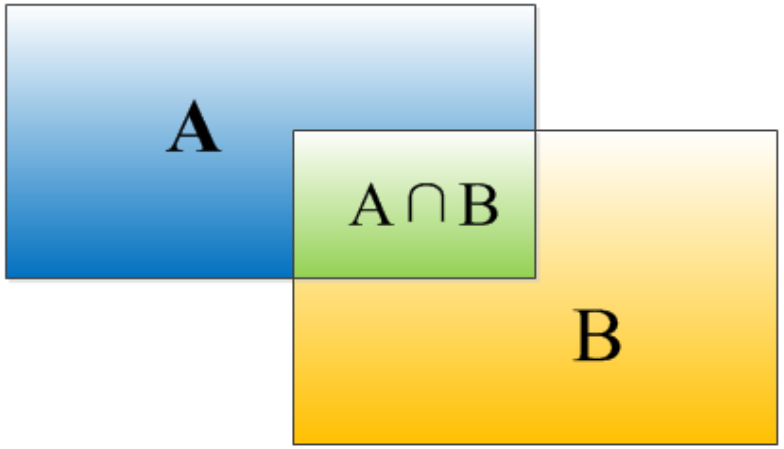
Visual object detection algorithm

IOU（重叠联合比）

- $\mathrm{IOU}_{\mathrm{pred}}^{\mathrm{truth}}$ 表示的是预测框（Prediction）和真实框（Ground Truth）之间的重叠联合比（Intersection over Union）

- IOU定义了2个边界框（ bounding box ）(就是恰好框住对象的矩形框)的重叠度，计算为相交面积（∩）/相并面积（∪）

- $\mathrm{IOU}_{\mathrm{pred}}^{\mathrm{truth}} = \dfrac{\text{Area of Intersection}}{\text{Area of Union}}$



视觉对象检测的错误类型

- 对于模型给出的检测结果，都会根据以下标准，被判定为其中的一种：
- 正确的
  - 正确（Correct）：类别正确，IOU > 0.5
- 错误的
  - 定位错误（Localization）：类别正确，0.1 < IOU < 0.5
  - 相似性错误（Similar）：类别相似，IOU > 0.1
  - 其他错误（Other）：类别错误，IOU > 0.1
  - 背景误认（Background）：IOU < 0.1

视觉对象检测方法

- **R-CNN**
  - Region based convolutional networks for accurate object detection and segmentation, TPAMI, 2015.
  - Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.
- **Fast R-CNN**
  - Fast R-CNN, ICCV 2015.
- **Faster R-CNN**
  - Faster R-CNN, NIPS, 2015.
- **YOLOv1-->YOLOv3**
  - You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016.
- **SSD**
  - SSD: Single Shot MultiBox Detector, ECCV 2016.

http://wiki.icenter.tsinghua.edu.cn/icenterwiki/index.php/计算机视觉

参考资料

- [1] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 580-587, 2014.
- Region based convolutional networks for accurate object detection and segmentation, TPAMI 2016.
- [2] Girshick R. Fast R-CNN[C]. IEEE International Conference on Computer Vision, 2015: 1440-1448.
- [3] 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 & Machine Intelligence, 2016, 39(6):1137-1149.

R-CNN

- R-CNN: 全名叫Regions with CNN features / Region-based Convolutional Neural Networks
- 将卷积神经网络应用region proposal的策略，自底下上训练可以用来定位目标物和图像分割
- 当标注数据是比较稀疏的时候，在有监督的数据集上训练之后到特定任务的数据集上fine-tuning（微调参数，总体网络架构不变了）可以得到较好的性能。
- 用ImageNet上训练好的模型，在需要训练的数据上fine-tuning一下，检测效果很好。
- 突破性：当时在Pascal VOC数据集上测试性能最好，达到的效果比当时最好的DPM方法 mAP还要高上20点。
  https://www.rossgirshick.info/

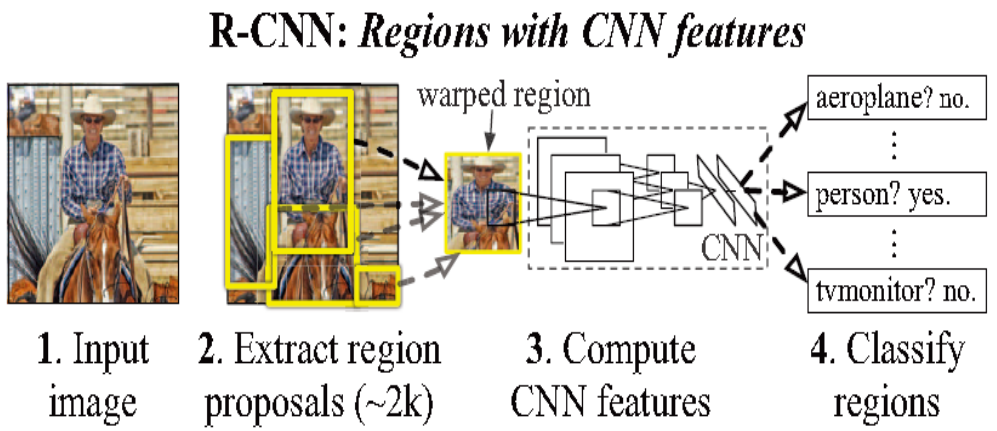


Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of **53.7% on PASCAL VOC 2010**. For comparison, [34] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%.

R-CNN

- 输入图像，提取提炼区域（region）：
  - 用选择性搜索（selective search）的算法去搜索一个‘fast mode’(快速模式)，对每一个提出的可能有对象的图像区域提取出一个4096维的特征向量。
  - 对于不是标准227*227像素的正方形的区域，使其标准化。最简单的方法是膨胀（dilate, 形态学算法）其最小外边框（设宽度=16 pixels），使整幅图像大小合适。
- 计算CNN特征：
  - CNN网络架构： 5个卷积层（Convolution Layers），2个全连接(Fully Connected Layers), 正如Yann Le Cun之前提出的LeNet算法。
- 区域分类：
  - 对每一个类预先训练好一个支持向量机（SVM），然后对之前提炼出来的特征向量（feature vector）用对应类的SVM去“打分”。
  - 贪心思想的“非极大值抑制”(non-maximum suppression)算法：如果一个区域和一个有更高打分的区域有交集（Intersection-over-Union （IoU））并且IoU的值>某个阈值，那么这个区域（得分相对低的）将被舍弃。

R-CNN的缺点

- 训练分为3个步骤的流水线（对候选区提取特征的微调卷积网络，训练线性SVM作为对象探测器，处理proposal计算卷积特征，边界框（BBOX）回归运算）；
- 训练时间和空间开销大。要从每一张图像上提取大量proposal，还要从每个proposal中提取特征，并存到磁盘中；
- 测试时间开销大。要从每个测试图像上，提取大量proposal，再从每个proposal中提取特征来进行检测过程；
- 速度慢。一个原因是在前向运算时对每一个候选区域的对象分别计算，并没有用共享权值或共享模型参数的方法加快。

Fast R-CNN改进R-CNN

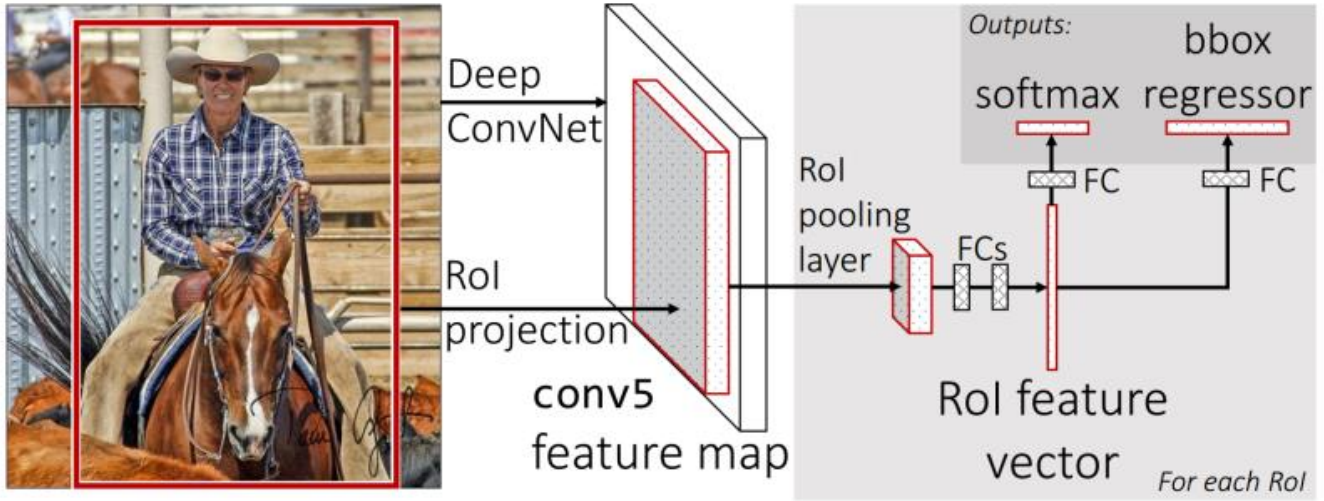
- 1. 比R-CNN更高的检测质量（mAP）；
- 2. 把多个任务的损失函数写到一起，实现单级的训练过程;
- 3. 在训练时可更新所有的层;
- 4. 不需要在磁盘中存储特征。

Fast R-CNN

1. 使用外部算法（选择性搜索SS ）来找出候选区域（2000个object proposal），找出感兴趣的区域（Regions of Interest, RoI），映射到特征空间里；
2. 缩放图片的scale得到图片金字塔，得到conv5的特征金字塔；
3. 对于每个scale的每个ROI，求取映射关系，在conv5中crop出对应的patch；并用一个单层的空间金字塔池化层（SPP） layer（称为RoI pooling layer）来统一到一样的尺度，因为后续的全连接层输入的所有向量有同样的大小；
4. 连续续经过两个全连接层得到特征，特征又分别共享到到两个新的全连接层，分别对应两个优化目标
   - 第一个优化目标是分类，使用softmax，
   - 第二个优化目标是边界框回归（bbox regression），使用了一个smooth的L1-loss（一次函数和小量时二次函数的结合）。

Fast R-CNN优点

- Fast R-CNN 实现了端到端的联合训练（end-to-end joint training） (single stage)
- R-CNN用SVM训练特征时需要中间大量的磁盘空间存放特征，Fast RCNN没有了SVM这一步，所有的特征都暂存在显存中，不需要额外的磁盘空间。
- R-CNN中因为ROI-centric的原因，测试时间开销大，Fast R-CNN进一步通过single scale(pooling->spp just for one scale) testing和SVD（奇异值分解）(降维)分解全连接来提速。

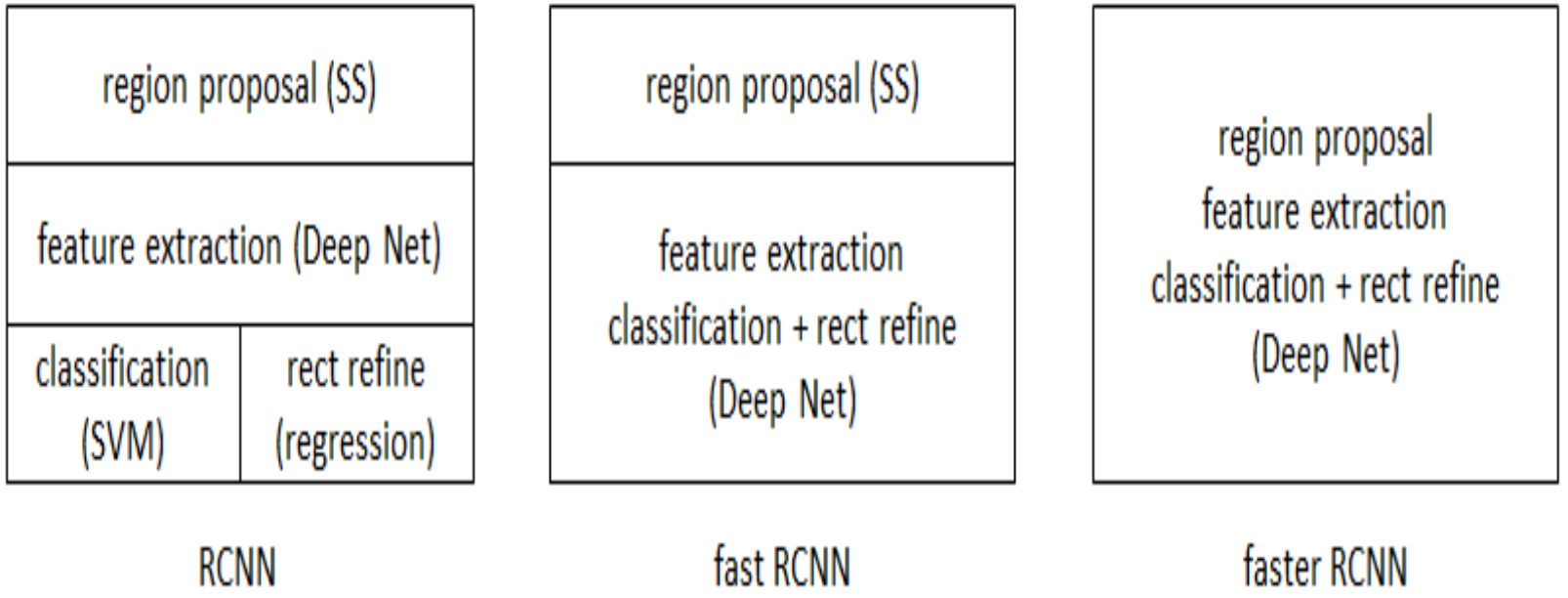


Faster R-CNN改进Fast R-CNN

- Faster R-CNN速度更快，精确度更高。
- Faster R-CNN中，每个网络可以独立训练或联合训练。
- 模型有4个损失函数：
  - RPN（区域生成网络）分类是否对象；
  - RPN 边界框提议；
  - Fast R-CNN 对象分类；
  - Fast R-CNN 边界框回归 。

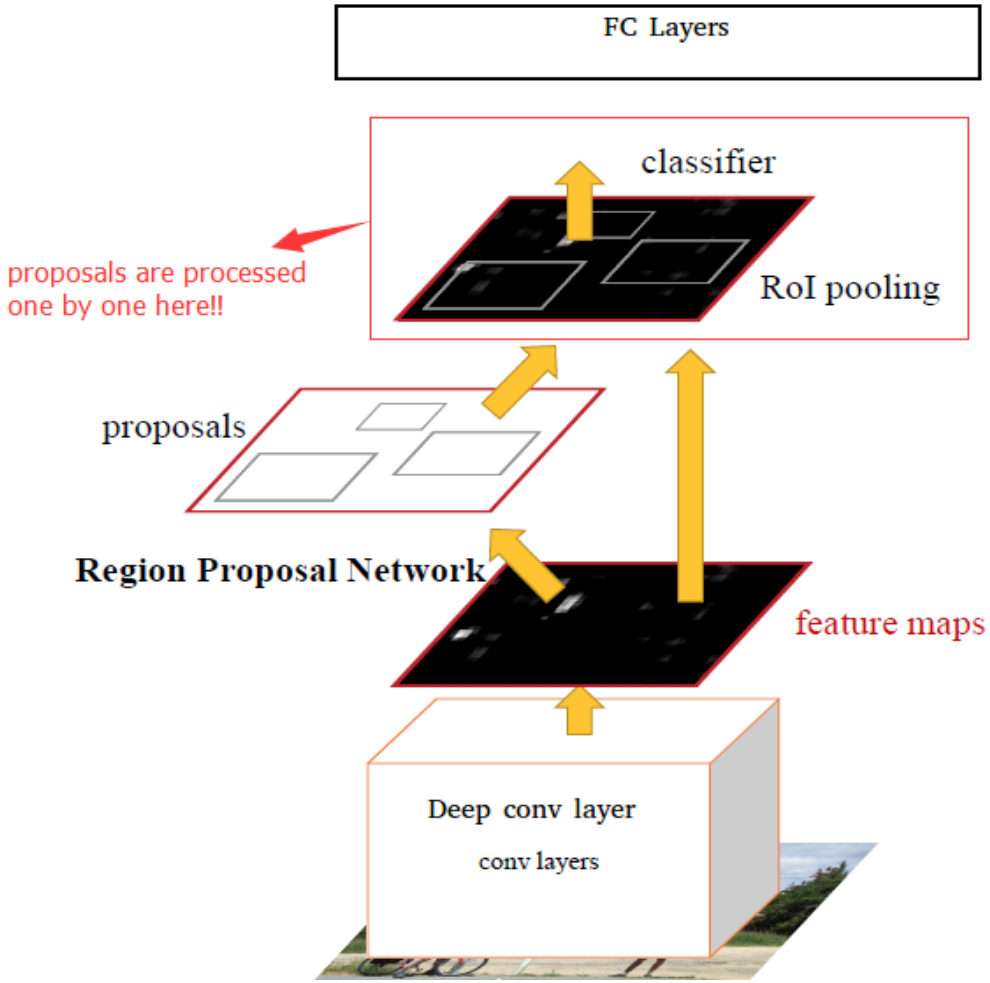
Faster R-CNN

- Faster RCNN可以简单地看做“RPN+fast R-CNN“的系统，用RPN代替fast R-CNN中的Selective Search方法。
- RPN区域生成网络



Faster R-CNN

- Faster R-CNN包含2个模块：
  - RPN(Region Proposal Network):在深度卷积层基础上给出一系列的矩形候选区域。
  - Fast R-CNN RoI 池化层：对每个proposal 区域进行分类，提取proposal定位。
- ***主要思想***是用最后一个卷积层去推断候选区域。



Faster R-CNN效果

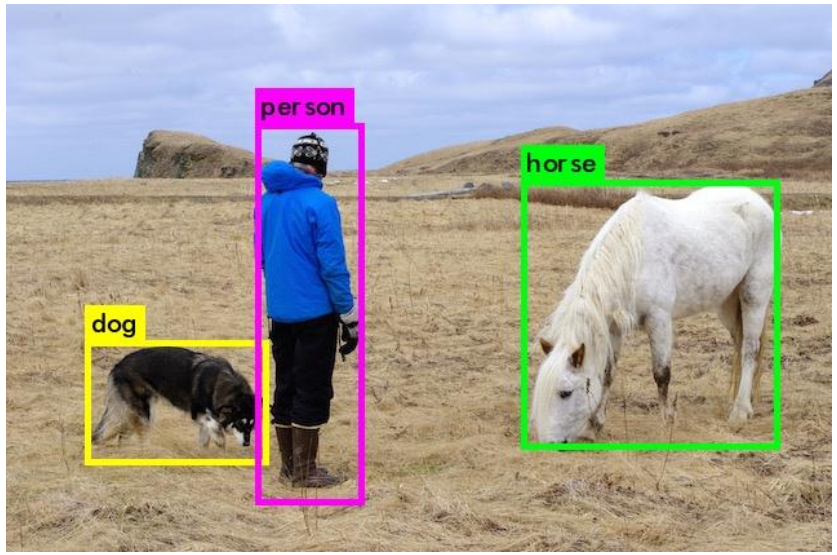
- Faster R-CNN用一个101层的resnet架构，称为ResNet101
- 对每幅图像（包括proposals）的处理速度是R-CNN的250倍，是Fast R-CNN的10倍。
- 精确度和Fast R-CNN一样，都比R-CNN高。

YOLO对象检测的算法

YOLO: You Only Look Once

YOLO算法

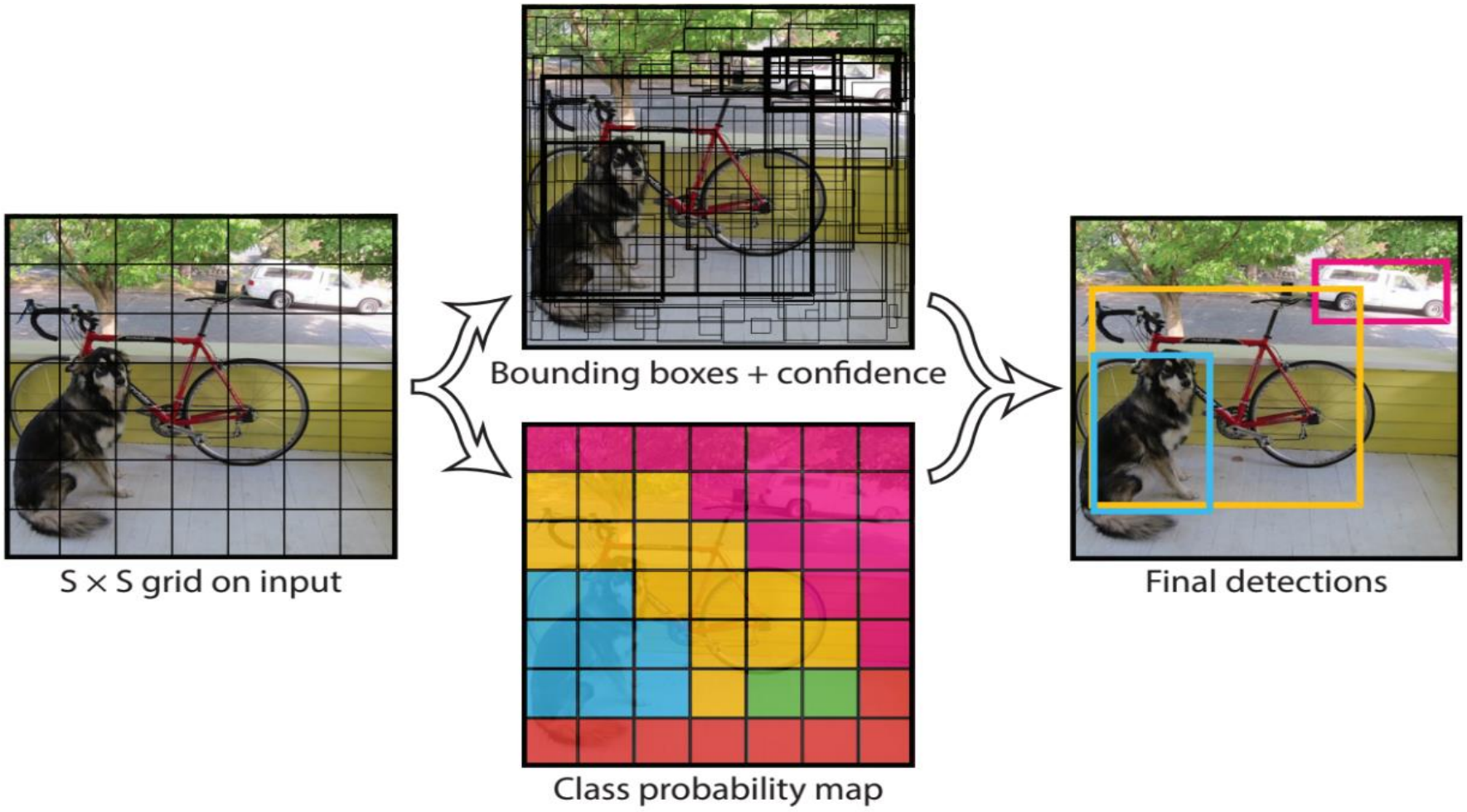
- YOLO算法将对目标检测任务的认识由分类问题（Classification）化简为回归问题（Regression）
- 在保证精度不过多损失的前提下，极大地提高了检测速度。
- 运算速度高，在Titan X GPU上的运行速度可以达到45 FPS（实时）



参考资料

- [1] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. Computer Vision and Pattern Recognition, 2016:779-788.
- [2] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision, 2016:21-37.

YOLO v1


Bounding boxes + confidence
S × S grid on input
Class probability map
Final detections

SSD对象检测的算法

SSD: Single Shot MultiBox Detector

SSD

- SSD方法的核心：
  - 预测对象（ predict object ）及其归属类别的score（得分）
  - 在 feature map上使用小的卷积核去predict一系列bounding boxes的box offsets
- 为了得到高精度的检测结果：
  - 在不同层次的 feature maps（特征图谱）上去 predict object、box offsets,
  - 得到不同aspect ratio（纵横比）的predictions。
- 改进设计：
  - 能够在当输入分辨率较低的图像时，保证检测的精度。
  - 整体端到端（end-to-end）的设计，训练也变得简单。
  - 在检测速度、检测精度之间取得较好的折衷。
- SSD，比YOLOv1方法，还要快，还要精确。
- SSD，在保证速度的同时，mAP指标与使用region proposals 技术的方法（如 Faster R-CNN）相媲美

图像语义分割方法

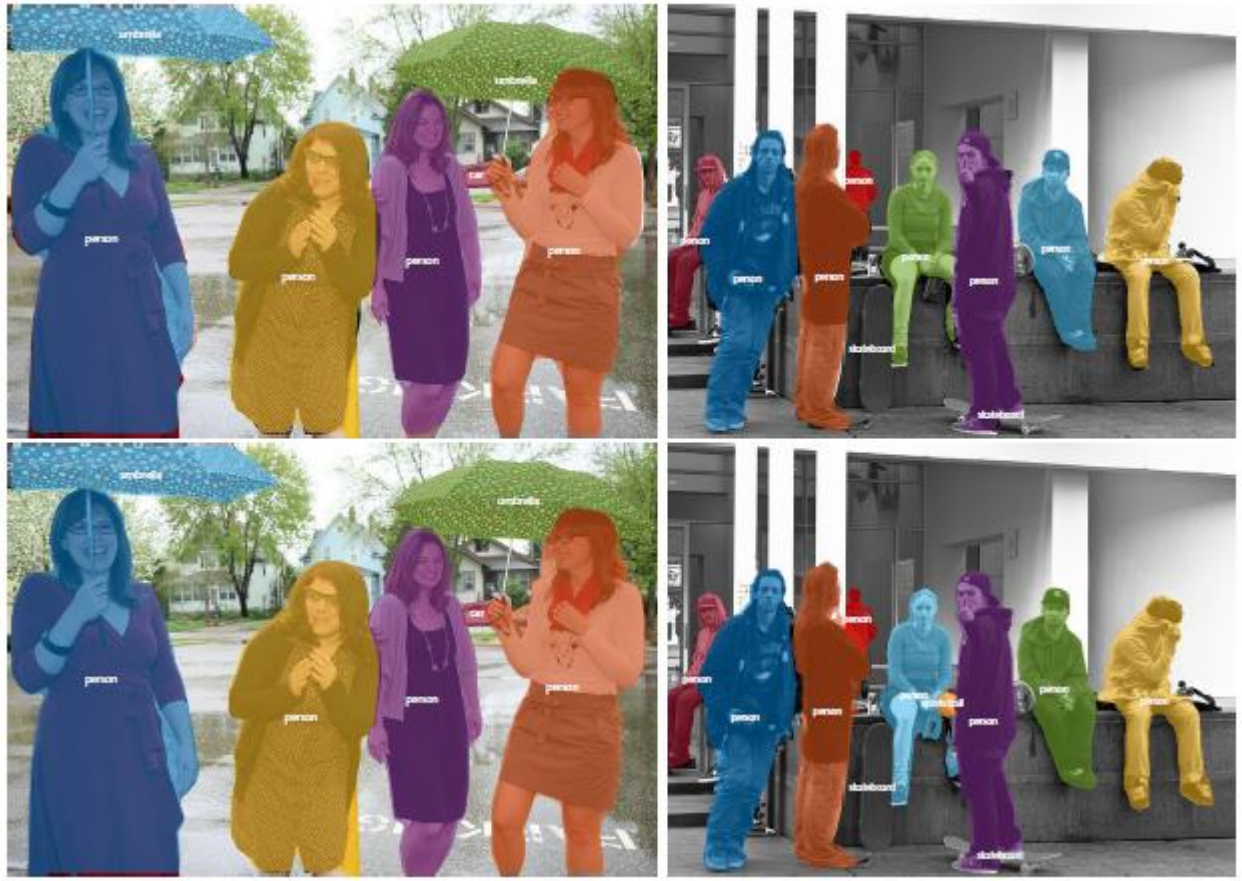
Semantic Segmentation

参考资料

- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, Mask R-CNN, ICCV 2017.
- Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, Ross Girshick, Learning to Segment Every Thing, CVPR 2018.
- Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollár, Panoptic Segmentation, CVPR 2019.
- Xinlei Chen, Ross Girshick, Kaiming He, Piotr Dollár, TensorMask: A Foundation for Dense Object Segmentation, 2019.

Semantic Segmentation

- TensorMask and Mask R-CNN



参考资料

- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
- [5] Huang et al., Speed/accuracy trade-offs for modern convolutional object detectors[C], CVPR 2017. (https://arxiv.org/abs/1611.10012)

谢谢指正！

zhenchen@tsinghua.edu.cn