【一图看懂】计算机视觉识别简史：从 AlexNet、ResNet 到 Mask RCNN

2017-04-30新智元

新智元编译

来源：medium

作者：Đặng Hà Thế Hiển

编译：新智元编辑部

。

【新智元导读】 Medium 用户 Đặng Hà Thế Hiển 制作了一张信息图示，用专业、简洁并且最有吸引力的方式——信息图示，讲述计算机视觉（CV）物体识别的现代史。不仅总结了 CV 6 大关键技术和目标识别的重要概念，整个信息图示从 2012 年 AlexNet 赢得了 ILSVRC（ImageNet 大规模视觉识别挑战赛）说起，总结了至今关键的 13 大模型及其概念，比如 VGGNet、ResNet、Inception 到最近的 Mask RCNN。作者特别强调，所有参考文献都精挑细选，以便读者能够知道从哪里找到有关细节的解释。

最近，物体识别已经成为计算机视觉和 AI 最令人激动的领域之一。即时地识别出场景中所有的物体的能力似乎已经不再是秘密。随着卷积神经网络架构的发展，以及大型训练数据集和高级计算技术的支持，计算机现在可以在某些特定设置（例如人脸识别）的任务中超越人类的识别能力。

我感觉每当计算机视觉识别方面有什么惊人的突破发生了，都得有人再讲一遍是怎么回事。这就是我做这个图表的原因。它试图用最简洁的语言和最有吸引力的方式讲述物体识别的现代史。故事开始于 2012 年 AlexNet 赢得了 ILSVRC（ImageNet 大规模视觉识别挑战赛）。信息图由 2 页组成，第 1 页总结了重要的概念，第 2 页则勾画了历史。每一个图解都是重新设计的，以便更加一致和容易理解。所有参考文献都是精挑细选的，以便读者能够知道从哪里找到有关细节的解释。

下载地址：https://github.com/Nikasa1889/HistoryObjectRecognition/find/master

Modern History of Object Recognition Infographic

MiniMap

2012 AlexNet RCNN OverFeat 2013 ZFNet SPPNets

YOLO Fast RCNN InceptionNet VGGNet 2014 MultiBox

2015 ResNet Faster RCNN 2016 SSD 2017 MaskRCNN

Image Classification

Classify an image based on the dominant object inside it.

datasets: MNIST, CIFAR, ImageNet

Object Localizatio

Predict the im that conta dominant ob image classif be used to object in the i

datasets: Ima

Semantic Segmentation

Label each pixel of an image by the object class that it belongs to, such as human, sheep, and grass in the example

datasets: PASCAL, COCO

Instance Segmenta

Label each p image by t class and instance that

datasets: PAS

1958

1965

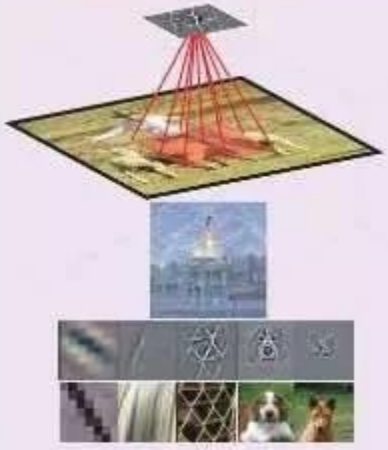
1970s

1993

1995

2000s

Important CNN Concepts



Feature[4,5,8] *(pattern, activation of a neuron, feature detector)*

A hidden neuron that is activated when a particular pattern (feature) is presented in its input region (receptive field).

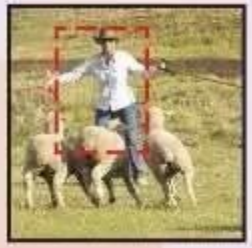
The pattern that a neuron is detecting can be visualized by (1) optimizing its input region to maximize the neuron's activation (deep dream), (2) visualizing the gradient or guided gradient of the neuron activation on its input pixels (back propagation and guided back propagation), (3) visualizing a set of image regions in the training dataset that activate the neuron the most.

Receptive Field[2] *(input region of a feature)*

The region of the input image that affects the activation of a feature. In other words, it is the region that the feature is looking at.

Generally, a feature in a higher layer has a bigger receptive field, which allows it to learn to capture

Important Object Re



Bounding b

proposal, box prop

A rectangular r contains an o generated by selective searc

A bounding bo vector, either st x1, y1), or (mo and its width a usually accom likely the box co

Offset =Δ(x, y, w, h)
Distance = ||Δ||

The difference measured by th tions. w and distance calcul

Intersection over Union

(IoU, Jaccard similarity):

A metric that measures the

Non M

A com bound

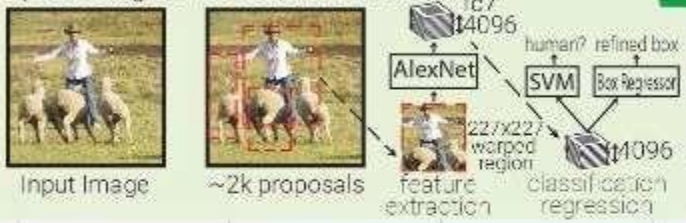
Region Proposals or Sliding Windows

RCNN and OverFeat represent two early competing ways to do object recognition: either classify regions proposed by another method (RCNN, FastRCNN, SPPNet), or classify a fixed set of evenly spaced square windows (OverFeat). The first approach has region proposals that fit the objects better than the other grid-like candidate windows but is two orders of magnitude slower. The second approach takes advantage of the convolution operation to quickly regress and classify objects in sliding-windows fashion.

Multibox ended this competition by introducing the ideas of prior box and region proposal network. Since then, all state-of-the-art methods now has a set of prior boxes (generated based on a set of sliding windows or by clustering ground-truth boxes) from which bounding box regressors are trained to propose regions that better fit the object inside. The new competition is between the *direct classification* (YOLO, SSD) and *refined classification* approaches (FasterRCNN, MaskRCNN).

2012

Region-based ConvNet (RCNN) is a natural combination of heuristic region proposal method and ConvNet feature extractor. From an input image, ~2000 bounding box proposals are generated using selective search. Those proposed regions are cropped and warped to a fixed-size 227x227 image. AlexNet is then used to extract 4096 features (fc7) for each warped image. An SVM model is then trained to classify the object in the warped image using its 4096 features. Multiple class-specific bounding box regressors are also trained to refine the bounding box proposal using the 4096 extracted features.



RCNN

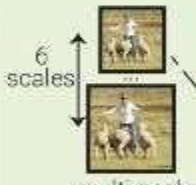
AlexNet

The mod
the deve
in 2012
margin.
orthogon
can be c
ConvNet

AlexNet
data aug
tion. It pr
glorious
vision.

OverFeat

OverFe
evenly sp
scales o
class-agn
and refin
layer (33
fc layers
multi-sca
pixels wh
usually not well
exhaustive poolin
input, which resu
spaced 12pixels i



2013

ZFNet

ZFNet is the ILSVRC 2013 winner, which is basically AlexNet with a minor modification: use 7x7 kernel instead of 11x11 kernel in the first Conv layer to retain more information.

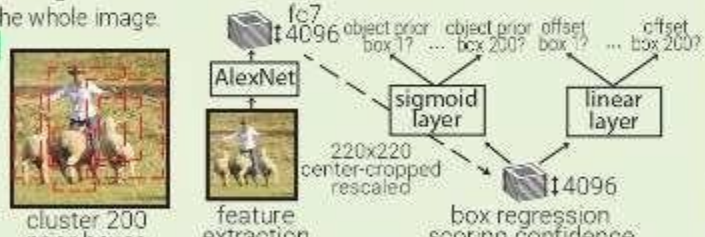
SPPNet (Spatial Pyramid Pooling net) is essentially an enhanced version of RCNN by introducing two important concepts: adaptively-sized pooling (the SPP layer), and computing feature volume only once. In fact, the Fast-RCNN embraced these ideas to fasten RCNN with minor modifications.

SPPNet

SPPNet uses selective search to propose 2000 region proposals per image. It then extracts a common global feature volume from the entire image using ZFNet-Conv5. For each region proposal, SPPNet uses spatial pyramid pooling (SPP) to pool features in that region from the global feature volume to generate its fixed-length representation. This representation is used for training the object classifier and box regressors. Pooling features from a common global feature volume rather

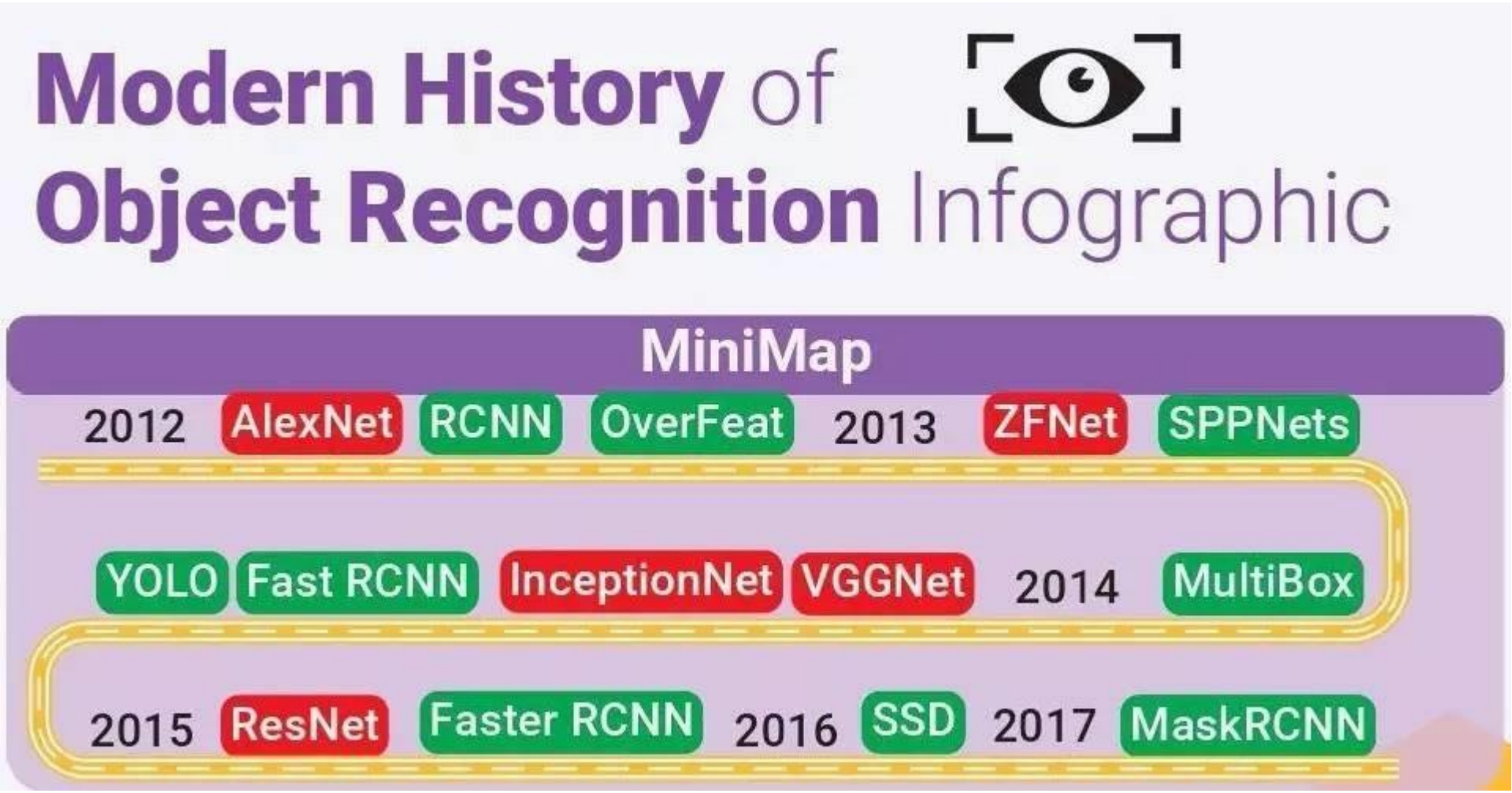
MultiBox

MultiBox is not an object recognition but a ConvNet-based region proposal solution. It popularized the ideas of region proposal network (RPN) and prior box, proving that ConvNet can be trained to propose better region proposals than heuristic approaches. Since then, heuristic approaches have been gradually fading out and replaced by RPN. MultiBox first clusters all ground truth box locations in the whole dataset to find 200 centroids that it uses as prior boxes' centers. Each input image is center cropped and rescaled to 220x220. Then it uses AlexNet to extract 4096 features (fc7). A 200-sigmoid layer is added to predict the object confidence score, and 4x200-linear layer is added to predict center offset and scale of box proposal from each prior box. Note that box regressors and confidence scorers look at features extracted from the whole image.

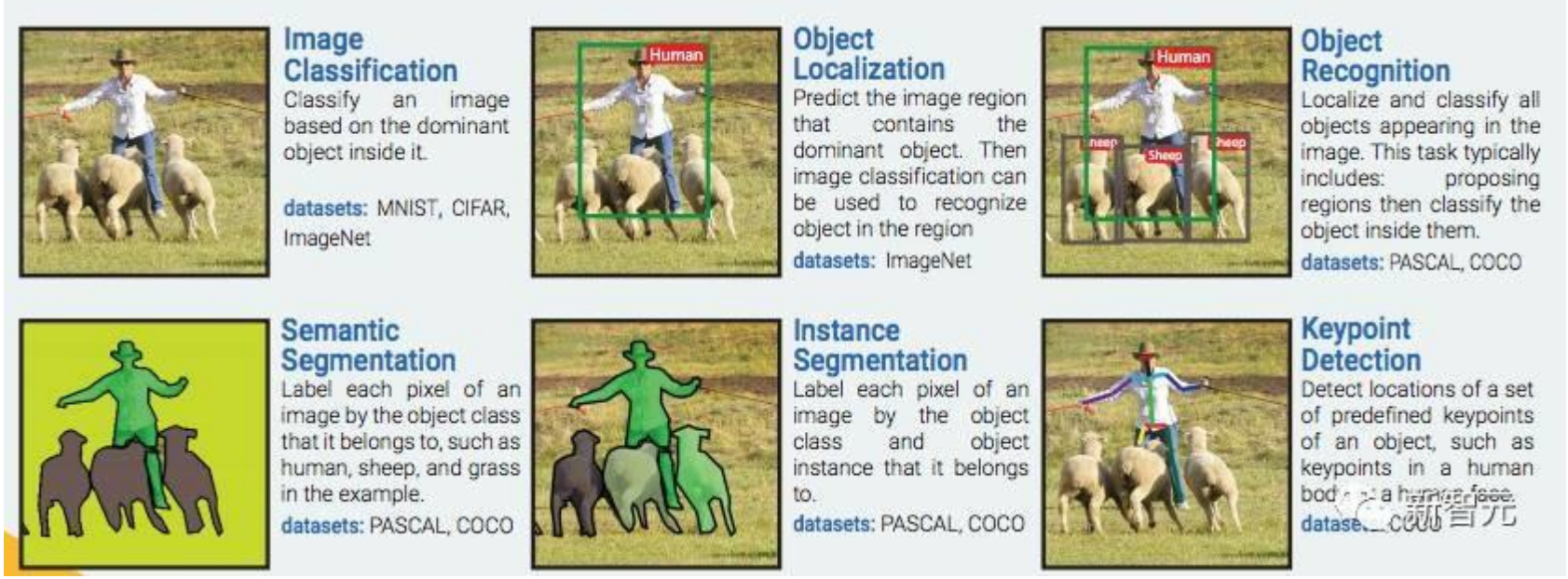


Although not ar
common Con
simplicity and
large-kernel co
strictly uses 3x
with 2x2 maxpo

点击图片放大查看及保存



计算机视觉 6 大关键技术



- 图像分类：根据图像的主要内容进行分类。数据集：MNIST, CIFAR, ImageNet
- 物体定位：预测包含主要物体的图像区域，以便识别区域中的物体。数据集：ImageNet
- 物体识别：定位并分类图像中出现的所有物体。这一过程通常包括：划出区域然后对其中的物体进行分类。数据集：PASCAL, COCO

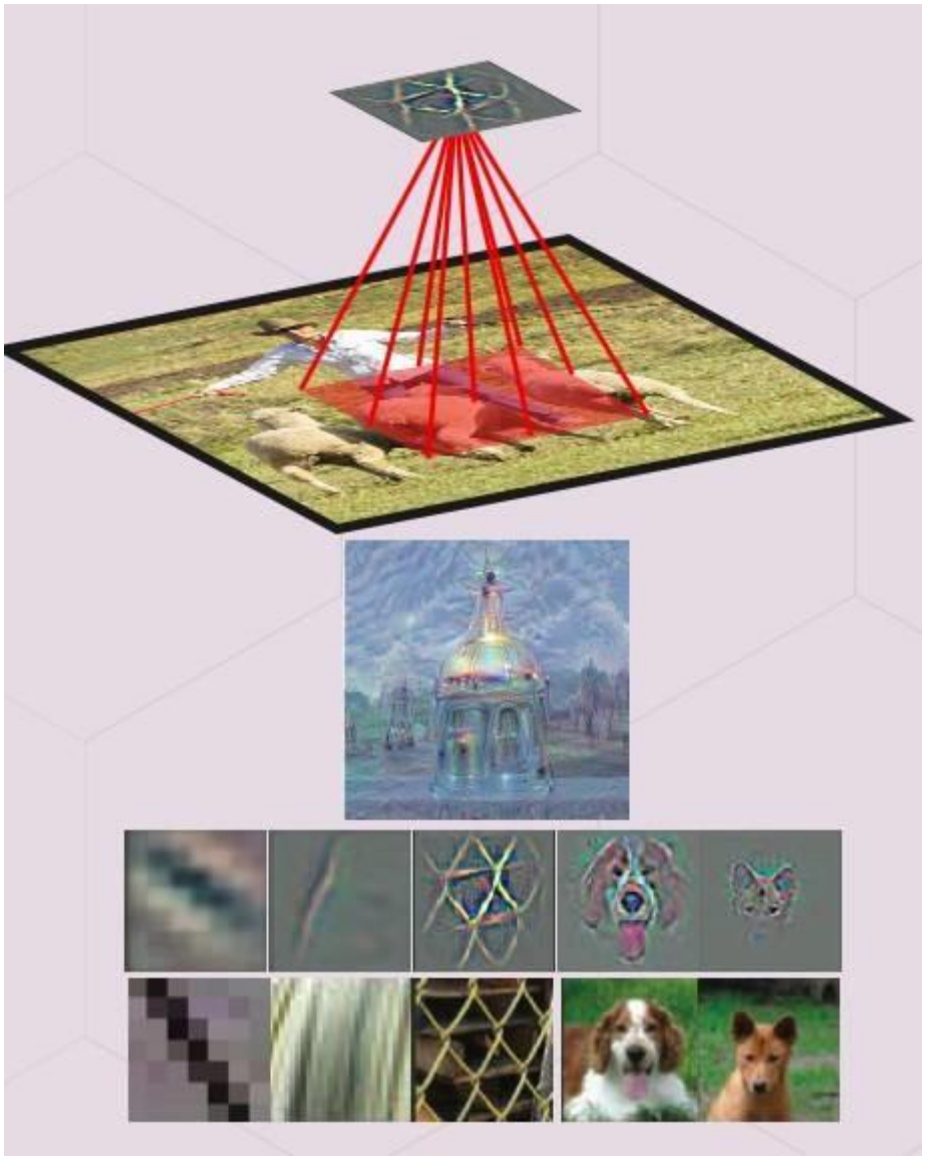
- 语义分割：把图像中的每一个像素分到其所属物体类别，在样例中如人类、绵羊和草地。数据集：PASCAL, COCO
- 实例分割：把图像中的每一个像素分到其物体类别和所属物体实例。数据集：PASCAL, COCO
- 关键点检测：检测物体上一组预定义关键点的位置，例如人体上或者人脸上的关键点。数据集：COCO

关键人物

这种图列出了物体识别技术中的关键人物：J. Schmidhuber； Yoshua Bengio ；Yann Lecun； Georey Hinton ；Alex Graves ；Alex Krizhevsky ；Ilya Sutskever ；Andrej Karpathy； Christopher Olah ；Ross Girshick； Matthew Zeiler ；Rob Fergus ；Kaiming He ； Pierre Sermanet ；Christian Szegedy ；Joseph Redmon ；Shaoqing Ren ；Wei Liu ；Karen Simonyan；Andrew Zisserman；Evan Shelhamer ；Jonathan Long ；Trevor Darrell； Springenberg ；Mordvintsev ；V. Dumoulin ；Francesco Visin； Adit Deshpande ……

重要的 CNN 概念

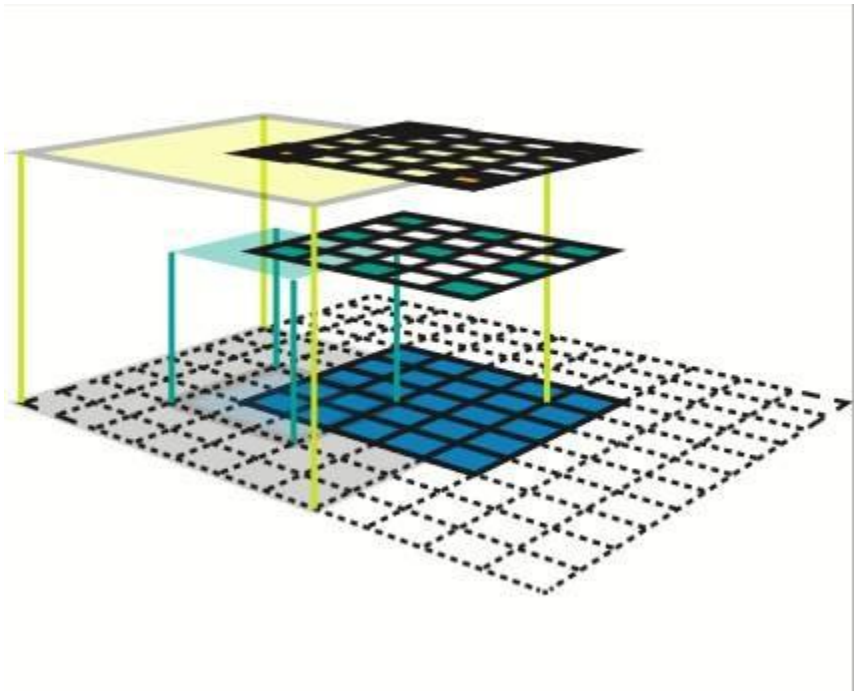
1. 特征（图案，神经元的激活，特征探测）



当一个特定的图案（特征）被呈现在输入区（接受域）中时，一个隐藏的神经元就被会被激活。

神经元识别的团可以被进行可视化，其方法是：1）优化其输入区，将神经元的激活（deep dream）最大化；2）将梯度（gradient）可视化或者在其输入像素中，引导神经元激活的梯度（反向传播以及经引导的反向传播）3）将训练数据集中，激活神经元最多的图像区域进行可视化。

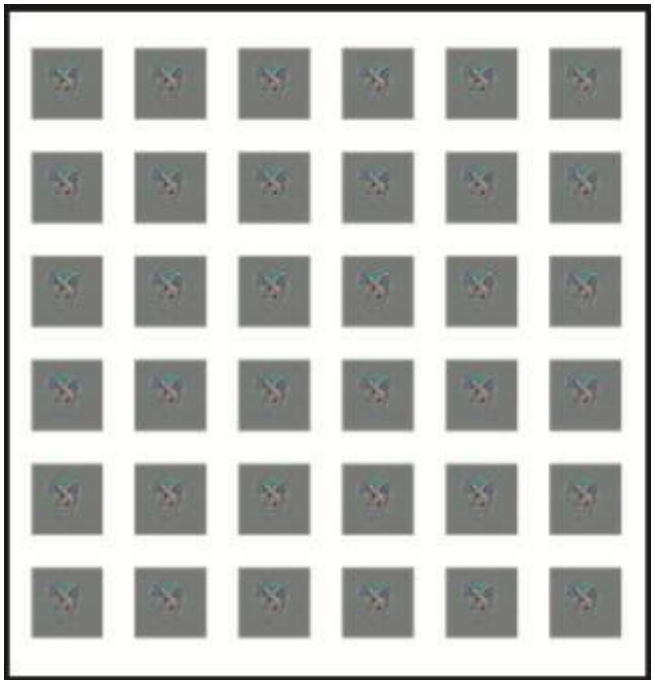
2. 感受野（特征的输入区）



输入图像区会影响特征的激活。换句话说，它就是特征参考的区域。

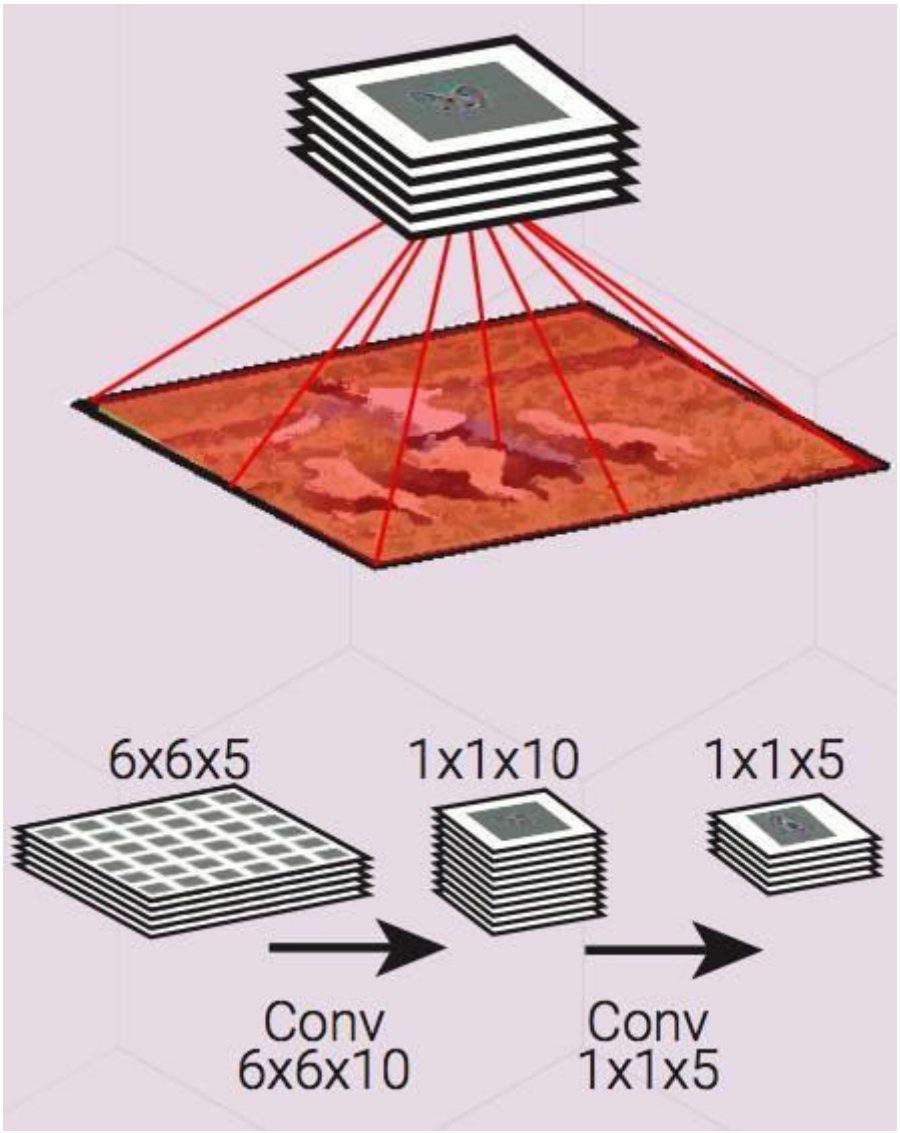
通常，越高层上的特征会的接受域会更宽，这能让它能学会捕捉更多的复杂/抽象图案。ConvNet 的架构决定了感受野是如何随着层数的改变而改变的。

3. 特征地图（feature map，隐藏层的通道）



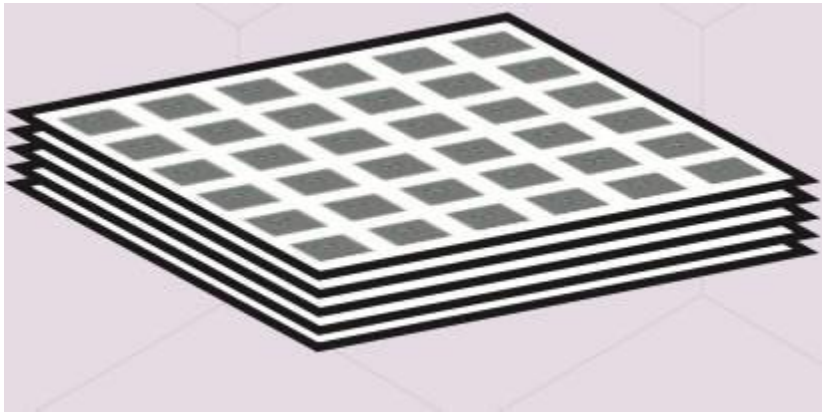
指的是一系列的特征，通过在一个滑动窗口（例如，卷积）的方式，在一个输入地图中的不同位置应用相同的特征探测器来创造。在相同的特征地图上的特征，有着一致的可接收形状，并且会寻找不同位置上的相同图案。这构成了 ConvNet 的空间不变性。

4. 特征量（卷积中的隐藏层）


6x6x5
1x1x10
1x1x5
Conv
6x6x10
Conv
1x1x5

这是一组特征地图，每一张地图会在输入地图中的一些固定位置搜寻特定的特征。所有的特征的接受域大小都是一样的。

5.作为特征量的全连接层

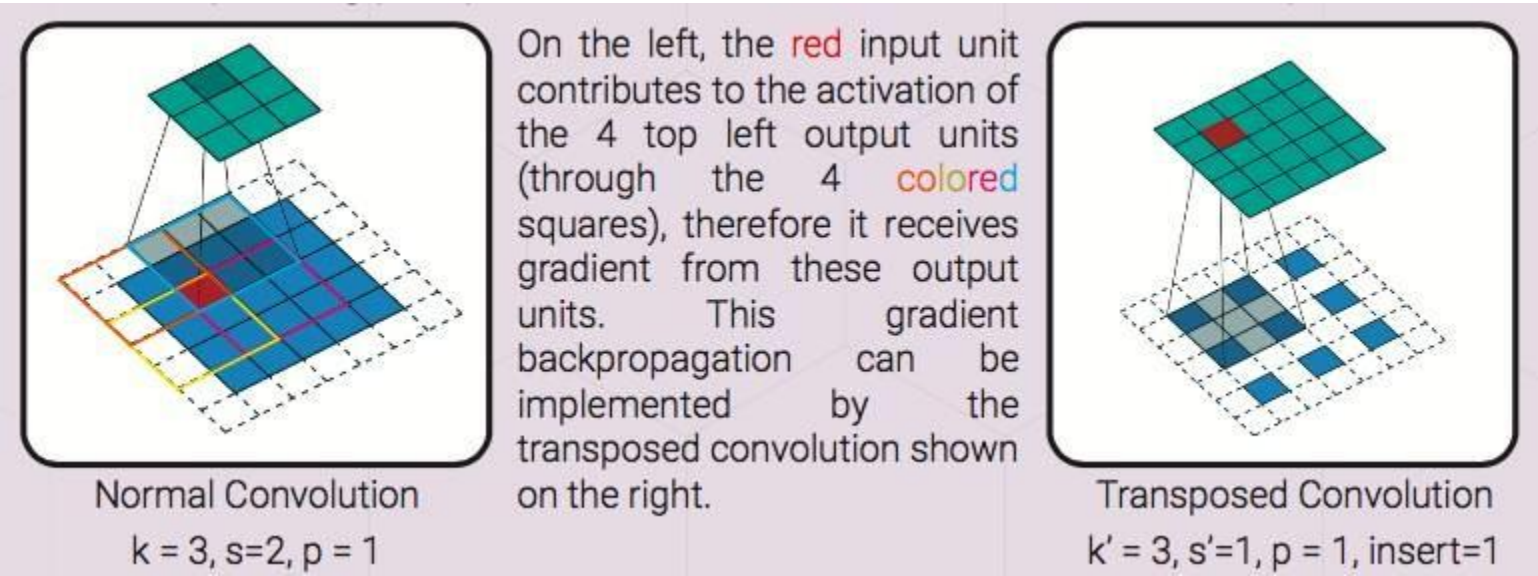


全连接层（fc layers，在识别任务中通常粘附在一个 ConvNet 的尾端），这一特征量在每一张特征滴入上都有一个特征，其接收域会覆盖整张图像。全连接层中的权重矩阵 W 可以被转化成一个 CNN 核。

将一个核 wxhxk 卷积成一个 CNN 特征量 wxhxd 会得到一个 1x1xk 特征量（=FC layer with k nodes)。将一个 1x1xk 的过滤核卷积到一个 1x1xd 特征量，得到一个 1x1xk 的特征量。通过卷积层替换完全连接的图层可以使 ConvNet 应用于任意大小的图像。

6. 反卷积

这一操作对卷积中的梯度进行反向传播。换句话说，它是卷积层的反向传递。反向的卷积可以作为一个正常的卷积部署，并且在输入特征中不需要任何插入。



左图，红色的输入单元负责上方四个单元的激活（四个彩色的框），进而能从这些输出单元中获得梯度。这一梯度反向传播能够通过反卷积（右图）部署。

7. 端到端物体识别管道（端到端学习/系统）

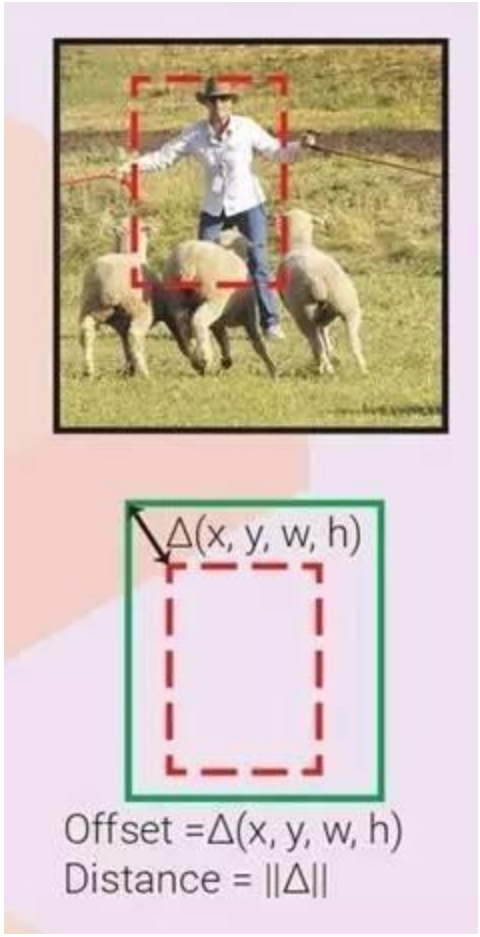
这是一个包含了所有步骤的物体识别管道（预处理、区域建议生成、建议分类、后处理），可以通过优化单个对象函数来进行整体训练。单个对象函数是一个可差分的函数，包含了所有的处理步骤的变量。这种端到端的管道与传统的物体识别管道的完全相反。在这些系统中，我们还不知道某个步骤的变量是如何影响整体的性能的，所以，么一个步骤都必须要独立的训练，或者进行启发式编程。

重要的目标识别概念

1. Bounding box proposal

提交边界框（Bounding box proposal，又称兴趣区域，提交区域，提交框）

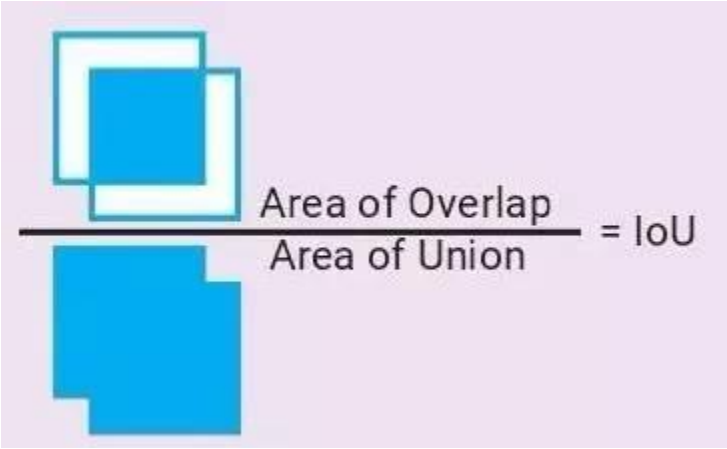
输入图像上的一个长方形区域，内含需要识别的潜在对象。提交由启发式搜索（对象、选择搜索或区域提交网络 RPN）生成。



一个边界框可以由 4 元素向量表示，或表达为 2 个角坐标（x0,y0,x1,y1），或表达为一个中心坐标和宽与高（x,y,w,h）。边界框通常会配有一个信心指数，表示其包含对象物体的可能性。

两个边界框的区别一般由它们的向量表示中的 L2 距离在测量。w 和 h 在计算距离前会先被对数化。

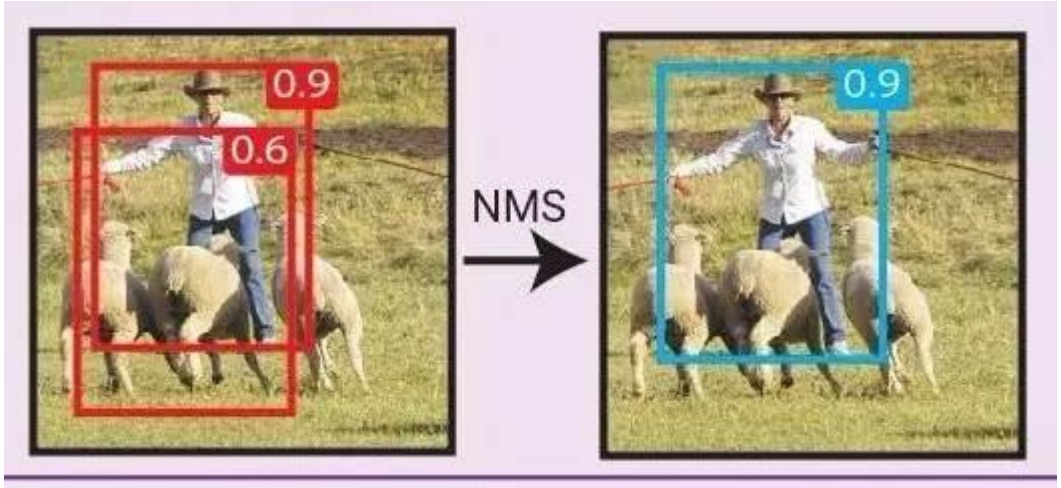
2. Intersection over Union



重叠联合比（Intersection over Union，又称 IoU，Jaccard 相似度）

两个边界框相似度的度量值=它们的重叠区域除以联合区域

3. 非最大抑制（Non Maxium Suppression，又称 NMS）



一个融合重叠边界框（提交或侦测出的）的一般性算法。所有明显和高信度边界框重叠的边界框（IoU > IoU_threshold）都会被抑制（去除）。

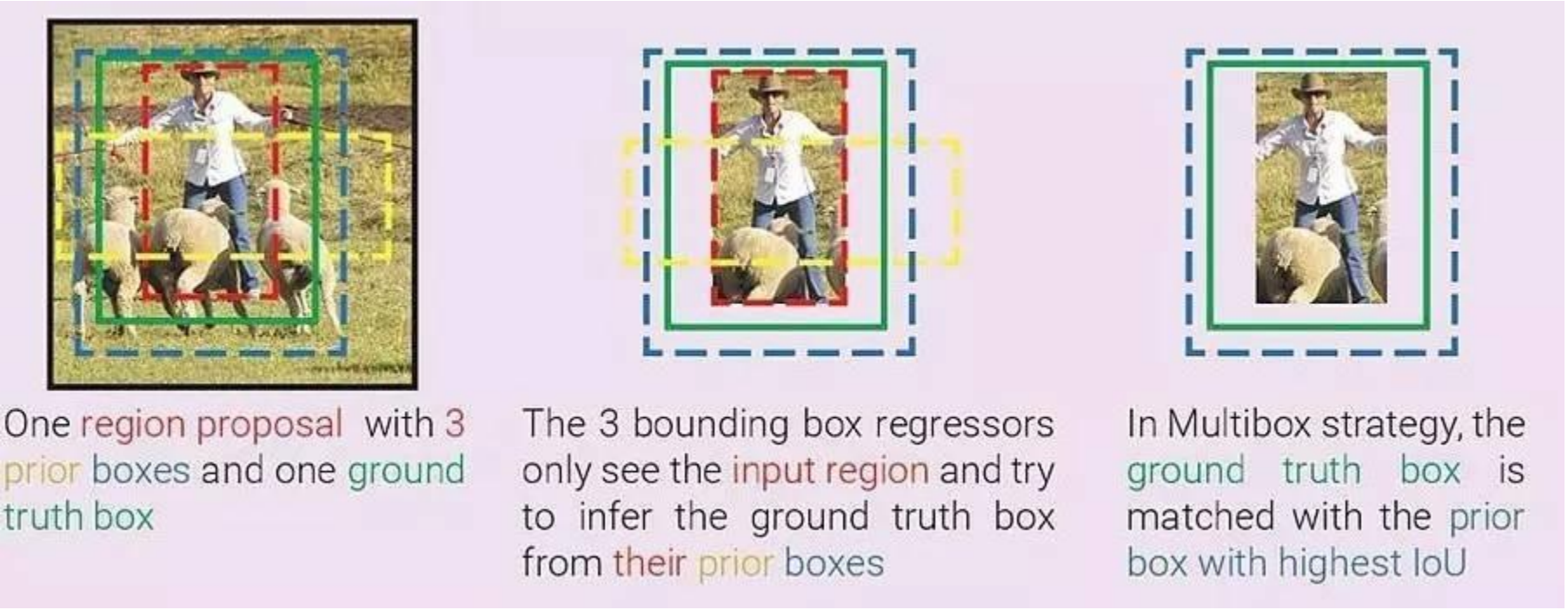
4. 边界框回归（边界框微调）

观察一个输入区域，我们可以得到一个更适合隐含对象的边界框，即使该对象仅部分可见。下图显示了在只看到一部分对象时，得出真实边界框（ground truth box）的可能性。因此，可以训练回归量，来观察输入区域，并预测输入区域框和真实框之间的 offset ∆（x,y,w,h）。如果每个对象类别都有一个回归量，就称为特定类别回归量，否则就称为不可知类别（class-agnostic，一个回归量用于所有类别）。边界框回归量经常伴有边界框分类器（信度评分者），来评估边界框中在对象存在的可信度。分类器既可以是特定类别的，也可以是不可知类别的。如果不定义首要框，输入区域框就扮演首要框的角色。

5. 首要框（Prior box，又称默认框、锚定框）

如果不使用输入区域作为唯一首要框，我们可以训练多个边界框回归量，每一个观测相同的输入区域，但它们各自的首要框不同。每一个回归量学习预测自己的首要框和真实框之间的 offset。这样，带有不同首要框的回归量可以学习预测带有不同特性（宽高比，尺寸，位置）的边界框。相对于输入区域，首要框可以被预先定义，或者通过群集学习。适当的框匹配策略对于使训练收敛是至关重要的。

6. 框匹配策略



我们不能指望一个边界框回归量可以预测一个离它输入区域或首要框（更常见）太远的对象边界框。因此，我们需要一个框匹配策略，来判断哪一个首要框与真实框相匹配。每一次匹配对回归来说都是一个训练样本。可能的策略有：（多框）匹配每一个带有最高 IoU 的首要框的真实框；（SSD，FasterRCNN）匹配带有任何 IoU 高于 0.5 的真实框的首要框。

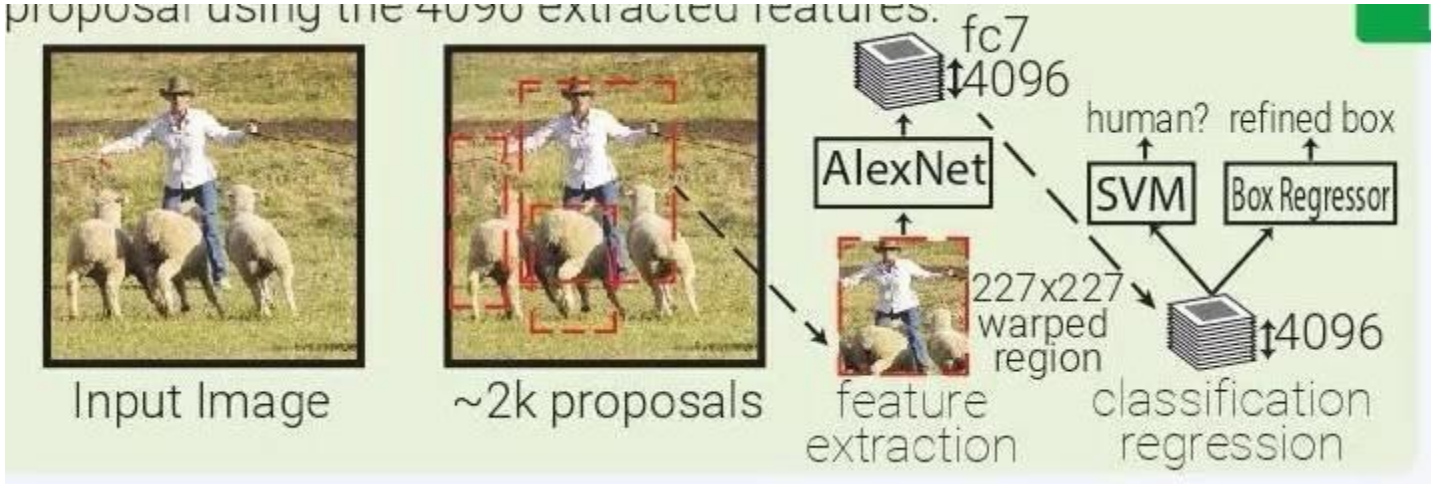
7. 负样本挖掘（Hard negative example mining）

对于每个首要框，都有一个边界框分类器来评估其内部含有对象的可能性。框匹配之后，所有其他首要框都为负。如果我们用了所有这些负样本，正负之间本会有明显的不平衡。可能的解决方案是：随机挑选负样本（FasterRCNN），或挑选那些分类器判断错误最严重的样本，这样负和正之间的比例大概是 3：1 。

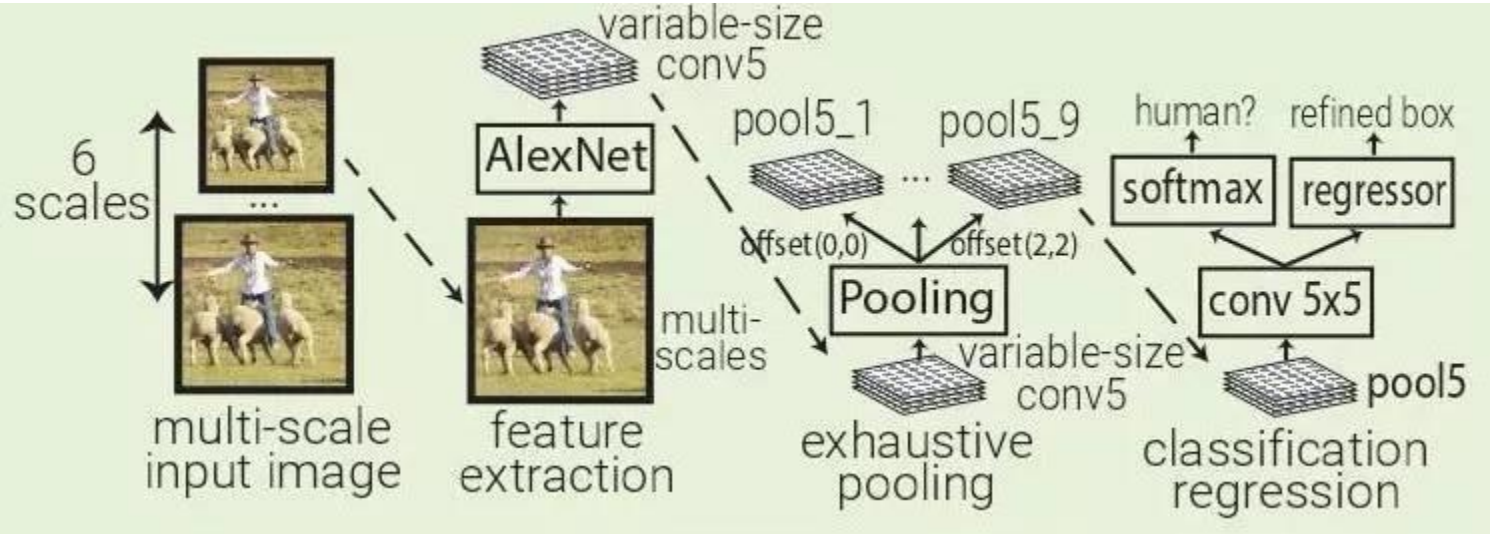
重要视觉模型发展：AlexNet→ZFNet →VGGNet →ResNet →MaskRCNN

一切从这里开始：现代物体识别随着 ConvNets 的发展而发展，这一切始于 2012 年 AlexNet 以巨大优势赢得 ILSVRC 2012。请注意，所有的物体识别方法都与 ConvNet 设计是正交的（任意 ConvNet 可以与任何对象识别方法相结合）。 ConvNets 用作通用图像特征提取器。

2012 年 AlexNet：AlexNet 基于有着数十年历史的 LeNet，它结合了数据增强、ReLU、dropout 和 GPU 实现。它证明了 ConvNet 的有效性，启动了 ConvNet 的光荣回归，开创了计算机视觉的新纪元。



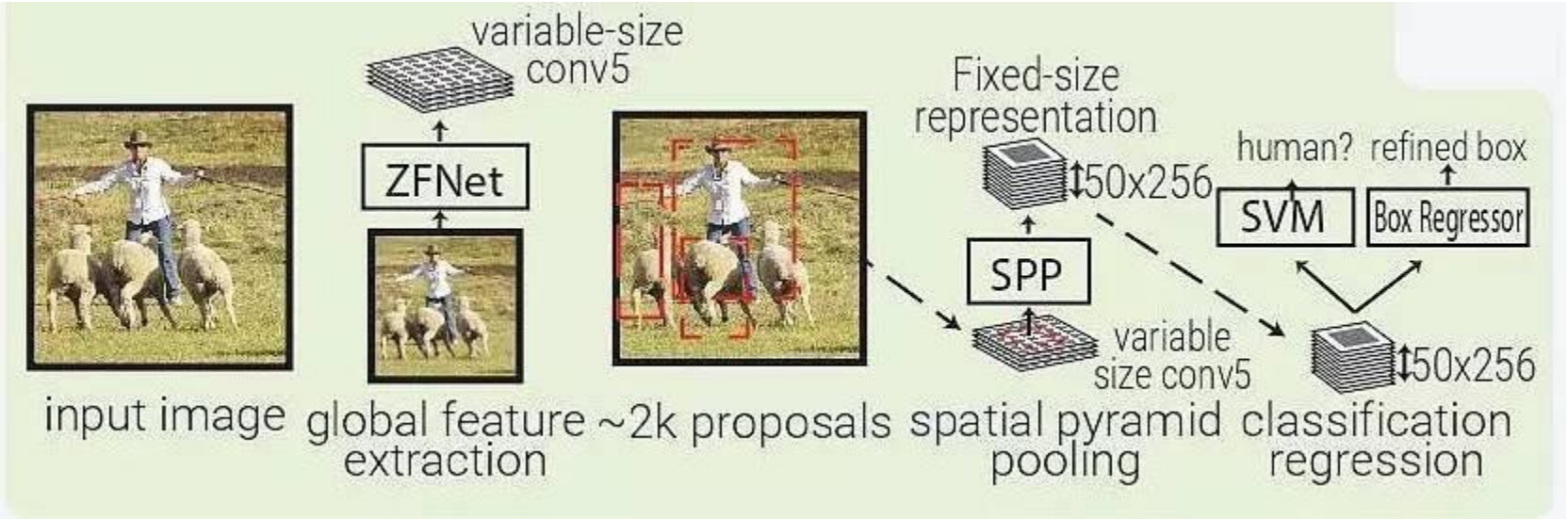
RCNN：基于区域的 ConvNet（RCNN）是启发式区域提案法（heuristic region proposal method）和 ConvNet 特征提取器的自然结合。从输入图像，使用选择性搜索生成约 2000 个边界框提案。这些被推出区域被裁剪并扭曲到固定大小的 227x227 图像。 然后，AlexNet 为每个弯曲图像提取 4096 个特征（fc7）。然后训练一个 SVM 模型，使用 4096 个特征对该变形图像中的对象进行分类。并使用 4096 个提取的特征来训练多个类别特定的边界框回归器来改进边界框。



OverFeat：OverFeat 使用 AlexNet 在一个输入图像的多个层次下的多个均匀间隔方形窗口中提取特征。训练一个对象分类器和一个类别不可知盒子回归器，用于对 Pool5 层（339x339 接收域窗口）中每 5x5 区域的对象进行分类并对边界框进行细化。OverFeat 将 fc 层替换为 1x1xN 的卷积层，以便能够预测多尺度图像。因为在 Pool5 中移动一个像素时，接受场移动 36 像素，所以窗口通常与对象不完全对齐。OverFeat 引入了详尽的池化方案：Pool5 应用于其输入的每个偏移量，这导致 9 个 Pool5 卷。窗口现在只有 12 像素而不是 36 像素。

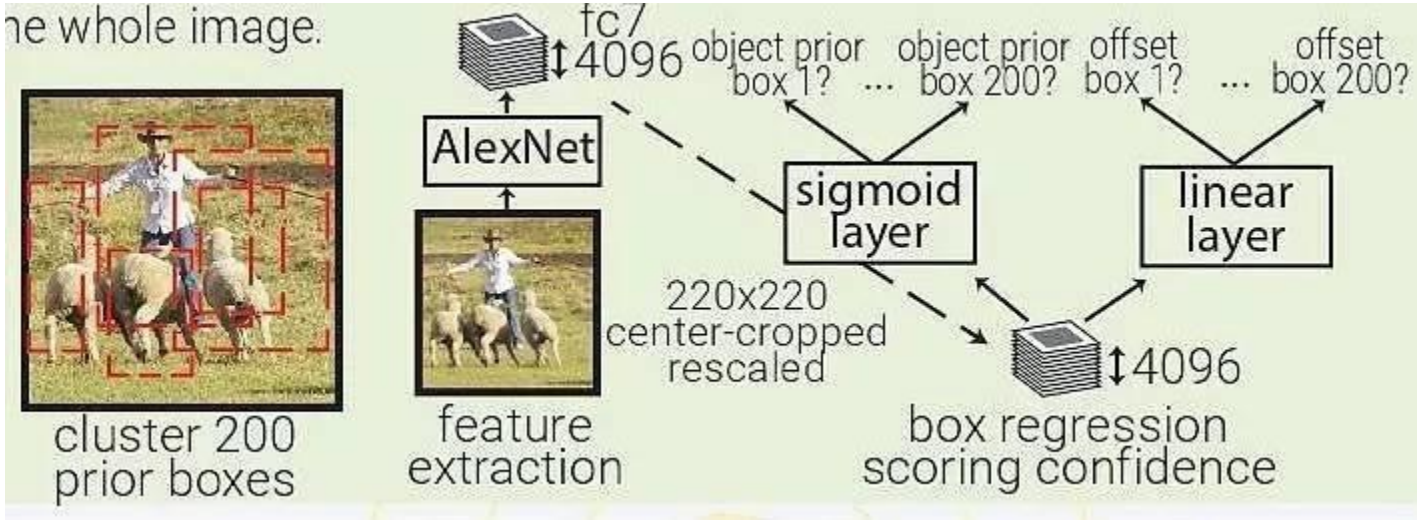
2013 年 ZFNet：ZFNet 是 ILSVRC 2013 的冠军得主，它实际上就是在 AlexNet 的基础上做了镜像调整（mirror modification）：在第一个卷积层使用 7×7 核而非 11×11 核保留了更多的信息。

SPPNet：SPPNet（Spatial Pyramid Pooling Net）本质上是 RCNN 的升级，SFFNet 引入了 2 个重要的概念：适应大小池化（adaptively-sized pooling，SPP 层），以及对特征量只计算一次。实际上，Fast-RCNN 也借鉴了这些概念，通过镜像调整提高了 RCNN 的速度。



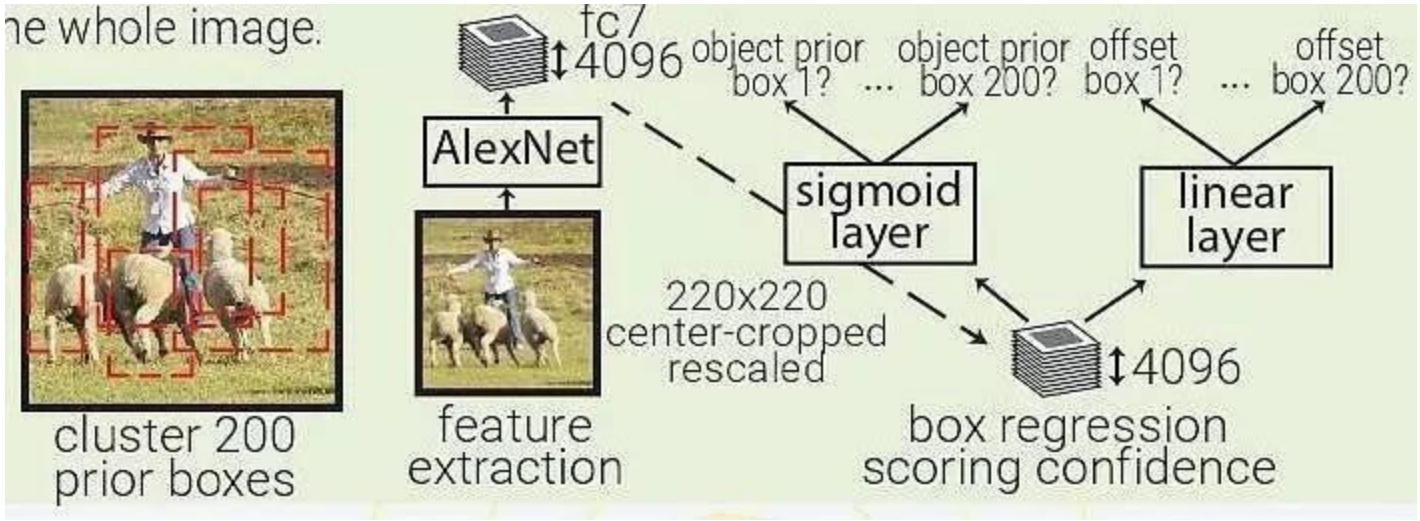
SPPNet 用选择性搜索在每张图像中生成 2000 个区域（region proposal）。然后使用 ZFNet-Conv5 从整幅图像中抓取一个共同的全体特征量。对于每个被生成的区域，SPPNet 都使用 spatial pyramid pooling（SPP）将该区域特征从全体特征量中 pool 出来，生成一个该区域的长度固定的表征。这个表征将被用于训练目标分类器和 box regressor。从全体特征量 pooling 特征，而不是像 RNN 那样将所有图像剪切

（crops）全部输入一个完整的 CNN，SPPNet 让速度实现了 2 个数量级的提升。需要指出，尽管 SPP 运算是可微分的，但作者并没有那么做，因此 ZFNet 仅在 ImageNet 上训练，没有做 finetuning。

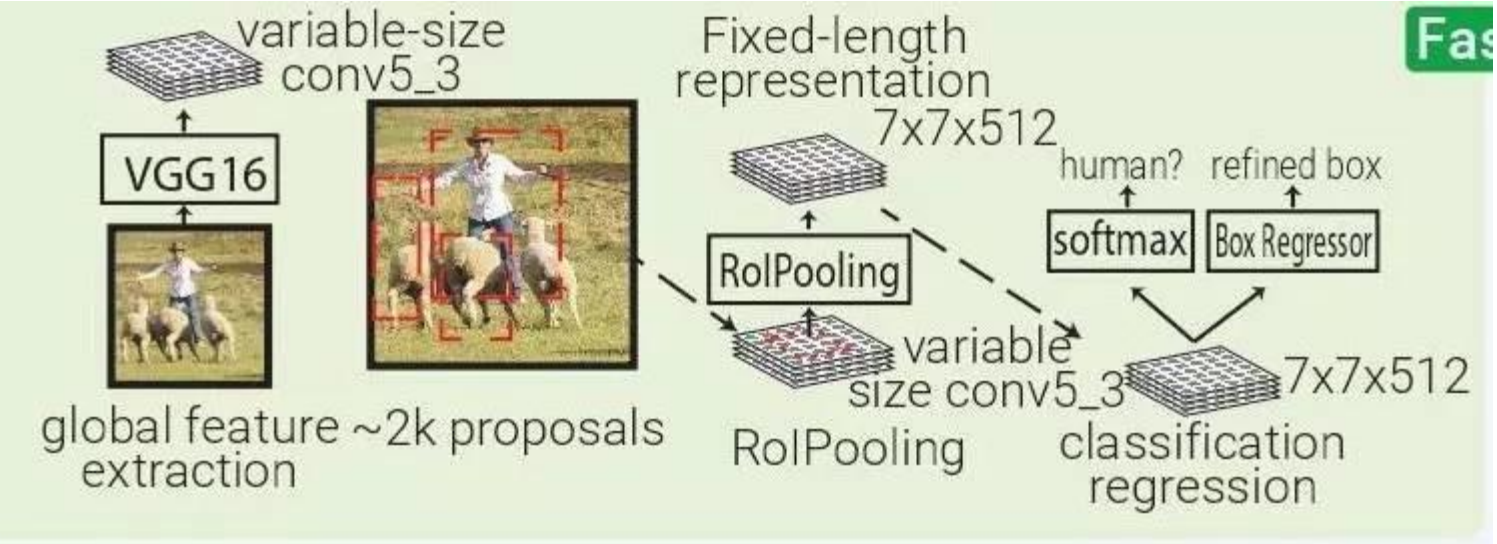


MultiBox：MultiBox 不像是目标识别，更像是一种基于 ConvNet 的区域生成解决方案。MultiBox 让区域生成网络（region proposal network，RPN）和 prior box 的概念流行了起来，证明了卷积神经网络在训练后，可以生成比启发式方法更好的 region proposal。自此以后，启发式方法逐渐被 RPN 所取代。MultiBox 首先将整个数据集中的所有真实 box location 聚类，找出 200 个质心（centroid），然后用将其用于 prior box 的中心。每幅输入的图像都会被从中心被裁减和重新调整大小，变为 220×220。然后，MultiBox 使用 ALexNet 提取 4096 个特征（fc7）。再加入一个 200-sigmoid 层预测目标置信度分数，另外还有一个 4×200-linear 层从每个 prior box 预测 centre offset 和 box proposal。注意下图中显示的 box regressors 和置信度分数在看从整幅图像中抓取的特征。

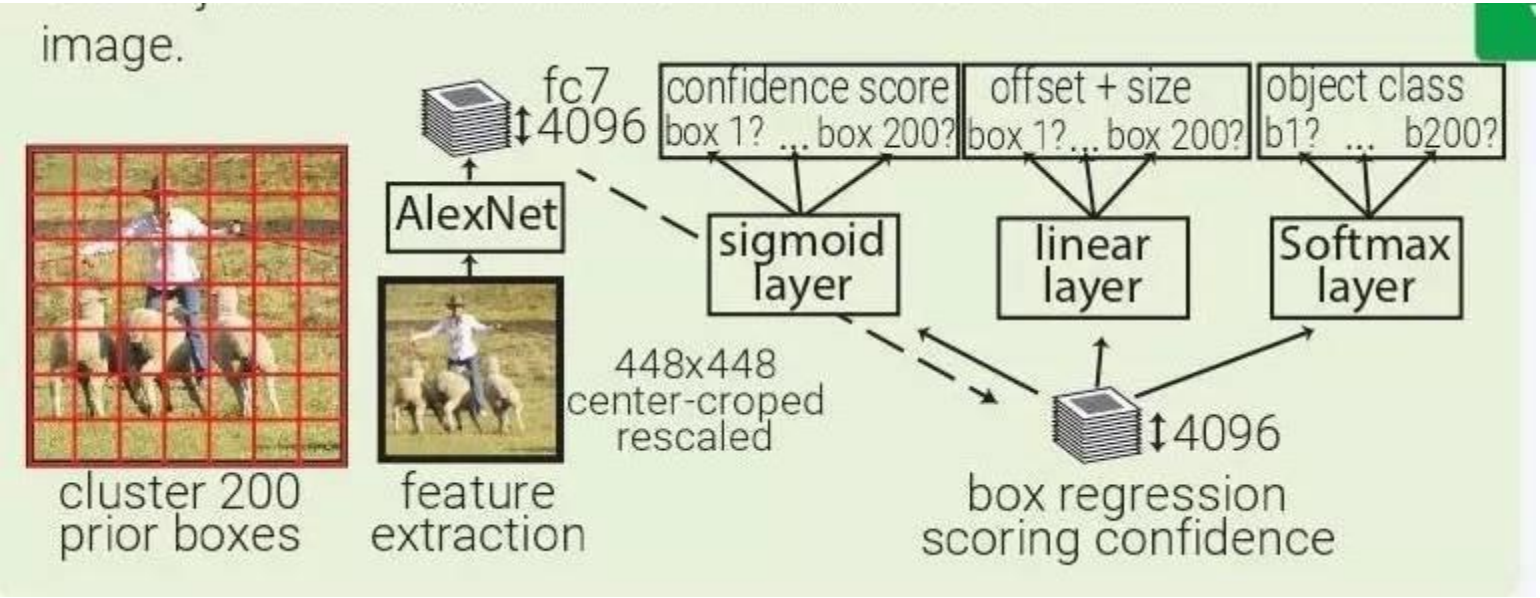
2014 年 VGGNet：虽然不是 ILSVRC 冠军，VGGNet 仍然是如今最常见的卷积架构之一，这也是因为它简单有效。VGGNet 的主要思想是通过堆叠多层小核卷积层，取代大核的卷积层。VGGNet 严格使用 3×3 卷积，步长和 padding 都为 1，还有 2×2 的步长为 2 的 maxpooling 层。



2014 年 Inception：Inception（GoogLeNet）是 2014 年 ILSVRC 的冠军。与传统的按顺序堆叠卷积和 maxpooling 层不同，Inception 堆叠的是 Inception 模块，这些模块包含多个并行的卷积层和许多核的大小不同的 maxpooling 层。Inception 使用 1×1 卷积层减少特征量输出的深度。目前，Inception 有 4 种版本。

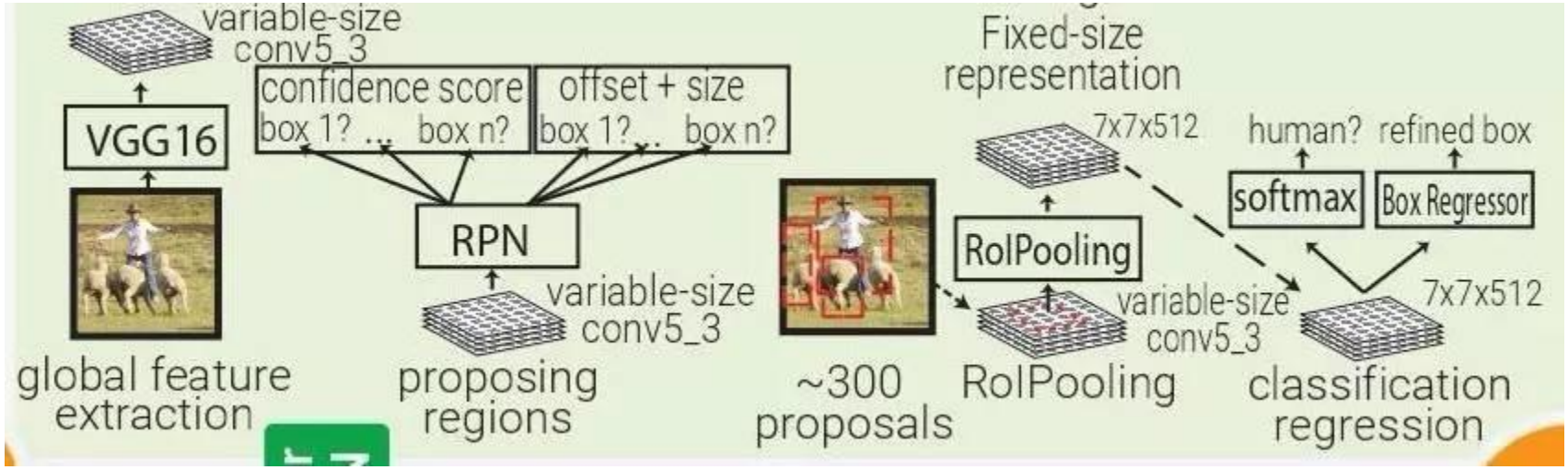


Fast RCNN：Fast RCNN 本质上 SPPNet，不同的是 Fast RCNN 带有训练好的特征提取网络，用 RoIPooling 取代了 SPP 层。

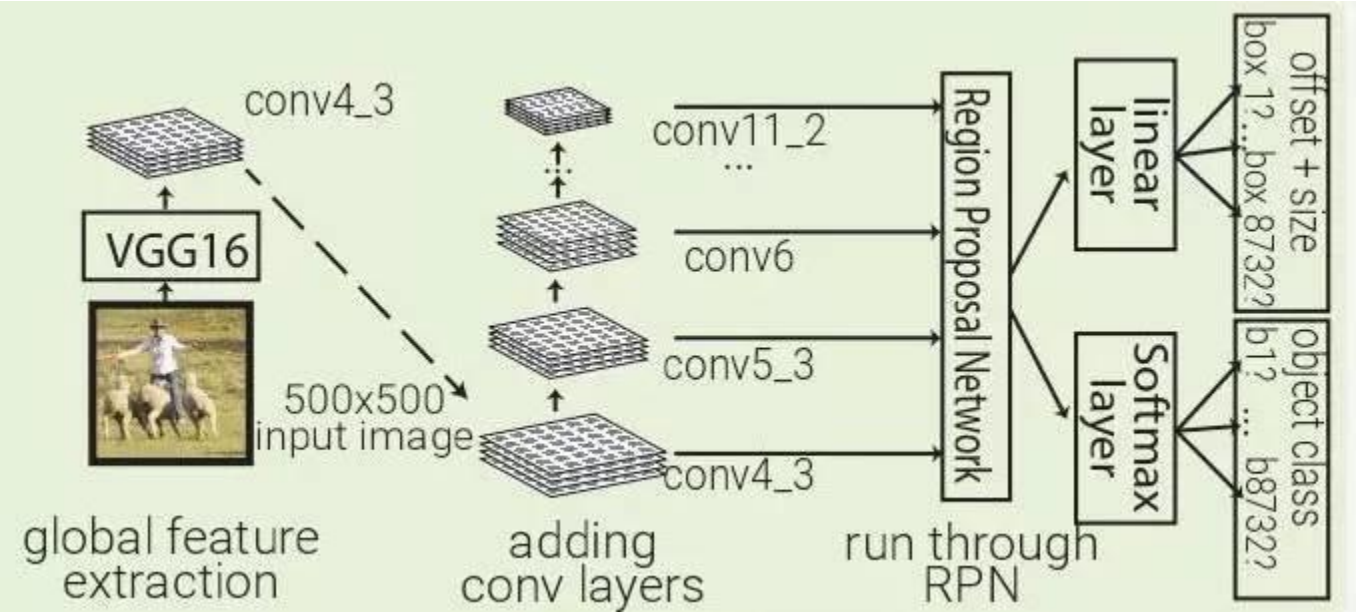


YOLO：YOLO（You Only Look Once）是由 MultiBox 直接衍生而来的。通过加了一层 softmax 层，与 box regressor 和 box 分类器层并列，YOLO 将原本是区域生成的 MultiBox 转为目标识别的方法，能够直接预测目标的类型。

2015 ResNet：ResNet 以令人难以置信的 3.6%的错误率（人类水平为 5-10%）赢得了 2015 年 ILSVRC 比赛。ResNet 不是将输入表达式转换为输出表示，而是顺序地堆叠残差块，每个块都计算它想要对其输入的变化（残差），并将其添加到其输入以产生其输出表示。这与 boosting 有一点关。



Faster RCNN：受 Multibox 的启发，Faster RCNN 用启发式区域生成代替了区域生成网络（RPN）。在 Faster RCNN 中，PRN 是一个很小的卷积网络（3×3 conv → 1×1 conv → 1×1 conv）在移动窗口中查看 conv5_3 全体特征量。每个移动窗口都有 9 个跟其感受野相关的 prior box。PRN 会对每个 prior box 做 bounding box regression 和 box confidence scoring。通过结合以上三者的 loss 成为一个共同的全体特征量，整个管道可以被训练。注意，在这里 RPN 只关注输入的一个小的区域；prior box 掌管中心位置和 box 的大小，Faster RCNN 的 box 设计跟 MultiBox 和 YOLO 的都不一样。



2016 年 SSD：SSD 利用 Faster RCNN 的 RPN，直接对每个先前的 box 内的对象进行分类，而不仅仅是对对象置信度（类似于 YOLO）进行分类。通过在不同深度的多个卷积层上运行 RPN 来改善前一个 box 分辨率的多样性。

2017 年 Mask RCNN：通过增加一支特定类别对象掩码预测，Mask RCNN 扩展了面向实例分割的 Faster RCNN，与已有的边界框回归量和对象分类器并行。由于 RolPool 并非设计用于网络输入和输出间的像素到像素对齐，MaskRCNN 用 RolAlign 取代了它。RolAlign 使用了双线性插值来计算每个子窗口的输入特征的准确值，而非 RolPooling 的最大池化法。

参考文献