问题

ROI Pooling 和 ROI Align 的区别是什么

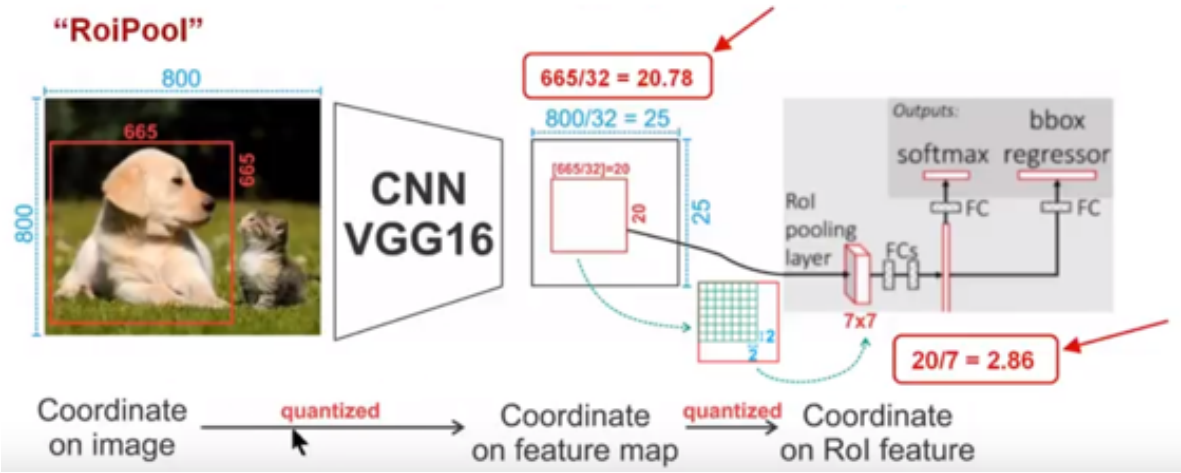
ROI Pooling 和 ROI Align 是什么

如果你对目标检测网络 Faster R-CNN 和实例分割网络 Mask R-CNN 网络比较熟悉的话，那你应该也对这个话题非常熟悉。

在区域建议网络 RPN 得到候选框 ROI 之后，需要提取该 ROI 中的固定数目的特征（例如Faster R-CNN 中的 7*7 ）输入到后面的分类网络以及边界回归网络的全连接层中。Faster R-CNN中使用的方法是 ROI Pooling，而对于像素位置精细度要求更高的 Mask R-CNN 对 ROI Pooling 进行改进，变成 ROI Align。

ROI Pooling和ROIAlign最大的区别是：前者使用了两次量化操作，而后者并没有采用量化操作，使用了双线性插值算法，具体的解释如下所示。

ROI Pooling 技术细节



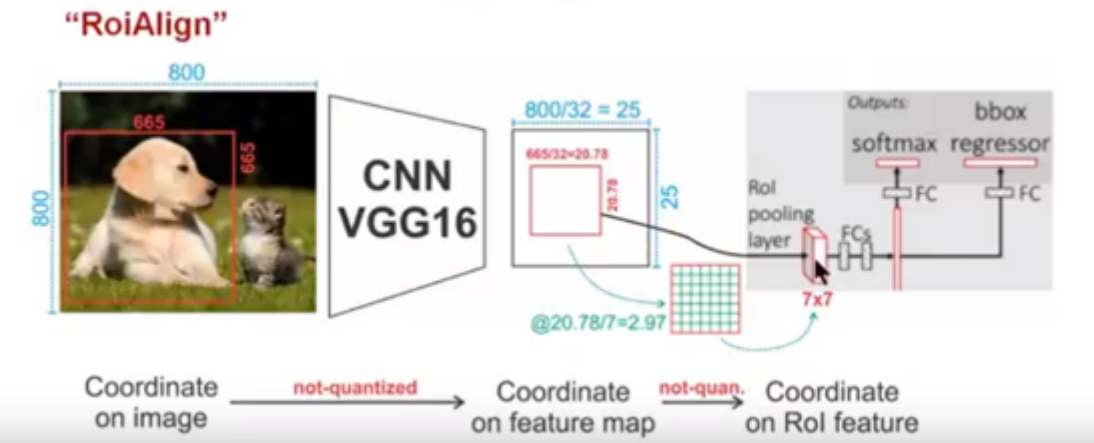
如上图所示，为了得到固定大小（7X7）的feature map，我们需要做两次量化操作：

1. 图像坐标 — feature map坐标
2. feature map坐标 — ROI feature坐标。

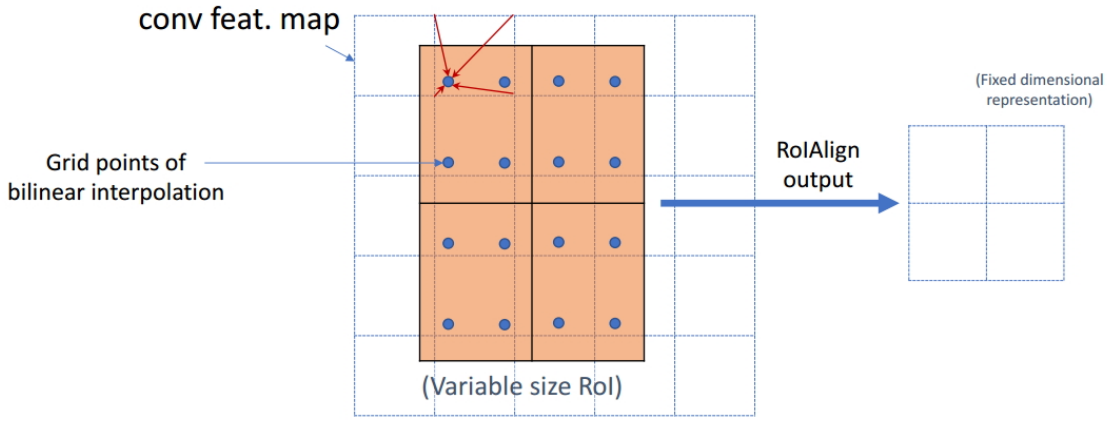
我们来说一下具体的细节，如图我们输入的是一张800x800的图像，在图像中有两个目标（猫和狗），狗的BB大小为665x665，经过VGG16网络后，我们可以获得对应的feature map，如果我们对卷积层进行Padding操作，我们的图片经过卷积层后保持原来的大小，但是由于池化层的存在，我们最终获得feature map 会比原图缩小一定的比例，这和Pooling层的个数和大小有关。在该VGG16中，我们使用了5个池化操作，每个池化操作都是2Pooling，因此我们最终获得feature map的大小为800/32 x 800/32 = 25x25（是整数），但是将狗的BB对应到feature map上面，我们得到的结果是665/32 x 665/32 = 20.78 x 20.78，结果是浮点数，含有小数，但是我们的像素值可没有小数，那么作者就对其进行了量化操作（即取整操作），即其结果变为20 x 20，在这里引入了第一次的量化误差；然而我们的feature map中有不同大小的ROI，但是我们后面的网络却要求我们有固定的输入，因此，我们需要将不同大小的ROI转化为固定的ROI feature，在这里使用的是7x7的ROI feature，那么我们需要将20 x 20的ROI映射成7 x 7的ROI feature，其结果是 20 /7 x 20/7 = 2.86 x 2.86，同样是浮点数，含有小数点，我们采取同样的操作对其进行取整吧，在这里引入了第二次量化误差。其实，这里引入的误差会导致图像中的像素和特征中的像素的偏差，即将feature空间的ROI对应到原图上面会出现很大的偏差。原因如下：比如用我们第二次引入的误差来分析，本来是2,86，我们将其量化为2，这期间引入了0.86的误差，看起来是一个很小的误差呀，但是你要记得这是在feature空间，我们的feature空间和图像空间

是有比例关系的，在这里是1:32，那么对应到原图上面的差距就是0.86 x 32 = 27.52。这个差距不小吧，这还是仅仅考虑了第二次的量化误差。这会大大影响整个检测算法的性能，因此是一个严重的问题。

ROI Align 技术细节



如上图所示，为了得到为了得到固定大小（7X7）的feature map，ROIAlign技术并没有使用量化操作，即我们不想引入量化误差，比如665 / 32 = 20.78，我们就用20.78，不用什么20来替代它，比如20.78 / 7 = 2.97，我们就用2.97，而不用2来代替它。这就是ROIAlign的初衷。那么我们如何处理这些浮点数呢，我们的解决思路是使用"双线性插值"算法。双线性插值是一种比较好的图像缩放算法，它充分的利用了原图中虚拟点（比如20.56这个浮点数，像素位置都是整数值，没有浮点值）四周的四个真实存在的像素值来共同决定目标图中的一个像素值，即可以将20.56这个虚拟的位置点对应的像素值估计出来。如下图所示，蓝色的虚线框表示卷积后获得的feature map，黑色实线框表示ROI feature，最后需要输出的大小是2x2，那么我们就利用双线性插值来估计这些蓝点（虚拟坐标点，又称双线性插值的网格点）处所对应的像素值，最后得到相应的输出。这些蓝点是2x2Cell中的随机采样的普通点，作者指出，这些采样点的个数和位置不会对性能产生很大的影响，你也可以用其它的方法获得。然后在每一个橘红色的区域里面进行max pooling或者average pooling操作，获得最终2x2的输出结果。我们的整个过程中没有用到量化操作，没有引入误差，即原图中的像素和feature map中的像素是完全对齐的，没有偏差，这不仅会提高检测的精度，同时也会有利于实例分割。这么细心，做科研就应该关注细节，细节决定成败。



参考资料

Mask R-CNN详解