

9.5 多尺度目标检测

在9.4节（锚框）中，我们在实验中以输入图像的每个像素为中心生成多个锚框。这些锚框是对输入图像不同区域的采样。然而，如果以图像每个像素为中心都生成锚框，很容易生成过多锚框而造成计算量过大。举个例子，假设输入图像的高和宽分别为561像素和728像素，如果以每个像素为中心生成5个不同形状的锚框，那么一张图像上则需要标注并预测200多万个锚框（ $561 \times 728 \times 5$ ）。

减少锚框个数并不难。一种简单的方法是在输入图像中均匀采样一小部分像素，并以采样的像素为中心生成锚框。此外，在不同尺度下，我们可以生成不同数量和不同大小的锚框。值得注意的是，较小目标比较大目标在图像上出现位置的可能性更多。举个简单的例子：形状为 1×1 、 1×2 和 2×2 的目标在形状为 2×2 的图像上可能出现的位置分别有4、2和1种。因此，当使用较小锚框来检测较小目标时，我们可以采样较多的区域；而当使用较大锚框来检测较大目标时，我们可以采样较少的区域。

为了演示如何多尺度生成锚框，我们先读取一张图像。它的高和宽分别为561像素和728像素。

```
%matplotlib inline
from PIL import Image
import numpy as np
import torch

import sys
sys.path.append("../")
import d2lzh_pytorch as d2l

img = Image.open('../doimg/catdog.jpg')
w, h = img.size # (728, 561)
```

我们在5.1节（二维卷积层）中将卷积神经网络的二维数组输出称为特征图。我们可以通过定义特征图的形状来确定任一图像上均匀采样的锚框中心。

下面定义 `display_anchors` 函数。我们在特征图 `fmap` 上以每个单元（像素）为中心生成锚框 `anchors`。由于锚框 `anchors` 中 x 和 y 轴的坐标值分别已除以特征图 `fmap` 的宽和高，这些值域在0和1之间的值表达了锚框在特征图中的相对位置。由于锚框 `anchors` 的中心遍布特征图 `fmap` 上的所有单元，`anchors` 的中心在任一图像的空间相对位置一定是均匀分布的。具体来说，当特征图的宽和高分别设为 `fmap_w` 和 `fmap_h` 时，该函数将在任一图像上均匀采样 `fmap_h` 行 `fmap_w` 列个像素，并分别以它们为中心生成大小为 `s`（假设列表 `s` 长度为1）的不同宽高比（`ratios`）的锚框。

```
d2l.set_figsize()
```

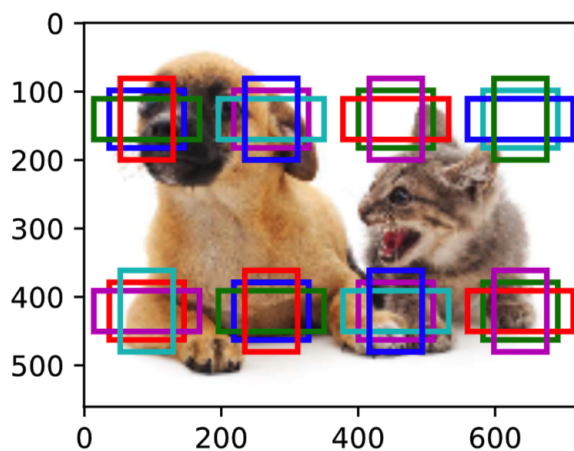
```
def display_anchors(fmap_w, fmap_h, s):
    # 前两维的取值不影响输出结果(原书这里是(1, 10, fmap_w, fmap_h), 我认为错了)
    fmap = torch.zeros((1, 10, fmap_h, fmap_w), dtype=torch.float32)

    # 平移所有锚框使均匀分布在图片上
    offset_x, offset_y = 1.0/fmap_w, 1.0/fmap_h
    anchors = d2l.MultiBoxPrior(fmap, sizes=s, ratios=[1, 2, 0.5]) + \
        torch.tensor([offset_x/2, offset_y/2, offset_x/2, offset_y/2])

    bbox_scale = torch.tensor([[w, h, w, h]], dtype=torch.float32)
    d2l.show_bboxes(d2l.plt.imshow(img).axes,
                    anchors[0] * bbox_scale)
```

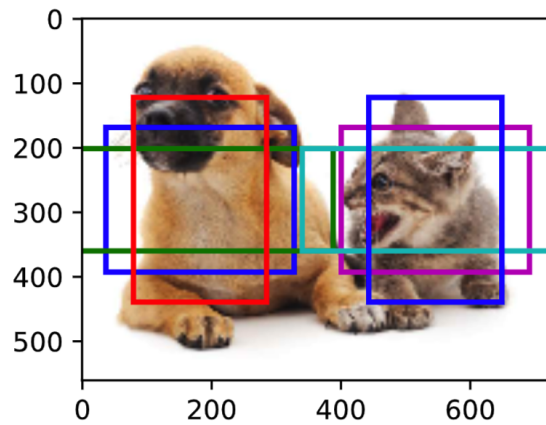
我们先关注小目标的检测。为了在显示时更容易分辨，这里令不同中心的锚框不重合：设锚框大小为0.15，特征图的高和宽分别为2和4。可以看出，图像上2行4列的锚框中心分布均匀。

```
display_anchors(fmap_w=4, fmap_h=2, s=[0.15])
```



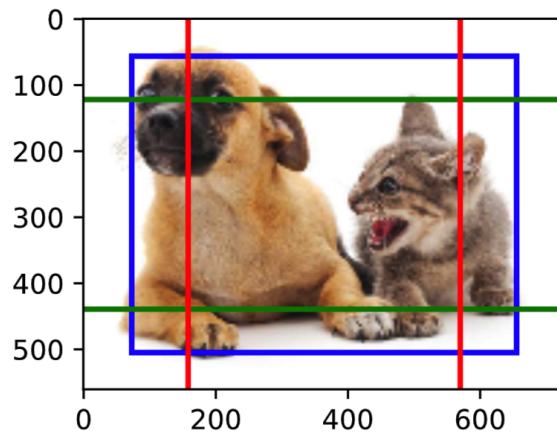
我们将特征图的高和宽分别减半，并用更大的锚框检测更大的目标。当锚框大小设0.4时，有些锚框的区域有重合。

```
display_anchors(fmap_w=2, fmap_h=1, s=[0.4])
```



最后，我们将特征图的宽进一步减半至1，并将锚框大小增至0.8。此时锚框中心即图像中心。

```
display_anchors(fmap_w=1, fmap_h=1, s=[0.8])
```



既然我们已在多个尺度上生成了不同大小的锚框，相应地，我们需要在不同尺度下检测不同大小的目标。下面我们来介绍一种基于卷积神经网络的方法。

在某个尺度下，假设我们依据 c 张形状为 $h \times w$ 的特征图生成 $h \times w$ 组不同中心的锚框，且每组的锚框个数为 a 。例如，在刚才实验的第一个尺度下，我们依据10（通道数）张形状为 4×2 的特征图生成了8组不同中心的锚框，且每组含3个锚框。接下来，依据真实边界框的类别和位置，每个锚框将被标注类别和偏移量。在当前的尺度下，目标检测模型需要根据输入图像预测 $h \times w$ 组不同中心的锚框的类别和偏移量。

假设这里的 c 张特征图为卷积神经网络根据输入图像做前向计算所得的中间输出。既然每张特征图上都有 $h \times w$ 个不同的空间位置，那么相同空间位置可以看作含有 c 个单元。根据5.1节（二维卷积层）中感受野的定义，特征图在相同空间位置的 c 个单元在输入图像上的感受野相同，并表征了同一感受野内的输入图像信息。因此，我们可以将特征图在相同空间位置的 c 个单元变换为以该位置为中心生成的 a 个锚框的类别和偏移量。不难发现，本质上，我们用输入图像在某个感受野区域内的信息来预测输入图像上与该区域位置相近的锚框的类别和偏移量。

当不同层的特征图在输入图像上分别拥有不同大小的感受野时，它们将分别用来检测不同大小的目标。例如，我们可以通过设计网络，令较接近输出层的特征图中每个单元拥有更广阔的感受野，从而检测输入图

像中更大尺寸的目标。

我们将在9.7节（单发多框检测（SSD））中具体实现一个多尺度目标检测的模型。

小结

- 可以在多个尺度下生成不同数量和不同大小的锚框，从而在多个尺度下检测不同大小的目标。
- 特征图的形状能确定任一图像上均匀采样的锚框中心。
- 用输入图像在某个感受野区域内的信息来预测输入图像上与该区域相近的锚框的类别和偏移量。