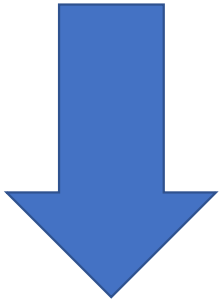
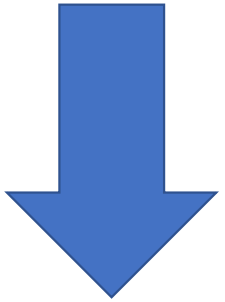
Complete Explanation

- https://www.youtube.com/watch?v=LZRfHkTNQqo

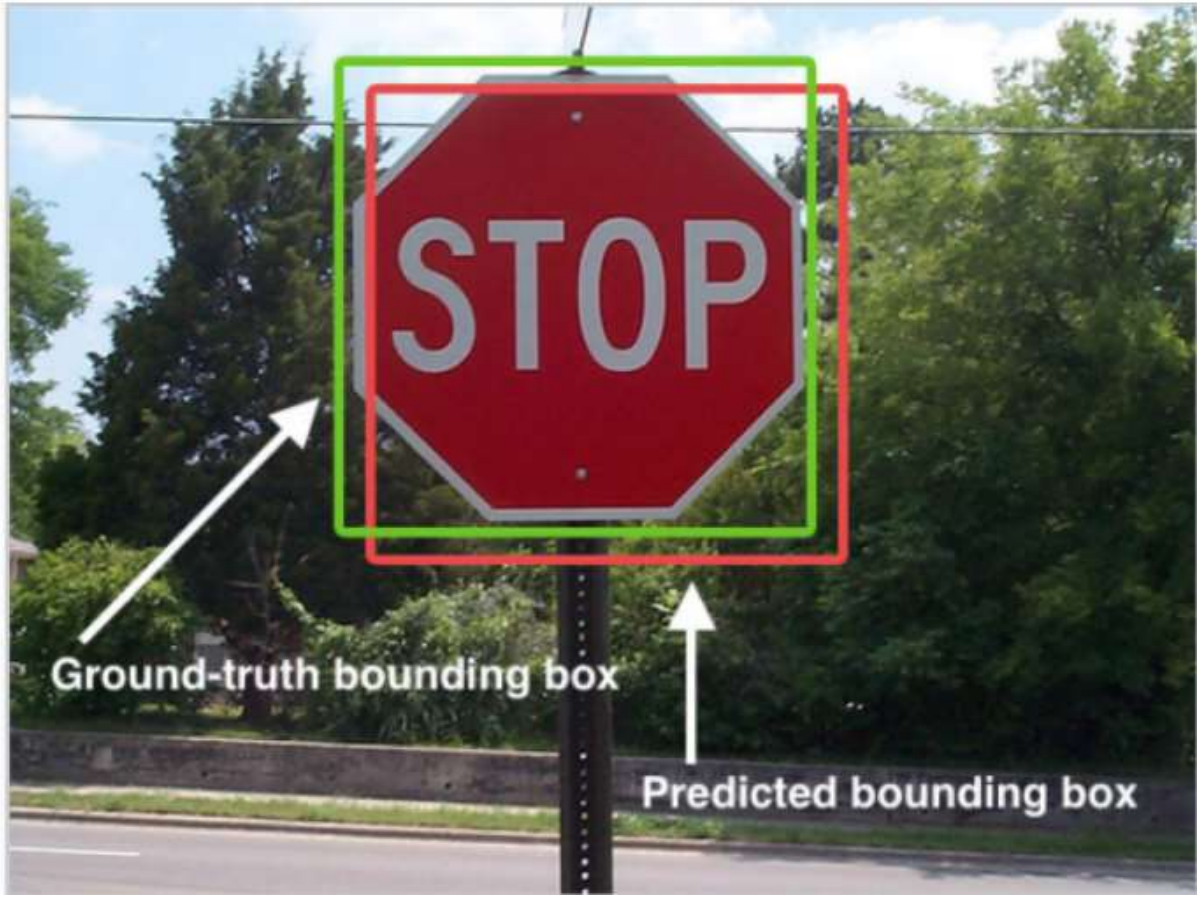
Single Shot Detection (SSD) + MobileNet

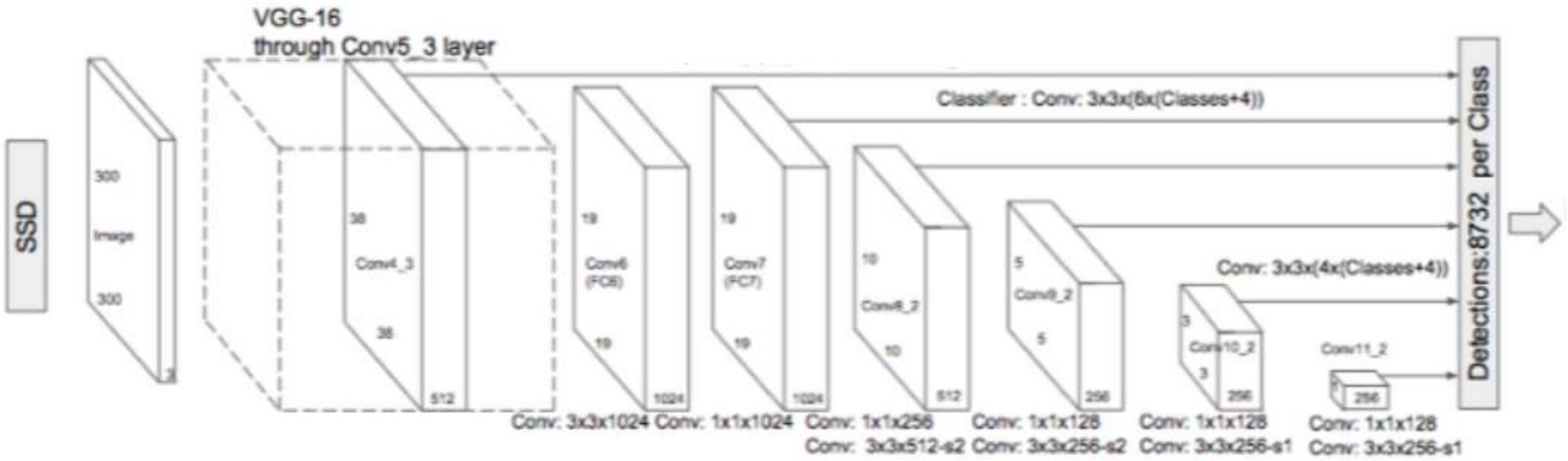




Localization

Classification


STOP
Ground-truth bounding box
Predicted bounding box



(from conv6 onwards) were added, thus enabling to extract features at

Depth-wise Separable Convolutions Architecture

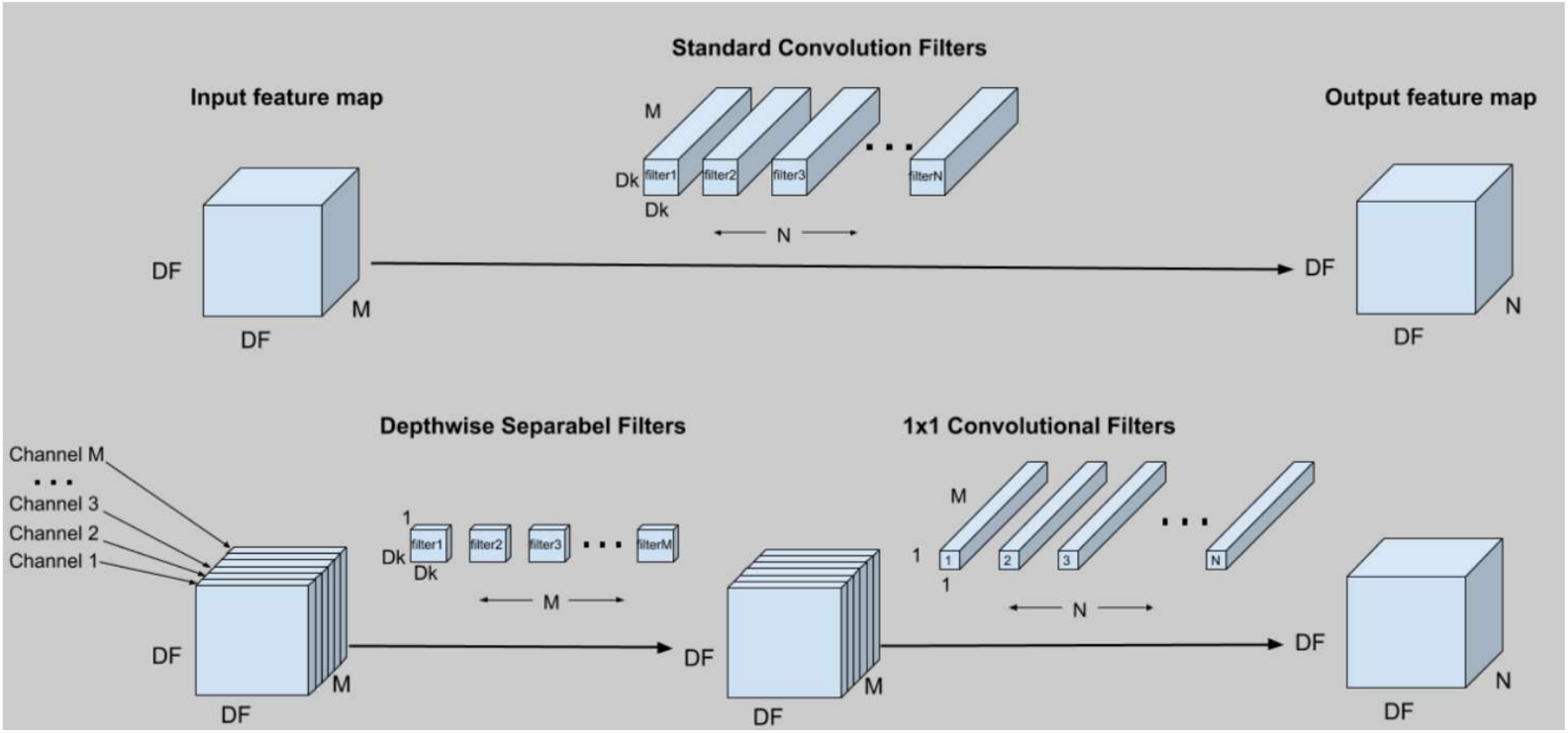


Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

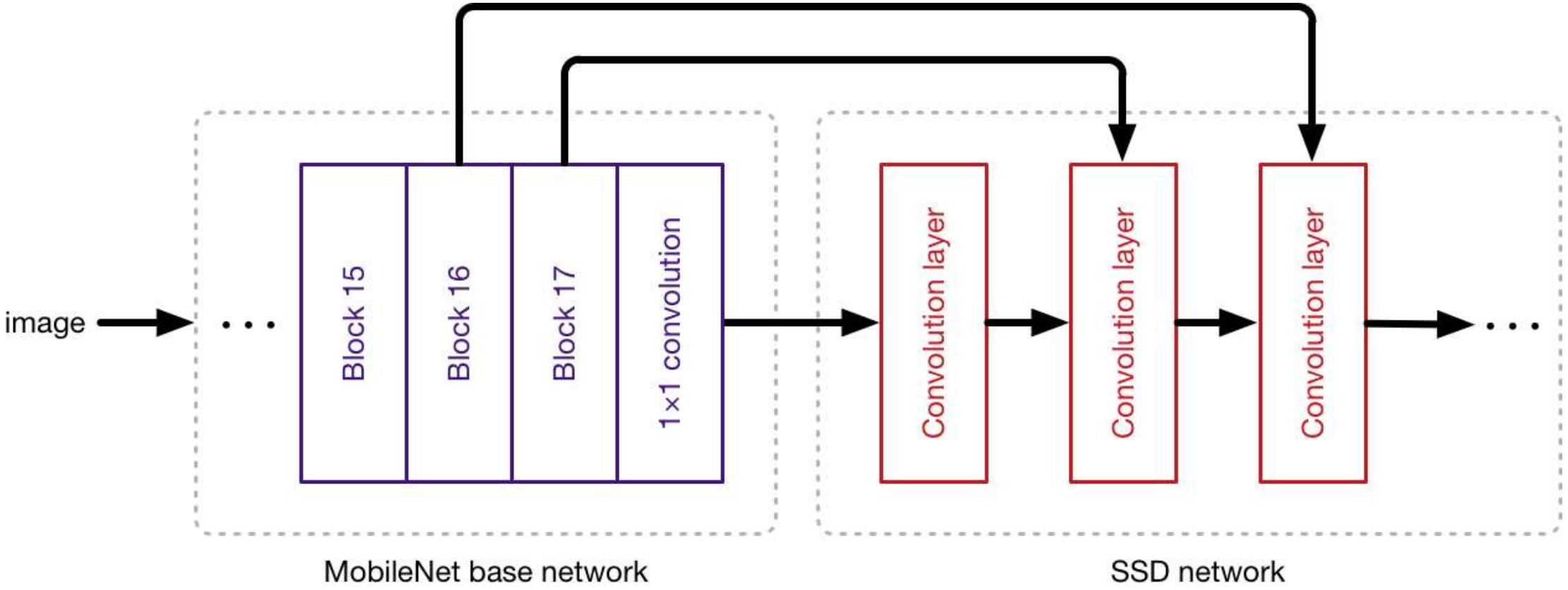


image
Block 15
Block 16
Block 17
1×1 convolution
Convolution layer
Convolution layer
Convolution layer
MobileNet base network
SSD network

What Input and Output for object detection module ?

Object Detection

- Output format:

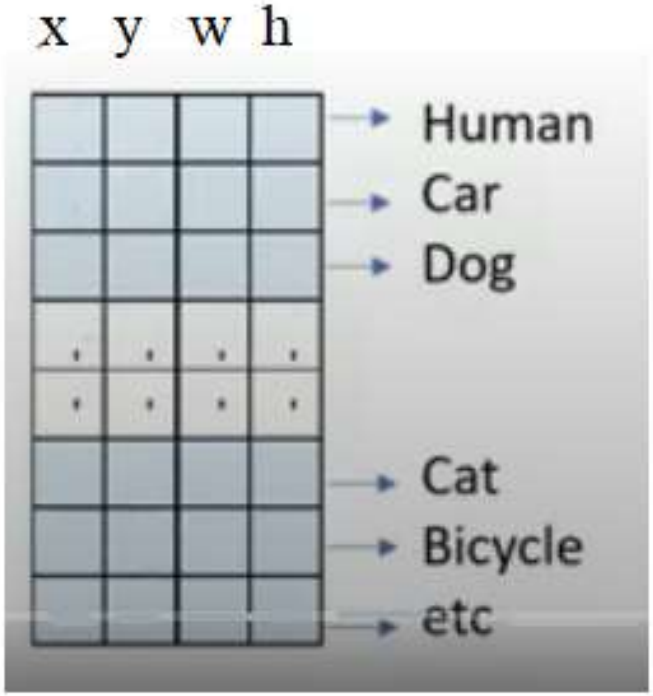
`x, y, w, h` are between [0, 1]. You can scale `x, w` by the width and `y, h` by the height of the image.

```
{
  detectedClass: "hot dog",
  confidenceInClass: 0.123,
  rect: {
    x: 0.15,
    y: 0.33,
    w: 0.80,
    h: 0.27
  }
}
```

SSD MobileNet:

- Run on image:

```
var recognitions = await Tflite.detectObjectOnImage(
  path: filepath,        // required
  model: "SSDMobileNet",
  imageMean: 127.5,
  imageStd: 127.5,
  threshold: 0.4,        // defaults to 0.1
  numResultsPerClass: 2,// defaults to 5
  asynch: true           // defaults to true
);
```


x y w h
Human
Car
Dog
Cat
Bicycle
etc