基于多头注意力的目标检测算法

摘要

本文提出了一种基于多头注意力的目标检测算法，名为MANet。它的主要目的是通过注意力机制将不同尺度特征层进行融合，增强上下文联系。为了达到这个目的，我们首先将Single Shot Detector（SSD）[1]的前馈基础网络替换为Residual-101（受Deconvolutional Single Shot Detector（DSSD）[2]启发），然后通过线性插值及注意力机制将不同尺度特征的信息进行融合，以提高目标检测的精度，特别是对于特殊背景类别的物体及小目标进行精度优化。本文的首要贡献是提出了一种融合注意力机制，其次是提出了一种多头的注意力融合方法。我们的最终模型MAnet检测器有效地统一了不同尺度之间的特征信息，使其能够以更高的精度检测不同尺寸的物体。我们采用320x320输入的MAnet在VOC 2007[3]测试中得到了80.1%的平均精度均值（mean average precision，mAP），68帧每秒（FramePerSecond，FPS）的实验结果，此结果表明MAnet优于其它以Feature Pyramid Network（FPN）[4]为代表的单阶段检测器。

1介绍

目标检测作为一个长期存在的基本又具有挑战性的问题，几十年来一直是计算机视觉研究领域的热点。目标检测的目的是确定给定的图像中是否存在给定类别的任何实例。给定一张图片，如果存在需要检测的对象，则返回每个对象实例的空间位置和范围（例如通过边界框）。作为图像理解和计算机视觉的基石，目标检测是更复杂或更高级别视觉任务（例如物体跟踪、图像捕获、实例分割等）的基础。此外，目标检测在人工智能和信息技术等领域也有广泛的应用，包括机器视觉，自动驾驶，人机交互等。

深度学习的发展极大地促进了视觉自动化和其他许多领域的发展。最近，通过深度学习从数据中自动学习特征表示的方法有效提高了目标检测的性能。神经网络是深度学习的基础，因此如何设计更好的神经网络已成为一个对性能影响的关键问题。近期的基于卷积神经网络（Convolutional Neural Network, CNN）的目标检测器可以分为两种：一种是两阶段检测器，例如 （Region-based CNN）R-CNN[5]、Region-based Fully Convolutional Networks (R-FCN) [6]、Feature Pyramid Network（FPN）[4]，另一种是单阶段检测器，例如 You only look once （YOLO）[7]、Single Shot Detector（SSD）[1] 和 RetinaNet[8]。前者先由算法生成一系列作为样本的候选框，再通过CNN进行样本分类；后者则不用产生候选框，直接将目标边框定位问题转化为回归问题处理。

在上述各种目标检测器中，为了在不牺牲太多精度的情况下保持实时速度，Liu等人[1]提出SSD，比YOLO快并且具有当时最先进的基于区域的目标检测器的准确性。SSD将YOLO的回归思想和Faster R-CNN[9]的anchor box机制结合，在不同卷积层的特征图上预测物体区域，输出离散化的多尺度、多比例的default boxes坐标，同时利用小卷积核预测一系列候选框的边框坐标补偿和每个类别的置信度。在整幅图像上各个位置用多尺度区域的局部特征图边框回归，保持了YOLO算法快速特性，同时也保证了边框定位效果和Faster R-CNN[9]类似。但SSD利用多层次特征分类，最后一个卷积层的感受野范围很大，使得小目标特征不明显。为了解决这些问题，在本文中，我们提出了一种新的单阶段检测架构MAnet，聚合了不同尺度的特征信息。我们的MAnet在PASCAL VOC 2007 上实现了80.1%的平均精度均值（mean average precision，mAP），同时保持了约68帧每秒（FramePerSecond，FPS）的快速推理速度。

我们工作的贡献可以归纳如下：

1. 提出了一个名为MAnet的框架，它融合不同尺度的特征信息，以获得更好的性能。
2. 为了融合不同尺度的特征信息，我们提出了一个新的注意力机制取名为融合注意力。
3. 为了更好地将多个不同尺度的特征信息融合，我们设计了不同的多头融合模块以生成更有效的特征表达。

2相关工作

在卷积神经网络出现之前，早期的目标检测方法通常是基于滑动窗口，例如Support Vector Machines with Histogram of Oriented Gradients（HOG-SVM）[10]，Deformable Part Model （DPM）[11]等。它们大多数被广泛应用于将感兴趣区域（Region Of Interest，ROIs）分类为各种类别。DPM[11]是其中最常用的方法之一，该模型由Felzenszwalb等人在2008年提出，正如其名，它是一种基于组件的检测算法,也是当时最好的算法,蝉联了三届PASCAL VOC目标检测冠军。它的基本思想是先提取DPM人工特征，再用latentSVM分类。这种特征提取方式存在明显的局限性：首先，DPM特征计算复杂，计算速度慢；其次，人工特征对于旋转、拉伸、视角变化的物体检测效果差。这些弊端很大程度上限制了算法的应用场景。为了解决这个问题，Ross Girshick al等人提出了一种基于深度神经网络的目标检测方法R-CNN[5],它大大超过了以往方法的性能，之后几乎所有最优的目标检测方法都是基于卷积神经网络。两阶段检测器和单阶段检测器是目前主流的两种目标检测方法。

两阶段检测器：其中Spatial Pyramid Pooling（SPPNet）[12],Fast R-CNN [13],Faster R-CNN[9] 和 R-FCN[6]是两阶段方法的一些代标性例子。它们大多都采用卷积网络的最顶层来检测不同尺度大小的物体。这些方法首先使用单独的生成器生成候选对象集合，例如Selective Search [14],Edge Boxes [15]和Region Proposal Network（RPN）。其中大多数是基于超像素合并（例如Constrained Parametric Min-Cuts（CPMC）[16]等）或基于滑窗算法（例如Edghe Boxes[15]等），它们的共同特点是主要用CNN作为分类器，不预测对象边界。尽管这些方法大大提高了检测精确度，但是计算量大，速度较慢。

单阶段检测器：考虑到两阶段检测器速度慢的缺点，单阶段检测器（例如YOLO[7],SSD[1]）引起了人们的关注。这些检测器通过使用一组不同尺度的固定尺寸的anchors覆盖整个图像，代替region proposal阶段。在具有高分辨率特征的浅层卷积层中检测小对象，而在具有低分辨率特征的深层卷积层中检测大对象。因此，单阶段检测器与Faster-RCNN[9]相比，可以用更小的输入尺寸来提取丰富的特征以降低计算成本。但是，这一类方法都没有注意每个位置的局部信息，并且较低卷积层中的信息并没有得到充分利用，因此，尽管单阶段方法比两阶段方法计算成本低，但是精确度仍然落后于两阶方法。

为了改善两阶段检测器和单阶段检测器的性能，学界已经提出了各种策略将多尺度信息聚合到目标检测器。Multi-scale CNN（MSCNN）[17]利用各种分辨率的特征图来检测不同尺度的物体，从而产生一组可变的感受野大小用来覆盖许多物体尺寸。Rainbow SSD（RSSD）[18]通过池化（pooling）加反卷积融合不同层的特征，不仅加强了不同层特征图之间的关系同时也增加了不同层特征图的数量，一定程度上解决了小物体检测的问题。Deconvolutional Single Shot Detector（DSSD）[2]使用额外的反卷积层增加特征映射层的分辨率并融合上下文信息。StairNet[19]引入了一个组合特征模块，它以自上而下的方式增强上下文语义信息，进一步推断了组合信息。

与以往工作不同，我们提出了一种新的架构，以实现不同层的全局上下文信息融合，从而提高目标检测的性能。

3方法

在本节中，我们首先对多头融合特征的概念进行解释，然后对MANet的各个添加的模块进行详细的介绍。

### 3.1多头融合特征

最近一些网络模型通过将特征图与邻近层组合在一起，来提高检测的精确度。但是，邻近层提供的信息不足以使当前层检测各种比例的对象。因此，我们提出了多头融合特征的概念，通过多头注意力集成不同大小的特征图，获取融合特征，以提取更准确的信息。

Luo等人[20]指出有效的感受野大小远小于理论上的感受野大小，因此SSD模型在小物体检测上性能表现较差，我们猜想SSD无法较好地检测小物体的原因是低层特征层的感受野较小，因此可以通过融合不同深度不同大小的特征图信息，增加有效的感受野，加强上下文联系，以提高检测性能。所以，我们提出了一种多头注意力，用来融合来自不同层不同大小分辨率的特征图的信息，弥补当前层的信息，从而提高检测的性能。

### 3.2MANet模型

在本小节中，我们首先回顾SSD结构；然后为了提高训练效率，MANet基础网络采用与DSSD[2]相同的策略，用残差网络代替原始SSD中使用的Visual Geometry Group（VGG）[21]网络；最后我们讨论如何通过添加融合注意力来创建多头网络，获得最终MANet模型的语义上下文信息。

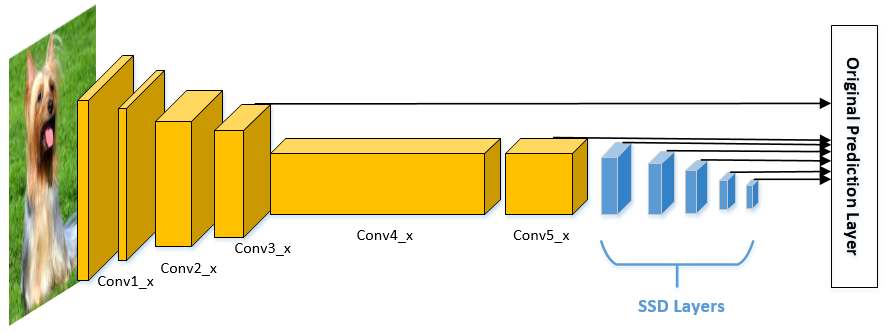


图1原始的SSD

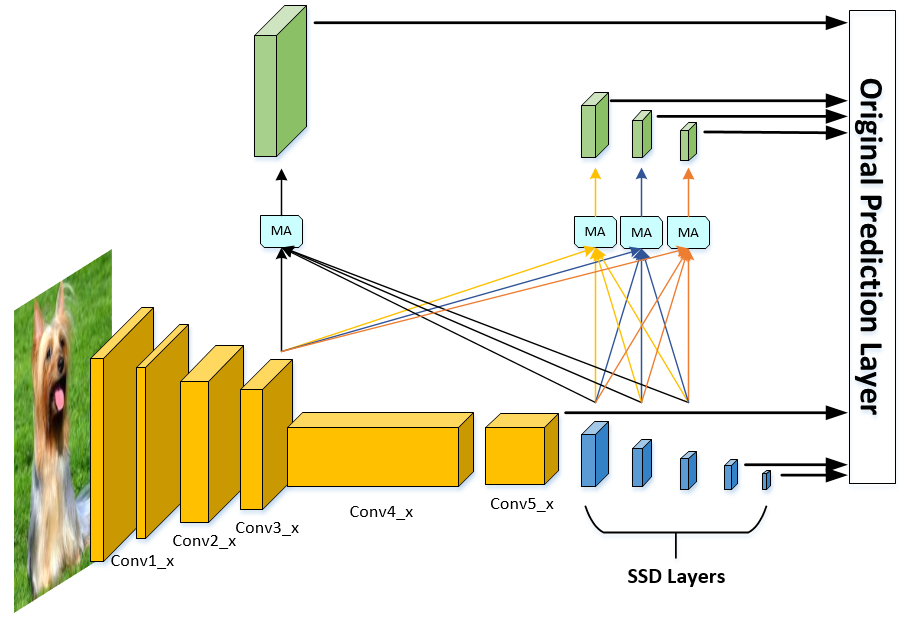


图2 MAnet模型

#### 3.2.1 SSD

Single Shot MultiBox Detector（SSD）[1]是 Wei Liu等人在2016年欧洲计算机视觉大会（ECCV2016）上提出的一种目标检测算法。它沿用了YOLO中直接回归 Bounding box（bbox）和分类概率的方法，同时又参考了Faster R-CNN[9]，大量使用anchor来提升识别准确度。通过把这两种结构相结合，SSD保持了很高的识别速度，还能把平均精度均值提升到较高的水平。同时，采用了特征金字塔结构进行预测检测结果，即检测时利用了conv4-3、conv-7（FC7）、conv6-2、conv7-2、conv8\_2、conv9\_2这些大小不同的特征图（feature maps），在多个特征图上同时进行softmax分类和位置回归。同时，它使用非最大抑制（Non Maximum Suppression，NMS）对预测后的结果进行处理，去掉冗余的检测框，以获得最终的检测结果。更多内容可以在[1]找到，其中使用VGG作为基础网络。

#### 3.2.2 Resnet-SSD

为了更好地训练网络模型，我们采用与DSSD[2]一样的策略，第一个修改是用ResNet[22]代替原始SSD中使用的VGG网络，但是我们选择的特征图尺度大小与SSD选择的特征图尺度一样。具体见表1所示。图1是用Resnet-101[22]作为基础网络的Resnet-SSD，如图1所示在Resnet-SSD中我们选择conv3\_x、conv5\_x、conv6\_x、conv7\_x、conv8\_x、conv9\_x输入到预测模块中，与SSD一样我们额外地添加了一系列的卷积特征层在基础网络的末端。这些特征层在尺寸上会逐渐减小，并允许在多个尺度上对检测结果进行预测。虽然Resnet-101层数比VGG -16层数深，但是通过实验结果得知，将SSD的基础卷积网络替换成残差网络并不会提高它的准确度反而有所下降，接下来我们将逐步添加不同的模块提高检测的准确度。

#### 3.2.3 Fusion attention

在目标检测的任务中，直观上可以理解，底层特征层的尺度往往较大所以对于小物体的检测效果要好，而越深层的特征层具有更精炼的语义信息但是对于小物体的检测效果并不是很好。因此，不同尺度特征之间关系对目标检测精度具有重要影响，不同尺度特征关系旨在捕获不同尺度之间依赖关系，增强各自特征表达能力。例如，DSSD[2]只整合了最邻近的特征信息，这是远远不够的。为了获得不同尺度的特征信息，以提供更全面的检测信息，本文引入了一个融合注意力模块，该模块主要利用任意不同尺度特征之间的关联，来互相增强各自特征的表达。接下来，我们详细介绍融合注意力模块的工作过程。

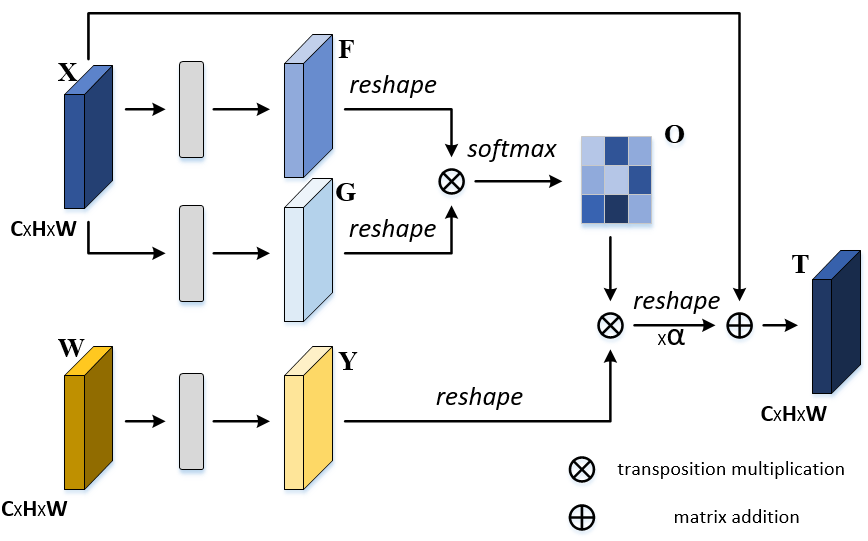


图3 Fusion attention

如图3所示，给定特征图,首先将其输入到具有批量归一化和ReLU层的卷积层，生成两个新的特征F和G，其中,然后将它们重新整形为,其中是特征的数量。然后,将特征F和特征G转置相乘，再经过softmax操作归一化获得该尺度特征的注意力图:

同时，再给定另一个特征图，将特征W送到具有批量归一化和ReLU层的卷积层，以生成新的特征,然后将其整形为,然后将注意力图O与特征Y转置相乘，并将结果重新整形为,最后，将它乘以一个尺度参数，并用特征X进行和运算，得到最终输出如下：

其中初始化为0并逐渐学会分配更多的权重，可以从上述公式推断出最后特征T是融合两种不同特征图的结果。

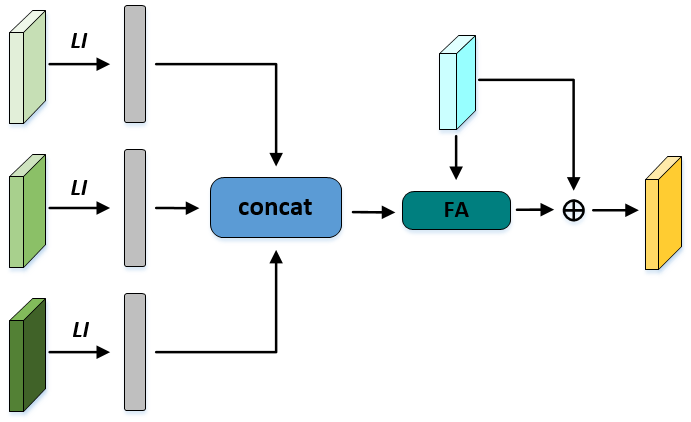
因此，融合注意力模块，具有融合不同特征之间的能力，并且根据融合注意力选择性地聚合特征。实现了特征之间的相互增益，更加有利于目标检测。

#### 3.2.4 Multiple模块

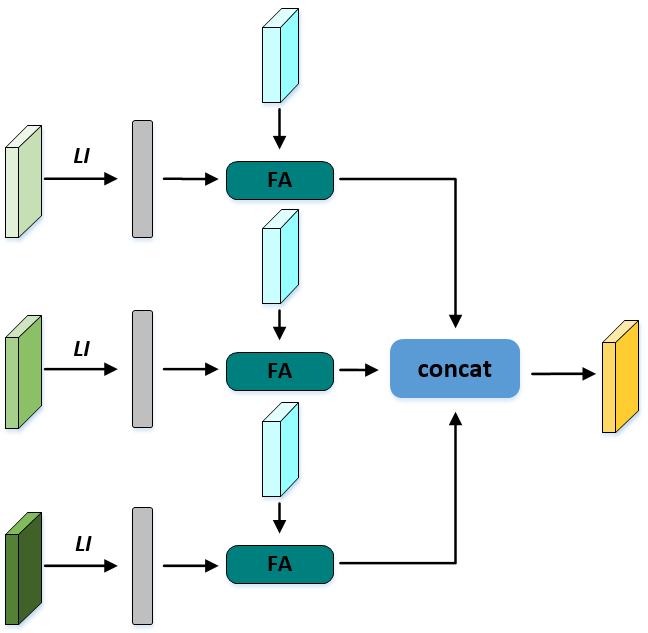
不同特征分辨率大小的特征图，包含着不同的特征信息，SSD[1]指出多尺度信息对于目标检测精度具有重要的影响。以往的做法通常直接将不同尺度的特征信息直接输入到预测模块，但是不同尺度各自包含的信息不足以有效的检测对象，因此我们通过融合其它不同尺度特征图的信息，弥补当前层的信息，以提高检测精确度。本文构建了一个多头注意力模块，用来融合多个不同尺度之间的信息，在全局上增强不同尺度各自的特征表达能力，以提高目标检测的精度。

为了减少计算量，我们在多头模块中舍弃了两个小比例的尺度（1x1,3x3），直接将这两个尺度输入到最后的预测模块中。如图所示，以5x5特征图为例，我们分别设计了三种不同的多头融合模型对各个尺度的特征进行融合。为了方便运算，我们将不同尺度的特征进行整合：首先将不同尺度的特征图统一到同一个尺度上，为了减少运算量，直接通过双线性插值完成统一尺度的操作(双线性插值不引入参数，计算量少)，然后通过1x1卷积将通道统一到同一维度。三个多头模型具体细节如下：

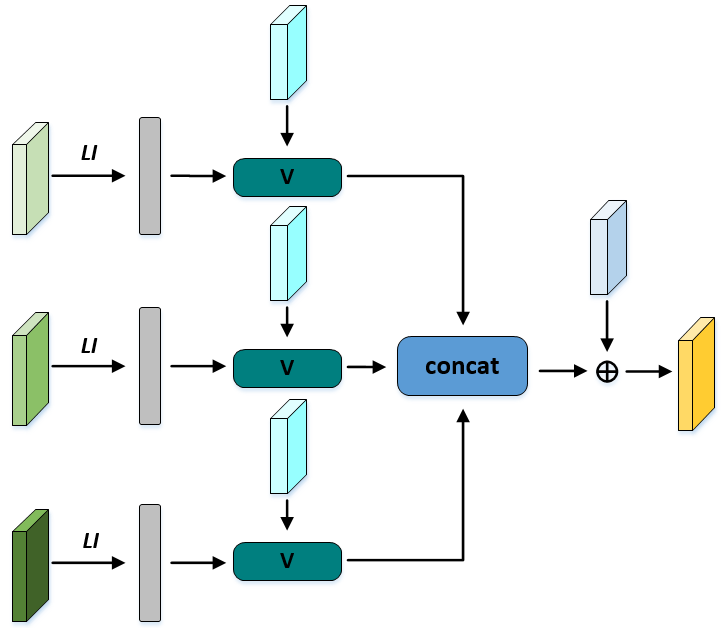
1. 模型a把整合后的特征张量进行concat操作，然后直接输入到Fusion attention模块中，得到最后融合结果。具体公式如下，其中是三个特征concat后的新的特征结果，FA是指Fusion attention 模块。
2. 模型b把整合后的特征张量，分别输入到Fusion attention模块中，然后将输出的结果进行concat操作得到最终的结果。公式如下所示，其中，是指不同尺度的特征分别与当前的特征尺度通过融合注意力进行融合得到新的特征，将,进行concat操作得到最终的结果。
3. 模型c与之前两个模型有一定的区别，没有直接用融合注意力，而是首先分别计算了不同尺度对应的,然后将不同尺度计算得到的结果进行 concat操作再与本层特征尺度进行相加得到终的结果。公式如下所示，其中就是融合注意力中的。



MAnet-a



MANet-b



MAnet-c

**图4** Multiple模块模型

## 4实验

我们在PASCAL VOC [3]数据集上评估我们的方法，该数据集具有20个对象类别。 对于PASCAL VOC[3]，按照协议，我们在VOC2007训练验证集和VOC 2012 训练验证集上对模型进行联合训练以及在VOC 2007测试集上进行测试。我们使用平均精度均值（mean average precision，mAP）对模型进行评估。为了提高实验效率，我们用Resnet-50[22]作为消融实验中的基础卷积神经网络。

### 4.1 Base network

我们最终实验的模型选择Resnet-101[22]作为特征提取网络，它在Imagenet[23]数据集上进行预训练。我们将conv\_5阶段的有效步幅从32个像素更改为16个像素，以增加特征图分辨率。基于SSD拟合残差，我们进一步引入残差模块(residual block)来添加额外的特征层，减小特征层的尺寸。

表I基础网络图

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| VGG | conv4\_3 | conv7 | conv8\_2 | conv9\_2 | conv10\_2 | conv11\_2 |
| Resolution  Depth | 3838  13 | 1919  20 | 1010  22 | 55  24 | 33  26 | 1  27 |
|  |  |  |  |  |  |  |
| Residual-101 | conv3\_x | conv5\_x | conv6\_x | conv7\_x | conv8\_x | conv9\_x |
| Resolution  Depth | 4040  23 | 2020  101 | 1010  104 | 55  107 | 33  110 | 1  113 |

表I显示了在原始VGG架构和resnet-101中选择的特征层，深度（Depth）是指所选特征层在网络中的位置（只考虑卷积层和池化层）。其中要注意的是，这两个网络中第一个预测层的深度。尽管Resnet-101包含101层，但是我们需要用密集的特征层去预测较小的目标，因此我们只能选择conv3\_x中的最后一个特征层作为第一个预测层。如果我们只考虑内核大小大于1的，这个数字将一直到9，意味着这一层的神经元的感受野可能是小于VGG中conv4\_3神经元的感受野。与Resnet-101[22]其它层相比，该层的特征表达能力较弱，预测性能较差。

### 4.2 Pascal VOC 2007

在这个实验中，由于内存限制，我们设置批量为32来训练模型，与原始SSD [1]模型相同，我们使用0.0001的权重衰减和0.9的动量。首先采用初始学习率为0.001的随机梯度下降（stochastic gradient descent，SGD）[23]优化器，但是由于损失急剧波动，使得训练过程不那么稳定。 为了解决这个问题，我们通过在前2次epoch（一次epoch指所有数据送入网络中完成一次前向计算及反向传播的过程）中以0.001的学习率训练模型来使用预热策略。 在预热阶段之后，它返回到原始学习速率计划，其在150 epoch和200 epoch时分别减少10倍。 训练到第250 epoch终止。

表II PASCAL VOC2007 测试结果

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Method | network | map | areo | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbk | person | plant | sheep | sofa | train | tv |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| HyperNet [25] | VGGNet | 76.3 | 77.4 | 83.3 | 75.0 | 69.1 | 62.4 | 83.1 | 87.4 | 87.4 | 57.1 | 79.8 | 71.4 | 85.1 | 85.1 | 80.0 | 79.1 | 51.2 | 79.1 | 75.7 | 80.9 | 76.5 |
| Faster [9] | ResNet101 | 76.4 | 79.8 | 80.7 | 76.2 | 68.3 | 55.9 | 85.1 | 85.3 | 89.8 | 56.7 | 87.8 | 69.4 | 88.3 | 88.9 | 80.9 | 78.4 | 41.7 | 78.6 | 79.8 | 85.3 | 72.0 |
| R-FCN [6] | ResNet101 | 80.5 | 79.9 | 87.2 | 81.5 | 72.0 | 69.8 | 86.8 | 88.5 | 89.8 | 67.0 | 88.1 | 74.5 | 89.8 | 90.6 | 79.9 | 81.2 | 53.7 | 81.8 | 81.5 | 85.9 | 79.9 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| SSD321[2] | ResNet101 | 77.1 | 76.3 | 84.6 | 79.3 | 64.6 | 47.2 | 85.4 | 84.0 | 88.8 | 60.1 | 82.6 | 76.9 | 86.7 | 87.2 | 85.4 | 79.1 | 50.8 | 77.2 | 82.6 | 87.3 | 76.6 |
| SSD300 [1] | VGGNet | 77.5 | 79.5 | 83.9 | 76.0 | 69.6 | 50.5 | **87.0** | 85.7 | 88.1 | 60.3 | 81.5 | 77.0 | 86.1 | 87.5 | 83.9 | 79.4 | 52.3 | 77.9 | 79.5 | 87.6 | 76.8 |
| DSSD321 [2] | ResNet101 | 78.6 | 81.9 | 84.9 | **80.5** | 68.4 | 53.9 | 85.6 | 86.2 | 88.9 | 61.1 | 83.5 | **78.7** | 86.7 | **88.7** | 86.7 | 79.7 | 51.7 | 78.0 | 80.9 | 87.2 | 79.4 |
| StairNet [19] | VGGNet | 78.8 | 81.3 | 85.4 | 77.8 | 72.1 | **59.2** | 86.4 | **86.8** | 87.5 | 62.7 | **85.7** | 76.0 | 84.1 | 88.4 | 86.1 | 78.8 | 54.8 | 77.4 | 79.0 | 88.3 | 79.2 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| MANet300(Ours)  MANet321(Ours) | ResNet101  ResNet101 | **79.5**  **80.1** | 82.3  **85.1** | 86.5 **87.1** | 76.6 80.2 | 72.5  **73.1** | 52.4  57.5 | 86.2  86.5 | 86.3  86.7 | **89.3**  88.0 | **64.2**  **64.2** | 84.4  83.7 | 77.0  76.0 | **88.5**  87.8 | 87.6  87.0 | **86.8**  86.4 | 80.1  **81.6** | 57.0  **58.3** | **82.3**  82.0 | 82.1  **82.3** | **88.4**  88.0 | 78.6 **80.3** |

表(II)显示了我们在PASCAL VOC2007测试集上检测的结果。通过将VGGNet[21]替换成Resnet-101[22]，在性能上低于原始的SSD，这可能和特征层的表达能力有关。原始的SSD仅使用不同深度的特征图进行预测，而不会融合低级和高级特征，我们用原始的SSD作为我们实验的基线，从表中得知我们在原始SSD[1]的基础上提高了2个百分点。同时MAnet与其它多尺度融合方法相比（如StairNet[19]和DSSD[2]），也具有一定的提高，这证明了MAnet的有效性。

综上所述，MAnet在测试任务中显示出在具有特定背景和小对象目标上有巨大改进。例如飞机、船、羊这些类通常有非常特殊的背景：飞机的天空、羊的草地等，同时植物通常都很小。这表明SSD[1]中小物体检测的弱点可以通过所提出的MANet模型更好的解决，并且对于具有独特上下文的类来说，性能更好。

### 4.3 VOC2007消融研究

本节中，为了验证每个模块对性能的影响，我们在VOC2007数据集上设置不同的模型进行测试，为了提高实验的效率，我们用Resnet-50[22]作为模型的基础网络，并将测试结果记录在表3和表四中。

#### 4.3.1 Multiple 研究

为了更好地融合不同感受野大小的特征图，我们设计了三个不同的多头融合模块，我们通过实验从中选择最优的模型。为了保证实验结果的准确性，三个模型训练的参数都是一样的。模型a直接把除本层外其它不同尺度特征直接concat，输入到FA模块中得到新的特征图，然后输入到预测模块中得到预测结果。模型b是将不同尺度的特征图输入到FA模块然后将结果concat得到新的特征图；与前两个模型不同的是模型c首先分别计算除本层外不同尺度特征与本层特征对应的V,然后将计算结果进行concat操作再与原始该尺度特征图进行相加得到融合后的特征图。通过表（III）得知，模型c的性能表现最优，同时计算复杂度相对于其它两个模型而言更简单，因此我们选择模型c作为MANet的多头融合模块。

表III 多头模型比较结果

|  |  |  |
| --- | --- | --- |
| Method | network | map |
| MANet(a) | ResNet50 | 78.97 |
| MANet(b) | Resnet50 | 79.21 |
| MANet(c) | Resnet50 | 79.32 |

#### 4.3.2 不同尺度影响

为了验证我们添加到SSD多尺度融合特征的有效性，我们对模型进了不同的设置，并在表(IV)中记录它们的评估结果。通过实验，逐渐增加不同尺度特征进行融合(1x1,3x3尺度除外，它们直接输入到最后的预测模块)，我们从大的尺度特征开始进行逐步融合，从表中可以看出，大尺度特征层对性能影响比较大（38x38和19x19尺度的增加分别提高了1个百分点左右），我们猜想这可能是因为底层特征层中包含的信息量多于高层特征。同时，我们可以看到随着更多不同尺度特征层的融入，结果正在逐步改进，这证明了不同尺度特征信息的融合有利于检测性能的提升，进一步证明了我们模型的有效性。

表IV尺度消融实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| c |  |  |  |  |
| Resolution  Results | 5x5 | 10x10 | 19x19 | 38x38 |
|  |  |  |  |  |
|  |  |  |  |  |
| 79.32  78.71  78.61  77.00  76.67 | √ | √  √ | √  √  √ | √  √  √  √ |
|  |  |  |  |  |

## 5结论

在本文中，我们提出了一种多尺度特征信息融合的方法，并展示了它在基准数据集上的有效性。虽然我们期望找到更高效、更有效的方法取得更多的改进，但是实验证明我们的模型已经可以在PASCAL VOC数据集[3]上取得很大的提升。我们的MAnet模型能够超越以往的SSD框架，特别是在具有特定背景的物体上，同时仍然保持着与其它检测器相当的速度。

参考文献

[1] Liu W , Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. 2015.

[2] Fu C Y , Liu W , Ranga A , et al. DSSD : Deconvolutional Single Shot Detector[J]. 2017.

[3] Lin T Y , Dollár, Piotr, Girshick R , et al. Feature Pyramid Networks for Object Detection[J]. 2016.

[4] Everingham M , Eslami S M A , Van Gool L , et al. ThePascalVisual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1):98-136.

[5] Girshick R B, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. computer vision and pattern recognition, 2014: 580-587.

[6] Dai J , Li Y , He K , et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. 2016.

[7] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.

[8] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.

[9] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. 2015.

[10] Dalal N , Triggs B . Histograms of Oriented Gradients for Human Detection[C]// null. IEEE Computer Society, 2005.

[11] Felzenszwalb P F , Mcallester D A , Ramanan D . A Discriminatively Trained, Multiscale, Deformable Part Model[C]// 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE, 2008.

[12] Purkait P, Cheng Z, Zach C. SPP-Net: Deep Absolute Pose Regression with Synthetic Views[J]. 2017.

[13] Girshick R . Fast R-CNN[J]. Computer Science, 2015.

[14] Uijlings J R, Sande K E, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.

[15] Zitnick C L , Dollar P . Edge Boxes : Locating Object Proposals from Edges[C]// ECCV. Springer, Cham, 2014.

[16] Carreira J , Sminchisescu C . CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7):1312-1328.

[17] Cai Z , Fan Q , Feris R S , et al. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection[J]. 2016.

[18] Jeong J , Park H , Kwak N . Enhancement of SSD by concatenating feature maps for object detection[J]. 2017.

[19] Woo S , Hwang S , Kweon I S . StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection[J]. 2017.

[20] Luo W , Li Y , Urtasun R , et al. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[J]. 2017.

[21] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[22] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.

[23] Cherry J M, Adler C, Ball C, et al. SGD: Saccharomyces Genome Database.[J]. Nucleic Acids Research, 1998, 26(1):73-9.

[24] Russakovsky O , Deng J , Su H , et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2014, 115(3):211-252.

[25] Kong T , Yao A , Chen Y , et al. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection[J]. 2016.