一种改进的多门控特征金字塔网络

赵彤, 刘洁瑜*, 沈强
火箭军工程大学导弹工程学院, 陕西 西安 710025

摘要 特征金字塔网络(FPN)在融合不同尺度特征图时采用上采样和相加的方法,然而经过上采样的特征图的空间层级化信息丢失严重,简单地进行相加必然引入一定的误差。同时,FPN 结构的深层特征信息前向传递性较差,其对更浅层的辅助效果基本消失。对此,结合长短时记忆(LSTM)网络在处理上下文信息上的优势对 FPN 结构进行改进,在不同深度的特征层之间建立一条自上而下的记忆链接,建立多门控结构对记忆链上的信息进行过滤和融合以产生表征能力更强的高级语义特征图。最后,将改进的 FPN 结构加入到 SSD(Single Shot MultiBox Detector)算法框架中,提出新的特征融合网络——MSSD(Memory SSD),并在 Pascal VOC 2007 数据集上进行验证。实验表明,该改进取得了较好的测试结果,相比于目前较先进的检测算法也有一定的优势。

关键词 机器视觉; 目标检测; 特征金字塔; 长短时记忆网络; 记忆滤波通道; SSD 网络
中图分类号 TP751.1 **文献标识码** A **doi**: 10.3788/AOS201939.0815005

An Improved Multi-Gate Feature Pyramid Network

Zhao Tong, Liu Jieyu*, Shen Qiang
College of Missile Engineering, Rocket Force University of Engineering, Xi'an, Shaanxi 710025, *China*

Abstract The feature pyramid network (FPN) adopts the method of upsampling and addition when fusing different scale feature maps. However, the spatial stratification information of the upsampled feature map is seriously lost, so that direct addition will inevitably make certain errors. At the same time, the deep feature information of the FPN structure is poorly forward-transferred, and its auxiliary effect to the shallower layer basically disappears. This paper uses the advantages of Long Short-Term Memory (LSTM) network in processing context information to improve the FPN structure. A top-down memory chain is established between feature layers of different depths, and a multi-gate structure is constructed to filter and fuse the information on the memory chain to generate a higher semantic feature map with stronger representation ability. Finally, the improved FPN structure is added to the SSD (Single Shot MultiBox Detector) algorithm framework, and a new feature fusion network, MSSD (Memory SSD), is proposed and verified on the Pascal VOC 2007 data set. Experiments show that the improved algorithm has achieved better test results, and it has certain advantages compared with the current advanced detection algorithms.

Key words machine vision; target detection; feature pyramid; long short-term memory network; memory and filter channel; single shot multibox detector network
OCIS codes 150.0155; 110.3080; 100.4996; 150.1135

1 引 言

目标检测在当今社会上有着广泛应用,如无人驾驶、导弹制导和野外搜救等,其主要任务是对图像中特定目标进行定位和判别。传统的基于手工特征的检测方法虽然取得了较好的效果,但是处理过程繁琐,并且针对不同类型的图像需要选择合适的检测特征,当面对类别较多、内容复杂的图像时,检测效果很差。

2014 年,Girdhivk 等[1]提出的基于区域卷积的神经网络(R-CNN)使用卷积神经网络(CNN)来提取图像特征,相对于传统方法在精度和速度上取得巨大突破,在 PSCAL VOC 2012[2]数据集上平均检测精度达到 53.3%。于是,如何构建 CNN 以产生更有表征能力的特征成为深度学习目标检测算法的重要发展方向。SPPNet[3]、Fast R-CNN[4]、Faster

收稿日期: 2019-03-13; **修回日期**: 2019-04-03; **录用日期**: 2019-04-22
基金项目: 国家自然科学基金青年基金(61503392)
* **E-mail**: 601080018@qq.com

R-CNN[5]等基于 R-CNN 的算法均在单个输入尺度上计算最顶层特征图来预测候选边界，但由于最顶层特征具有固定尺度的接受场，图像中与接受场尺度相差较大的目标的检测误差较大。尤其是对一些小尺度目标来说，最顶层的特征层甚至忽略了其特征信息。2016 年 Liu 等[6]提出了 SSD(Single Shot MultiBox Detector)算法，采用一种多尺度预测的思想，在多个不同深度的特征层上同时来预测候选边界，以适应图像中不同尺度的目标。此方法兼顾了感受野和接受场对目标尺度的适应性，不过其在不同的特征层上独立进行预测，忽视了深层特征对浅层特征的辅助作用。为此，众多学者们都对网络结构进行改进来融合不同深度的特征层。Lin 等[7]提出了特征金字塔网络(FPN)，其利用深度卷积网络固有的多尺度、多层次特征建立自上而下横向连接的结构，从而构建出更具表征能力的高级语义特征图。目前较为先进的检测网络均采用了 FPN 结构，如 YOLOV3[8]，RetinaNet[9]，RefineDet[10]。2017 年 Fu 等[11]提出的 DSSD(Deconvolutional Single Shot Detector)采用了 FPN 的思想，并把 SSD 的基准网络从 VGG(Visual Geometry Group)换成了 Resnet-101，增强了特征提取能力，然后使用反卷积层减少了深层特征图上采样的结构损失，最终提升了目标检测精度(尤其是小物体)，但是其速度下降了很多。Li 等[12]提出的 FSSD(Feature Fusion Single Shot Multibox Detector)先将 VGG16[13]顶层的三个特征提取层进行融合后，再产生 6 个额外特征层，最后进行回归，其在保证 SSD 速度的条件下提升了精确度。Jeong 等[14]提出的 RSSD(Rainbow Single Shot Detector)利用池化和反卷积结构建立了双向融合结构，解决了一个目标匹配到多框的问题同时提升了检测精度。2019 年，Zhao 等[15]提出的 M2Det 在 FPN 的基础上建立更深的 TUM(Thinned U-shape Module)结构，在 VGG16 基础特征图上并联了三路 TUM 结构分别来提取浅层特征、中层特征及深层特征，其精度在 COCO(Common Objects in COntext)数据集上取得了极大提升。Liu 等[16]在 FSSD 的基础上建立了多任务检测算法，结合场景信息特征进行融合，提升了空对地目标的检测。

虽然这些借鉴 FPN 特征融合思想的算法取得了较好的效果，但很少有学者对特征融合时深层信息上采样产生的结构误差进行研究。为此，本文针对 FPN 融合结构进行改进，在不同深度的特征层之间建立一条自上而下的记忆链，将深层的特征信息更有效地保留下来，同时结合长短时记忆(LSTM)网络[17]的门控思想，利用不同深度的特征信息对记忆链上的结构误差进行过滤，对有效特征进行融合以产生表征能力更强的高级语义特征图。在 SSD 算法框架对改进的结构进行验证，提出一种具有选择性和记忆性的网络结构——MSSD(Memory SSD)，并在 VOC2007 数据集上进行验证。实验结果显示，本改进算法在 FPS(Frames Per Second)下降有限的情况下，在精度上相比传统的 FPN 结构有了较大提升，输入 300 pixel×300 pixel 图像时，平均检测精度达到 79.0%。

2　改进的 FPN 算法

传统的 FPN 算法如图 1 所示，左侧自底向上的结构为深度卷积神经网络的卷积过程，BN 为批量标准化。随着卷积层的增加，特征层的尺度逐渐减小，每个像素位置包含的语义信息更强。而后，在右侧建立了横向连接，自上而下地对左侧特征层进行上采样和融合。图 1 中‘⊕’为融合单元，其结构如标注框内所示。先对上层特征层进行双线性插值的上采样，而后与左侧经过 1×1 的卷积操作的特征层直接相加。为消除相加时造成的混叠效应，在之后增加一个 3×3 的卷积操作，最后输出稳健性更强、精准度更高的特征图[7]。因为浅层特征可以提供更加准确的位置信息，而多次的降采样和上采样操作使得深层网络的定位信息存在误差。FPN 巧妙地将处理过的低层特征和处理过的高层特征进行累加，这样就构建了一个更深的、融合多层信息的特征金字塔，并在不同的特征进行输出，最终有效地提高了检测精度。FPN 的输出特征图表示为

$$\varphi_n = g(\phi_n, \phi_{n+1}) = f^{3\times3}\{\phi_n + f^{1\times1}[v(\phi_{n+1})]\}, \tag{1}$$

$$\phi_n = f_n(\phi_{n-1}) = f_n\{f_{n-1}[\cdots f_1(I)]\}, \tag{2}$$

式中：φ_n 为第 n 个经 FPN 结构融合后的特征图；g 为 FPN 结构函数；v 是双线性插值函数；$f^{1\times1}$ 为卷积核为 1×1 大小的卷积层；$f^{3\times3}$ 为卷积核为 3×3 大小的卷积层；ϕ_n 为特征提取网络的第 n 层特征图；I 为原始图像；f_n 为特征提取网络的第 n 个卷积函数。

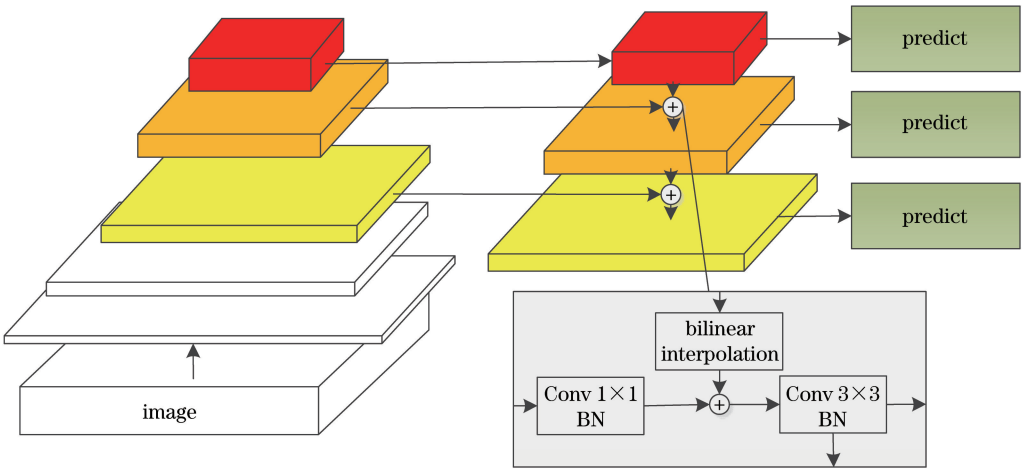

图 1　FPN 结构

Fig. 1　FPN structure

FPN 在融合不同深度的特征层时采用直接相加的做法，然而经过上采样后的深层特征图空间结构信息受损，必然会产生噪声。一方面由于深层特征层拥有更大的感受野，其特征包含大量的高级语义信息，因此当深层特征被噪声污染时，其对浅层特征中的位置信息必然造成负面干扰，影响目标检测。同时，深层特征包含的部分全局信息对浅层的小区域信息来说并不总有正面影响，当全局信息中包含一类以上目标时，必然仅对其中的某一类目标的定位有辅助作用。随着其自上而下融合，噪声又被进一步累加，即使 FPN 在不同层进行有监督的训练，仍然不可避免地对浅层特征图的细粒度产生影响。另一方面，FPN 在自上而下的融合过程中不相邻的特征层之间信息传递能力较低。在经过多层融合后，与当前层相邻较远的特征层的影响基本消失，这些特征层的辅助作用便被忽略。因此，本研究结合 LSTM 的记忆和筛选能力，对 FPN 进行改进。

改进的网络结构如图 2 所示。图 2 中‘⊕’为反卷积融合单元，其结构如标注框内所示。为减小噪声引入，增强高层特征表征能力，采用反卷积对上层特征层进行上采样，而后与当前层的特征层相加，通过一个 3×3 的卷积单元减弱混叠效应。延续 FPN 自上而下的横向连接结构，并加入记忆和滤波通道，此结构对融合后特征图的有效信息进行保留，对无效信息进行滤除。同时，在所有的特征层之间建立记忆链，增强上下文信息的传递效率。最后输出对上下文信息提取能力更强的高级特征图，在其上进行下一步的边框回归。

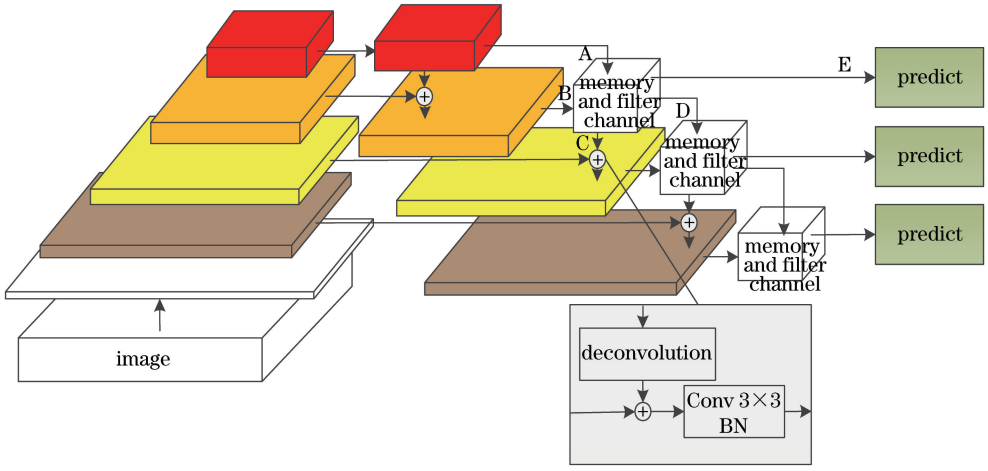

图 2　改进的 FPN 结构

Fig. 2　Improved FPN structure

反卷积融合单元如

$$\varphi_n = g(\phi_n, \phi_{n+1}) = f^{3\times3}[\phi_n + w(\phi_{n+1})] \quad (3)$$

所示，式中，w 为反卷积函数，其余参数同(1)式。

记忆和滤波通道的结构如图 3 所示。该结构有

两个输入通道和三个输出通道，A 输入为当前层之前的所有特征层的有效记忆信息，B 输入为当前层的融合特征图，C 输出为通过记忆和滤波通道筛选后的特征信息，D 输出为融合当前层的记忆信息，E 输出为最终输出的增强特征图。本结构采用 LSTM 网络的门控思想，从 A 端开始的箭头实线为记忆链，其贯穿网络中每个参与到边框回归的特征层，并与这些特征层进行线性的信息交互。在每次进入记忆和滤波通道时，都对保留的特征信息进行反卷积操作。I 区域为遗忘门，当新的特征信息加入进来后，利用这些信息对记忆链中前层的特征进行筛选，去除无用信息和噪声。其中 3×3的卷积操作跨通道整合了融合后的特征信息，增加了模型的深度，在一定程度上提升了模型的非线性。Sigmoid 函数的输出为 0 到 1，表示记忆链中信息的传递程度，0 表示信息截断，1 表示完全传递。II 区域为输入门，其对新特征进行筛选，将有效特征提取出来对记忆链中的信息进行更新。其中 Tanh 函数用来生成候选的特征信息，Sigmoid 函数决定候选信息的传递量，两个函数共同作用完成对记忆链的更新。III 区域为输出门，其利用新加入的特征信息对更新后的记忆链进行选择。更新后的记忆链中包含当前层和当前层之前的所有特征层的融合信息，相比原始 B 的输入包含更多上下文信息。采用 Tanh 函数对记忆链的信息进行压缩处理，提升参数的稳健性，采用 Sigmoid 函数对其进行过滤。C 输出链接下一个新加入的特征层，D 输出链接下一个记忆和滤波通道。E 输出则参考 MS-CNN(Multi-Scale Deep Convolutional Neural Network)[18]改善任务分支网络的技巧，在 C 输出后增加 1×1 的卷积层，再进行边框的回归。由于各个门结构和卷积层承担不同的选择功能，因此本结构中的所有单元不共享参数。输出的特征图可表示为

$$\varphi_C = F_{III}(\varphi_D), \tag{4}$$

$$\varphi_D = w(\phi_A) \times F_I(\phi_B) + F_{II}(\phi_B), \tag{5}$$

$$\varphi_E = \mathrm{relu}[f_0^{1\times1}(\varphi_D)], \tag{6}$$

式中：ϕ_A、ϕ_B 为记忆和滤波通道的输入特征图；φ_C、φ_D、φ_E 为输出特征图；F_I、F_{II} 和 F_{III} 分别为遗忘门、输入门和输出门；relu[]为激活函数；$f_n^{1\times1}$ 为第 n 个卷积核大小为 1×1 的卷积层。三个门结构的计算方法可表示为

$$F_I(\lambda) = \mathrm{Sigmoid}[f_1^{3\times3}(\lambda)], \tag{7}$$

$$F_{II}(\lambda) = \mathrm{Sigmoid}[f_2^{3\times3}(\lambda)] \times \mathrm{Tanh}[f_3^{3\times3}(\lambda)], \tag{8}$$

$$F_{III}(\lambda) = \mathrm{Sigmoid}[f_4^{3\times3}(\phi_B)] \times \mathrm{Tanh}[f_3^{3\times3}(\lambda)], \tag{9}$$

式中，Sigmoid 和 Tanh 为激活函数。

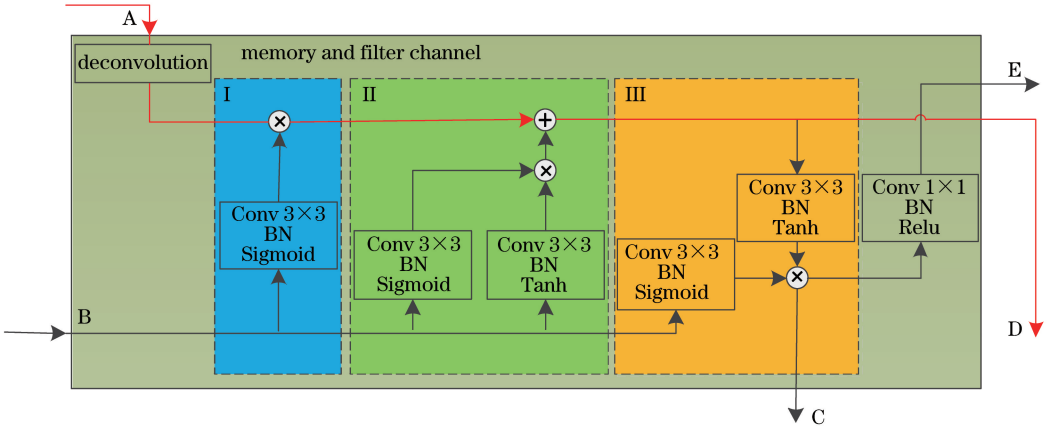

图 3　记忆和滤波通道的结构

Fig. 3　Memory and filter channel structure

3　MSSD 网络

为验证改进的 FPN 网络的有效性，以 SSD 网络为基础，提出 MSSD 网络。MSSD 算法是在 SSD 算法的基础上加入改进的 FPN 结构，基础网络依旧沿用 VGG16。而后，在其中 4 个附加层上建立记忆和滤波通道。MSSD 的网络结构如图 4 所示，下面，对 MSSD 网络的具体模块进行介绍。

3.1　基础特征提取网络

MSSD 的基础特征提取网络沿用 VGG16，并与传统 SSD 网络一样在 Conv4_3 层之后增加了 6 个额外的特征层，在不同感受野和接受场对目标进行特征的提取。也在 ResNet101[19]上的同尺度的特征层上进行测试，对比结果显示其在精度上有少量

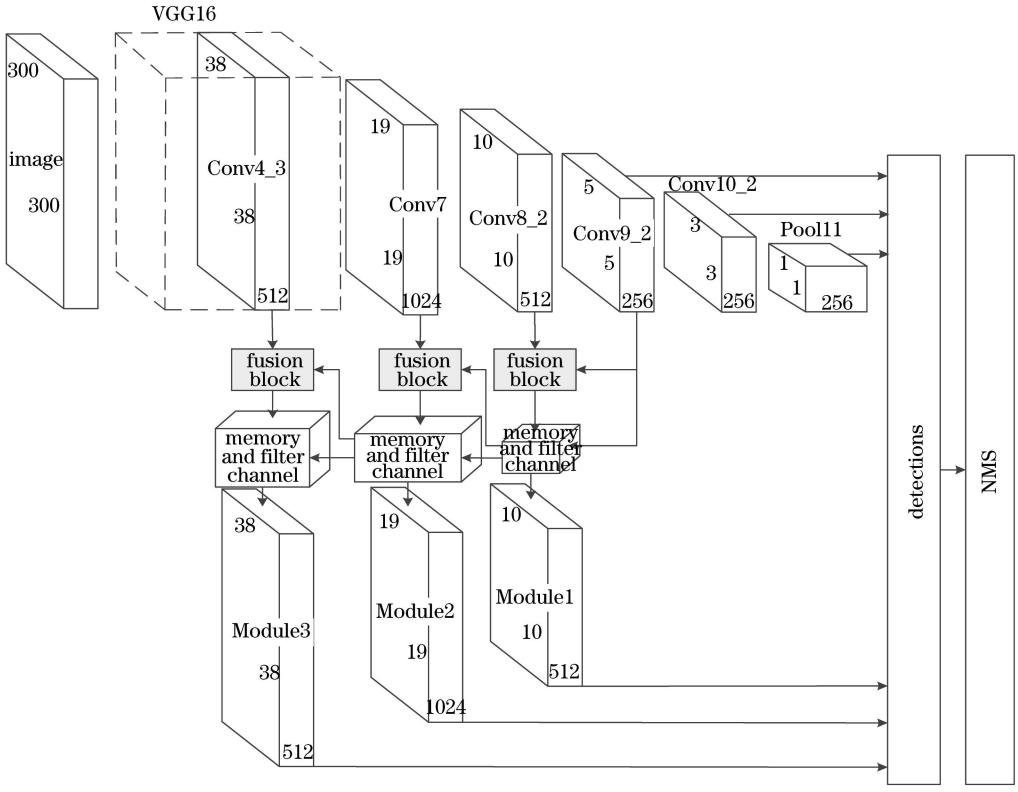

图 4　MSSD 网络结构

Fig. 4　MSSD network structure

提升，但是 FPS 大幅下降。虽然 ResNet 对图像的提取能力更强，但其网络结构更深，复杂度更高，综合考虑后，依旧采用 VGG16 进行特征提取。

3.2　Fusion block 及记忆和滤波通道

该结构即改进的 FPN，Fusion block 如(3)式所示。传统的 FPN 在融合当前层特征时采用 1×1 的卷积层进行降维处理，而后与上层网络经过双线性插值后相加。而 Fusion block 为提升信息的利用率，去除了 1×1 的卷积层，相应地为保持维度一致，利用反卷积对上层网络进行上采样。相比于双线性插值来说，反卷积是网络通过训练习得的，文献[20]也指出通过反卷积后进行融合可以得到较好的细节，获得尽可能强的语义信息。本网络 3 组反卷积的参数输入通道、输出通道、卷积核尺度和步长分别为[256，512，2，2]、[512，1024，1，2]和[1024，512，2，3]。记忆和滤波通道结构是结合 LSTM 网络提出的。LSTM 网络在长序列依赖问题上提供了一个有效的解决方案，而在 CNN 中，随着网络的加深，特征图的感受野增加，特征也从细节敏感变到语义敏感。整个过程可以看作为视角上升的过程，所有特征层可以看作为一组更加复杂的序列信息。本网络中 Conv10_2 和 Pool11 两层由于尺寸较小，上采样重建难度较大，夹杂噪声较多，仅在 Conv4_3、Conv7、Conv8_2 和 Conv9_2 这 4 个特征层上建立横向链接。为匹配 SSD 网络的接受域，最终得到的 Module1、Module2 和 Module3 特征层的尺度和维度与 Conv4_3、Conv7 和 Conv8_2 保持一致。

3.3　锚点(Anchors)

传统的 SSD 网络在每个特征层的锚点上生成 4～6个默认框(Default Boxes)。在本网络中，增加了锚点的覆盖度同时兼顾网络的复杂度，在每个特征层的锚点上生成同样数目的默认框，其比例参照文献[9]，在每个尺度的特征层上设置 3 个横纵比[1/2，1，2]，每个横纵比分为[2^0，$2^{1/3}$]两个级别。经计算，对于 300 pixel×300 pixel 的输入图像，产生的默认框覆盖的区域范围为 30 pixel×30 pixel 到 330 pixel×330 pixel，满足覆盖所有图像的目标。采用文献[6]的匹配方案将真实框(Ground Truth Boxes)匹配到默认框，交并比(IOU)设置为 0.5。边框回归的计算方式为计算边界框(Bounding Boxes)与匹配到目标的默认框的长，宽和中心位置 4 个参数的偏移度。

3.4　模型的训练

在模型训练过程中，首先对数据进行增广处理。

对训练集随机进行水平翻转，尺度变化，亮度变化和旋转变化，增加网络的稳健性。训练的损失函数同文献[6]，位置损失采用 smooth L1，分类损失采用 Log loss，并采用难样本挖掘（Hard Negative Mining）对正负样本进行平衡。优化器采用传统的随机梯度下降法（SGDR），学习率设置为 0.001，动量值为 0.9。主干网络 VGG16 采用在 ImageNet 上的预训练模型，其余卷积层采用均值为 0、方差为0.01的高斯权重填充，偏执值初始化为 0。训练在 1 个 GPU 上进行，批次设为 32，共进行 140000 批次的训练。

4　实　　验

实验在 Ubuntu16.04 系统的 Pytorch 框架下运行，并使用 CUDA8.0 和 cuDNN5.0 来加速训练。计算机搭载的 CPU 为 Corei7-8700k，显卡为 NVIDIA GTX1080Ti，内存为 32 G。网络的性能通过平均精度和检测速度（即 FPS）来测试。在 PASCAL VOC2007 和 PASCAL VOC2012 的训练集上对 MSSD 网络进行训练，训练集包含 16551 幅图像，20 类目标。在 PASCAL VOC2007 的测试集上对网络进行测试，其中测试集包含 4952 幅图像。

4.1　算法有效性研究

为验证改进结构的有效性，设置了若干组对比实验。其中实验的图像输入尺寸统一设置为 300 pixel×300 pixel，输出的检测框与真实框的 IOU 阈值统一为 0.5，FPS 为计算机在空载状态时在 NVIDIA GTX1080Ti 上测得。

4.1.1　与 FPN 结构对比

在 MSSD 网络的基础上将 Conv4_3、Conv7、Conv8_2 和 Conv9_2 这 4 个特征层的增强方式改为 FPN，其中横向链接的 1×1 卷积层的通道数同文献[8]的 FPN 结构，设置为 256。将此 FPN-SSD 结构与 MSSD 网络进行对比。MSSD 选择在 Conv4_3、Conv7、Conv8_2 和 Conv9_2 这 4 个特征层上进行特征增强，这里设置同时在 6 个特征层上进行增强对比实验，如表 1 所示。从表中可以看出，MSSD 网络取得了更好的检测结果，在牺牲一定速度的情况下相比原始 SSD 网络，精度提升了 1.5%。当 SSD 网络加入 FPN 时，检测精度仅仅提升了 0.2%，相对于改进的 FPN 在 mAP（Mean Average Precision）上相差 1.3%。但是在速度上，FPN 结构有明显的优势，这是由于 FPN 结构首先采用了 1×1 的卷积对 SSD 的特征层进行降维处理。MSSD 在 6 个特征层进行预测时，精度有所下降，这是由于 Conv10_2 和 Pool11 这两个特征层的尺度较小（分别为 3×3 和 1×1），在进行上采样时加入的噪声信息的影响更大。

表 1　加入 FPN 结构的 SSD 网络与 MSSD 网络的性能对比

Table 1　Comparison between SSD network with FPN structure and MSSD network

Method	Feature layers	mAP /%	FPS
SSD[6]	Conv4-Pool11	77.5	46(Titan X)
SSD+FPN	Conv4-Conv9	77.7	53.9
MSSD	Conv4-Pool11	78.8	27.1
MSSD	Conv4-Conv9	79.0	31.7

4.1.2　特征融合结构

MSSD 的特征融合结构相比 FPN 去除了 1×1 的卷积层，并将双线性插值的上采样改为反卷积，特征融合采用相加的方式（Sum）。为验证该改进的有效性，保留其他网络结构，分别测试是否去除 1×1 的卷积层和是否采用反卷积的不同情况，并与原始网络进行对比，如表 2 所示，其中√表示使用。最后，采用参考文献[21]在经过门控后进行特征融合时采用的Max融合方式对本特征融合结构进行对

表 2　融合结构的改进效果对比

Table 2　Comparison of improved effects of fusion structure

Conv 1×1	Feature fusion mode	Deconv or bilinear interpolation	mAP /%	FPS
√	Sum	Deconv	78.9	31.2
	Sum	Deconv	79.0	31.7
√	Sum	Bilinear interpolation	78.5	40.1
	Sum	Bilinear interpolation	78.5	40.2
	Max	Deconv	78.2	36.4
	Max	Bilinear interpolation	76.6	42.2

比实验。从表中可以看出采用反卷积的算法相比于双线性插值方法在精度上有较大提升，但速度有所下降。这是由于反卷积需要在线对参数进行训练，复杂度相对较高。是否采用 1×1 的卷积层，对网络的精度和速度影响不大。当采用 Max 的融合方法时，算法检测速度基本不变，但是精度下降较多。

4.1.3　记忆和滤波通道

本结构是 MSSD 网络的核心，该结构主要包含三个门结构。本部分分别对三个门结构的有效性作对比实验，如表 3 所示。当不包含输入门时，网络训练时发散。当仅有输入门时，精度最低，这是由于其对记忆链上的信息整合能力最差，基本没有利用新加入的特征信息对记忆链进行滤波。遗忘门和输入门相结合的精度较高，这也与文献[17]的结论一致——遗忘门和输出门是 LSTM 结构最重要的两个部分。当三个门结构同时作用时，网络的精度最高。

表 3　记忆和滤波通道的有效性分析

Table 3　Analysis of effectiveness of memory and filter channels

Number	Forget gate	Input gate	Output gate	mAP /%
1	√			/
2		√		75.2
3			√	/
4	√	√		78.6
5	√		√	/
6		√	√	77.9
7	√	√	√	79

4.1.4　基础特征提取网络

MSSD 提出的特征提取网络采用了 VGG16。而相比于 ResNet 来说，VGG16 的特征提取能力明显不足，因此将特征提取网络改为 ResNet101 进行对比。采取文献[8]的方法，在 ResNet101 的第 3～5 个 block 上建立增强结构，三个特征层分别为 Conv3_12、Conv4_69 和 Conv5_9。对于增强结构，对比 FPN 和本研究的改进 FPN 结构，结果如表 4 所示。相比之下，以 ResNet101 为基础网络的 MSSD 取得了最高的检测精度(79.3%)，相比加入 FPN 结构的 SSD 提升了 0.6%，相比 VGG16 下的 MSSD 提升了 0.3%。但是 ResNet101 的网络层数更多(101 层)复杂度更高，因此 FPS 下降较多。其中 SSD＋FPN 结构在 ResNet101 的基础网络下 mAP 有较大提升，FPS 也大于以 VGG16 为基础网络的 MSSD。这是由于 ResNet101 的特征提取能力更强，网络深层的特征信息稳健性好，噪声更少，对 FPN 的干扰更少。FPN 结构大幅度减小了预测层的通道数(总共仅有 768 个通道)。

表 4　改变基础网络性能对比

Table 4　Performance comparison of different basic networks

Method	Network	mAP /%	FPS
SSD[11]	ResNet101	77.1	18.9(Titan X)
SSD＋FPN	ResNet101	78.7	39.1
MSSD	ResNet101	79.3	23.7
MSSD	VGG16	79.0	31.7

4.2　算法性能研究

将所提算法与其他先进的深度学习目标检测算法进行比较，如表 5 所示。Faster R-CNN 和 R-FCN[22] 的输入尺度为 600 pixel×600 pixel～1000 pixel×1000 pixel。在 300 pixel×300 pixel 的输入图像下，MSSD 算法的平均精度可以达到 79.0%，相比 SSD300 的 77.5%有了 1.5%的提升，可以与 YOLOV3 在 416 pixel×416 pixel 尺度的输入图像的精度相媲美。相比于在 FPN 下改进的、采用 ResNet101 基础网络的 DSSD300 的 78.6%也有 0.4%的提升。当输入图像尺寸为 512 pixel×512 pixel时，MSSD 网络在 VGG16 上精度可以达到 81.0%。

表 6 为其他在 SSD 网络上进行改进的算法对比。表中训练集均为 VOC2007 和 VOC2012，测试集为 VOC2007，所有算法均经过预训练，IOU 阈值均取 0.5。相比来说，MSSD300 在以 VGG16 为特征提取网络时兼顾了速度和精度，在 SSD 的众多改进网络中取得了更好的性能。在输入为 300 pixel×300 pixel 图像时，MSSD 的精度在各类算法中最高。

表 7 为在 300 pixel×300 pixel 尺度输入时的 SSD 算法和 MSSD 算法对小目标的检测精度对比。按照文献[26]中将 32 pixel×32 pixel 以下的目标定义为小目标进行测试(验证集的真实目标仅保留 32 pixel×32 pixel 以下的)。从表中可以看出改进的网络对小目标的检测也有部分提升(0.9%)。

表 5　各类先进深度学习算法在 VOC2007 数据集上精度对比

Table 5　Accuracy comparison of various advanced deep learning algorithms on VOC2007 dataset

Method	Network	mAP /%																			
		Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	Tv	Average
Faster[5]	VGG16	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6	73.2
ION[23]	VGG16	79.2	83.1	77.6	65.6	54.9	85.4	85.1	87.0	54.4	80.6	73.8	85.3	82.2	82.2	74.4	47.1	75.8	72.7	84.2	80.4	75.6
Faster[18]	ResNet101	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	76.4
MR-CNN[24]	VGG16	80.3	84.1	78.5	70.8	68.5	88.0	85.9	87.8	60.3	85.2	73.7	87.2	86.5	85.0	76.4	48.5	76.3	75.5	85.0	81.0	78.2
R-FCN[22]	ResNet101	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	80.5
SSD300[6]	VGG16	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	83.97	79.4	52.3	77.9	79.5	87.6	76.8	77.5
SSD512[6]	VGG16	84.8	85.1	81.5	73.0	57.8	87.8	88.3	87.4	63.5	85.4	73.2	86.2	86.7	83.9	82.5	55.6	81.7	79.0	86.6	80.0	79.5
DSSD321[11]	ResNet101	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	78.6
DSSD513[11]	ResNet101	86.6	86.2	82.6	74.9	62.5	89.0	88.7	88.8	65.2	87.0	78.7	88.2	89.0	87.5	83.7	51.1	86.3	81.6	85.7	83.7	81.5
MDSSD300[25]	VGG16	86.5	87.6	78.9	70.6	55.0	86.9	87.0	88.1	58.5	84.8	73.4	84.8	89.2	88.1	78.0	52.3	78.6	74.5	86.8	80.7	78.6
MDSSD512[25]	VGG16	88.8	88.7	83.2	73.7	58.3	88.2	89.3	87.4	62.4	85.1	75.1	84.7	89.7	88.3	83.2	56.7	84.0	77.4	83.9	77.6	80.3
YOLOV3[8]	Darknet53	85.5	85.5	75.6	70.0	66.5	87.6	87.7	89.4	64.3	83.5	73.6	85.9	86.9	86.2	83.3	56.2	75.3	78.0	86.4	77.8	79.2
MSSD300	ResNet101	81.2	87.2	78.7	72.7	53.4	86.4	85.6	89.1	63.1	84.5	80.0	87.5	88.9	84.8	78.8	54.5	80.9	83.2	87.1	77.4	79.3
MSSD300	VGG16	81.6	85.8	78.0	74.0	55.3	86.2	86.5	88.2	64.6	85.9	76.9	85.4	87.7	85.2	79.5	51.1	78.8	80.8	87.9	78.6	79.0
MSSD500	VGG16	87.6	87.2	83.7	75.5	57.8	86.7	88.4	89.5	66.3	84.6	78.9	86.8	88.1	86.4	83.3	58.0	81.4	81.0	88.4	78.1	81.0

表 6　各类基于 SSD 的改进算法的测试结果对比

Table 6　Comparison of test results of various improved algorithms based on SSD

Method	Network	FPS	GPU	# proposals	Input size /(pixel×pixel)	mAP /%
SSD300[6]	VGG16	46	Titan X	8732	300×300	77.2
SSD512[6]	VGG16	19	Titan X	24564	512×512	78.5
MDSSD300[25]	VGG16	38.5	1080Ti	44530	300×300	78.6
MDSSD512[25]	VGG16	17.3	1080Ti	-	512×512	80.3
DSSD321[11]	ResNet101	9.5	Titan X	17088	321×321	78.6
DSSD513[11]	ResNet101	5.5	Titan X	43688	513×513	81.5
RSSD300[14]	VGG16	35	Titan X	8732	300×300	78.5
RSSD512[14]	VGG16	16.6	Titan X	24564	512×512	80.8
MSSD300	ResNet101	23.7	1080Ti	8728	300×300	79.3
MSSD300	VGG16	31.7	1080Ti	8732	300×300	79.0
MSSD512	VGG16	17.3	1080Ti	24564	512×512	81.0

表 7　SSD 算法和 MSSD 算法对小目标的检测精度

Table 7　Detection accuracy of small targets by SSD algorithm and MSSD algorithm

Method	mAP /%
SSD[6]	55.4
MSSD	56.3

MSSD 算法的部分可视化结果如图 5 所示，其中左侧为 SSD 算法，右侧为 MSSD 算法，坐标轴代表图像像素值。图 5(a)中 SSD 算法漏检了人，而图 5(b)中的 MSSD 算法在目标有大量重合的情况下依然可以很好地检测到整个目标。图 5(c)与图 5(d)中都存在漏检的现象，SSD 算法漏检了人和一辆摩托，MSSD 算法只漏检了一辆摩托但是选框的准确度不够(将人也包含进了选框)。图 5(e)和图 5(f)中，存在小尺度目标，SSD 算法漏检而 MSSD 算法更好地检测到了汽车。图 5(g)和图 5(h)中 SSD 算法漏检了与沙发位置完全重合的人，而 MSSD 算法检测准确。所有的选框得分上，SSD 算法总体较高。总的来说，两种算法各有优点与不足，但是 MSSD 算法相对来说更具有优越性。

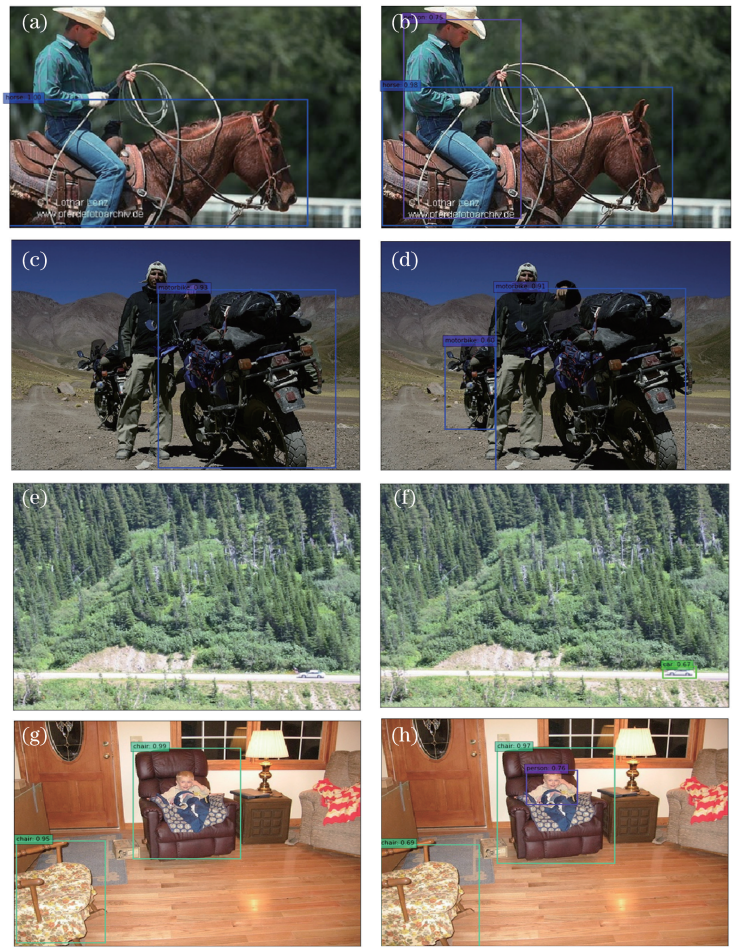

图 5　SSD 算法和 MSSD 算法的可视化对比。(a)(c)(e)(g) SSD 算法;(b)(d)(f)(h) MSSD 算法

Fig. 5　Visual comparison of SSD algorithm and MSSD algorithm. (a)(c)(e)(g) SSD algorithm; (b)(d)(f)(h) MSSD algorithm

5　结　　论

分析了传统的 FPN 在特征融合过程中存在的不足,为增强不同深度特征的融合效果,结合 LSTM 提出了一种多门控的记忆滤波结构。为验证改进的有效性,在 SSD 网络的框架上结合改进 FPN 提出了 MSSD 网络。在 Pascal VOC 2007 数据集上进行测试,证明了本改进的有效性。并与其他先进的深度学习目标检测算法对比,结果显示,所提算法具有更好的效果,为深度学习目标检测方向的研究者们提供了一种更有效的特征融合算法。

参　考　文　献

[1] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[2] Everingham M, Eslami S M A, van Gool L, *et al*. The PASCAL visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[3] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[4] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[5] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[6] Liu W, Anguelov D, Erhan D, *et al*. SSD: single

shot MultiBox detector[M]// Leibe B, Matas J, Sebe N, *et al*. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[7] Lin T Y, Dollar P, Girshick R, *et al*. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 936-944.

[8] Redmon J, Farhadi A. YOLOv3: an incremental improvement[J/OL]. (2018-04-08)[2019-01-30]. https://arxiv.org/abs/1804.02767.

[9] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2999-3007.

[10] Zhang S F, Wen L Y, Bian X, *et al*. Single-shot refinement neural network for object detection[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-23, 2018, Salt Lake City. New York: IEEE, 2018: 4203-4212.

[11] Fu C Y, Liu W, Ranga A, *et al*. DSSD: deconvolutional single shot detector[J/OL]. (2017-01-23)[2019-01-30]. https://arxiv.org/abs/1701.06659.

[12] Li Z X, Zhou F Q. FSSD: feature fusion single shot multibox detector[J/OL]. (2018-05-17)[2019-01-30]. https://arxiv.org/abs/1712.00960.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. (2015-04-10)[2019-02-02]. https://arxiv.org/abs/1409.1556.

[14] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection[J/OL]. (2017-05-26)[2019-02-01]. https://arxiv.org/abs/1705.09587.

[15] Zhao Q J, Sheng T, Wang Y T, *et al*. M2Det: a single-shot object detector based on multi-level feature pyramid network[J/OL]. (2019-01-06)[2019-02-01]. https://arxiv.org/abs/1705.09587.

[16] Liu X, Chen J, Yang D F, *et al*. Scene-coupled intelligent multi-task detection algorithm for air-to-ground remote sensing image[J]. Acta Optica Sinica, 2018, 38(12): 1215008.
刘星, 陈坚, 杨东方, 等. 场景耦合的空对地多任务遥感影像智能检测算法[J]. 光学学报, 2018, 38(12): 1215008.

[17] Graves A. Supervised sequence labelling with recurrent neural networks: long short-term memory[M]. Berlin, Heidelberg: Springer, 2012: 37-45.

[18] Cai Z W, Fan Q F, Feris R S, *et al*. A unified multi-scale deep convolutional neural network for fast object detection[M]// Leibe B, Matas J, Sebe N, *et al*. Lecture notes in computer science. Cham: Springer, 2016, 9908: 354-370.

[19] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[20] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[21] Zeng X Y, Ouyang W L, Yan J J, *et al*. Crafting GBD-net for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(9): 2109-2123.

[22] Dai J, Li Y, He K, *et al*. R-FCN: object detection via region-based fully convolutional networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. USA: Curran Associates Inc., 2016: 379-387.

[23] Bell S, Zitnick C L, Bala K, *et al*. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2874-2883.

[24] Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware CNN model[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1134-1142.

[25] Xu M L, Cui L S, Lü P, *et al*. MDSSD: multi-scale deconvolutional single shot detector for small objects[J/OL]. (2018-08-19)[2019-02-02]. https://arxiv.org/abs/1805.07009.

[26] Lin T Y, Maire M, Belongie S, *et al*. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, *et al*. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.