文章编号：2095-2163(2022)07-0090-06 中图分类号：TP391 文献标志码：A

一种基于共享多头模块的轻量型一阶段网络

肖贵明，丁德锐，梁 伟，魏国亮
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：一阶段目标检测网络 SSD 备受青睐，本文基于标准的 SSD 网络，提出了一种新颖的轻量型 SSD（Lightweight SSD，LSSD）网络构架；此外，还提出了 SMHM（Shared Multi-Head Module）模块，该模块使得所有输出层级共享网络头部。相对于标准的 SSD，本文提出的这种改进型一阶段网络构架（记为 SMHM-LSSD）具有更少的网络参数量、更快的速度、更高的检测精度。本文在香港中文大学和商汤科技推出的平台 mmdetection 上对 VOC0712 数据集进行实验，其中 VOC0712 训练集进行训练，VOC07 测试集进行测试。实验结果显示，本文提出的 LSSD 相比于 SSD 提高了 0.2%的检测精度，减少了 23.1 M 的参数量，提升了 5 fps/s 的速度；加入 SMHM 模块后，最高提升了 0.6%的检测性能，减少 28.7 M 的参数量，提升 8 fps/s 的速度，SMHM-LSSD 达到了 78.9%的均值平均精度。

关键词：目标检测；轻量型 SSD；SMHM；均值平均精度

A lightweight one-stage network based on shared multi-head modules

XIAO Guiming, DING Derui, LIANG Wei, WEI Guoliang
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 The one-stage object detection network SSD is very popular. Based on the standard SSD network, this paper proposes a novel LSSD (Lightweight SSD) network architecture. In addition, the SMHM (Shared Multi-Head Module) is proposed, which enables all output levels to share the network head. Compared with the standard SSD, the improved one-stage network architecture (denoted as SMHM-LSSD) proposed in this paper has fewer network parameters, faster speed, and higher detection accuracy. Experiments are conducted on the VOC0712 dataset on the platform mmdetection launched by the Chinese University of Hong Kong and SenseTime. The VOC0712 training set is used for training and the VOC07 test set is used for testing. The experimental results show that the LSSD proposed in this paper improves the detection accuracy by 0.2% compared with the SSD, reduces the parameter amount of 23.1 M, and improves the speed of 5 fps/s. After adding the SMHM module, the detection performance is improved by up to 0.6%. By reducing the parameter amount of 28.7 M and increasing the speed of 8 fps/s, SMHM-LSSD reached 78.9% of mAP.

【Key words】 object detection; Lightweight SSD; shared multi-head module; mean average precision

0 引 言

在计算机视觉中，目标检测一直是最基本也是最重要的研究方向之一。2012 年，深度学习开始快速发展，目标检测也随之迎来了快速发展期。现今，对目标检测可分为一阶段网络构架和二阶段网络构架。一阶段网络具有参数小、速度快的特性；二阶段网络相比一阶段网络具有高精度，参数大、速度慢的特性。YOLO V1 作为早期目标检测一阶段网络框架，具有良好的速度和不错的检测性能，但随着时间发展其检测精度稍有不足[1]。2016 年，SSD 横空出世，其同时拥有高速度和高检测精度[2]。但是 SSD 也存在问题：

（1）特征提取不充分，低层级特征冗余、语义信息不高。

（2）多个输出层的定位分支和分类分支特征共享，而定位偏向于边缘特征，分类偏向区域局部特征[3]。

（3）网络维持高精度的同时无法很好的做到实时性。

针对以上几个问题，多种对 SSD 的改进及其变种出现，主要集中在以下两方面：

（1）增加检测速度。文献[4]通过更改主干网络，提出 MobileNet v1，使用轻量型网络提升速度；文

基金项目：国家自然科学基金（61973219）；上海市“科技创新行动计划”国内科技合作项目（20015801100）资助。

作者简介：肖贵明（1995-），男，硕士研究生，主要研究方向：视觉目标检测；丁德锐（1981-），男，博士，教授，博士生导师，主要研究方向：随机非线性控制与滤波、智能优化算法、图像处理；梁 伟（1996-），男，博士研究生，主要研究方向：生成对抗网络、视觉跟踪；魏国亮（1973-），男，博士，教授，主要研究方向：随机控制、电机控制。

通讯作者：丁德锐 Email：deruiding2010@usst.ed.cn

收稿日期：2022-01-05

献[5]在其基础上进一步对网络构架进行修改，提出 MobileNet v2，网络构架更小，参数量更少，速度更快；文献[6]提出了新的轻量型网络 PeleeNet，相比 MobileNet 具有更少的参数量。但是提出的这些轻量型网络通常会使得检测精度下降。

(2)增加检测精度：文献[7]在 SSD 的网络基础上增加 FPN 块，用于增强特征融合，提升网络性能；文献[8]通过将浅层特征应用到深层特征中，进一步增强网络性能；文献[9]通过引入注意力机制，提升网络性能。值得指出的是，这些网络改进在增加 SSD 网络性能的同时，通常也增加了大量参数，使得网络速度变慢。

综上所述，本文致力于建立一种更轻、更快、更准的一阶段网络构架，同时兼顾速度、精度和参数量 3 方面需求，获得的创新点主要包括如下两个方面：

(1)针对速度和精度不能同时顾及的问题，本文在原有 SSD 的网络上进行了改进，提出了 LSSD 网络构架，在降低参数量，提升速度的同时，增加了检测精度。

(2)在提出 LSSD 网络构架的基础上进一步对网络头部进行了改进，提出了 SMHM 模块。在 SMHM 模块中将分类头部和定位头部分开训练，同时在分类头部中使用了注意力机制、双头机制、权重自适应、进行了参数共享，进一步降低参数量，提高检测精度。

1　SSD 网络构架的改进

1.1　LSSD 网络构架

相比于 SSD 网络构架，LSSD 主要在 VGG 的拓展层 Conv6 后面的网络层上进行了改进，如图 1 所示。

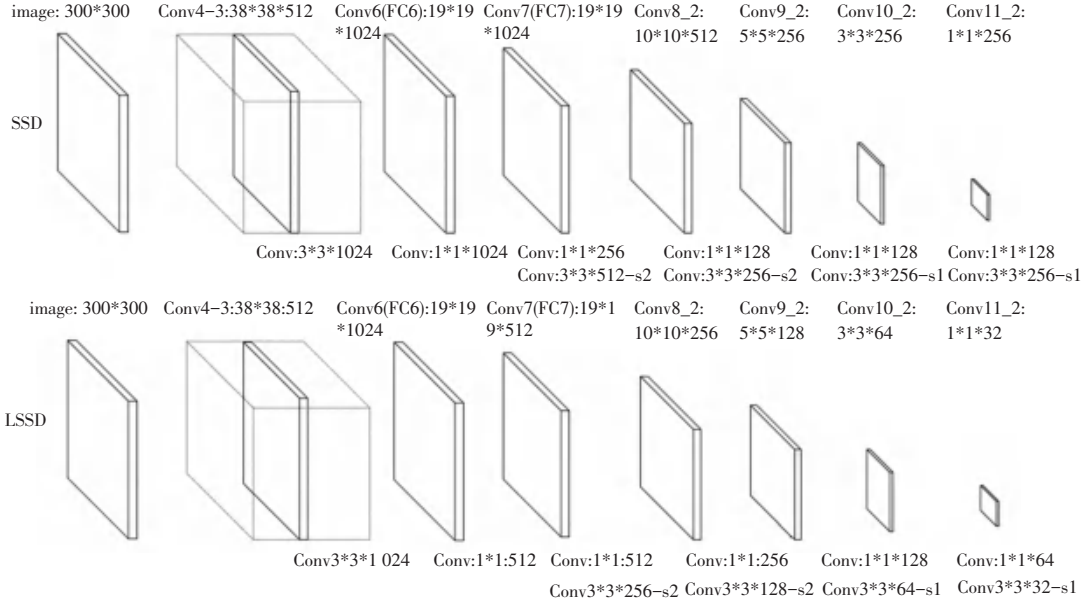

图 1　从 SSD 到 LSSD 的网络结构变化图

Fig. 1　Diagram of network structure change from SSD to LSSD

SSD 网络的通道数是先增加后减少，其中在 Conv6 后面的网络层次为拓展层。SSD 共有 6 层输出，分别为：Conv4-3：38 * 38 * 512、Conv7：19 * 19 * 1024、Conv8-2：10 * 10 * 512、Conv9-2：5 * 5 * 256、Conv10-2：3 * 3 * 256、Conv11-2：1 * 1 * 256；在拓展层的每层输出前，都会使用 Conv：1 * 1 * C 的卷积层，减少网络参数量；在 Conv7 ~ Conv8-2 之间使用 Conv：1 * 1 * 256；Conv8-2 ~ Conv9-2 之间使用 Conv：1 * 1 * 128；Conv9-2 ~ Conv10-2 之间使用 Conv：1 * 1 * 128；Conv10-2 ~ Conv11-2 之间使用 Conv：1 * 1 * 128。

SSD 输出层以及输出层间的卷积层的网络设计存在两个问题：

(1)过大的输出层，将会带来大量的参数，造成特征冗余，同时网络进行分类和定位时，提取过多的特征不具有表征性。

(2)输出层间的卷积层使用 Conv：1 * 1 * C，使通道数下降，在 Conv7 到 Conv9-2 中下降 4 倍，在 Conv9-2 ~ Conv11-2 中下降 2 倍；再通过 Conv：3 * 3 * C- S_x 恢复输出层的通道数。虽然这在一定程度上减少了参数量，但是相邻层级间参数的减少过多，必然会损失一定的细节。

针对上述两个问题，本文提出了一种新的 LSSD 网络构架。

针对输出层,相比于 SSD,提出的 LSSD 的变化主要体现在 Conv6 及后面的拓展层上。具体地,在 SSD 网络构架中,Conv6 ~ Conv7 时使用 Conv:1 * 1 * 1024 对增强特征的语义并没有什么影响,得到的 Conv7:19 * 19 * 1 024 中具有较多的低级特征。即 Conv6 的网络通道数已经达到最高,但是此时的层级仍然不是很高,特征的表征性不够好。为此,在本文提出的 LSSD 网络构架中,采用了 Conv:1 * 1 * 512,得到 Conv7:19 * 19 * 512,从而减少低级语义特征的冗余。随着层次的加深,语义信息逐渐增强,同时由于每张图片中的物体数量有限,需要的输出特征也并不会过多。因此,在本文构建的 LSSD 网络中, Conv7 之后的相邻输出层的通道数均相差 2 倍,分别为:Conv8-2:10 * 10 * 256、Conv9-2:5 * 5 * 128、Conv10-2:3 * 3 * 64、Conv11-2:1 * 1 * 32,进一步去除冗余特征,使得到的特征更具表征性。

针对输出层间的卷积层,输出层间的卷积层保留原有构架,并且卷积层 Conv:1 * 1 * C 不对输出层进行通道下降,卷积层 Conv:3 * 3 * C - S_x 通道数和 Conv:1 * 1 * C 的通道数保持两倍下降。具体地,在 Conv7 ~ Conv8-2 之间使用 Conv:1 * 1 * 512、Conv:3 * 3 * 256 - S_2; 在 Conv8-2 ~ Conv9-2 之间,使用 Conv:1 * 1 * 256、Conv:3 * 3 * 128 - S_2; 在 Conv9-2 ~ Conv10-2 之间,使用 Conv:1 * 1 * 128、Conv:3 * 3 * 64 - S_1; 在 Conv10-2 ~ Conv11-2 之间,使用 Conv:1 * 1 * 64、Conv:3 * 3 * 32 - S_1。这些改进的网络构架,使得相邻输出层间的特征传递得到良好的保留。

1.2 BAM 模块

为了在网络分类头部提取出更为有效的特征,在分类头部网络中增加了 BAM 模块,如图 2 所示,这是一种注意力机制,在该模块中,输入特征经过 Maxpool 和 AvgPool 后,得到 1 * 1 * C 的特征,然后通过 Conv、Relu、Conv 后再次得到 1 * 1 * C 的特征,这两次 Conv 主要是为了降低参数量,将得到的特征进行相加,通过 Sigmoid 得到最终的输出特征。由于 SMHM-LSSD 网络头部中的通道数为 128,通道数量并不多,所以本文在 Conv 的通道数上并未按 BAM 通用设置除以 16,而是除以 8,从而使得在参数下降的同时,保留更多的细节特征。

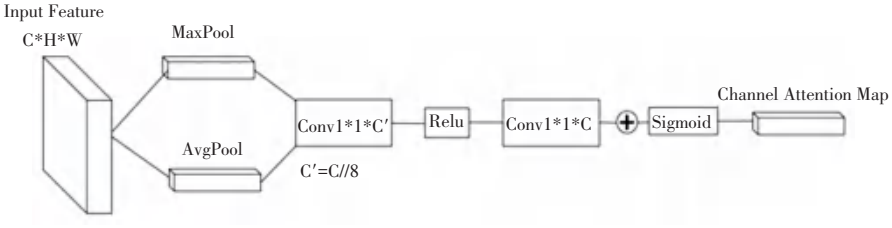

图 2 改进的 BAM 模块

Fig. 2 Improved BAM module

1.3 SMHM 模块

在 SSD 网络构架中,针对 6 个输出层,有 6 个网络头部,其中每个网络头部的分类和定位层共享特征参数,同时 6 个网络头部的 anchor 数量有所不同,分别为 4 和 6。在检测中,分类和定位是两个不同的任务,对特征的要求不同,其中分类更偏向于对特定区域敏感,而定位偏向于对物体边界更加敏感。如果将分类和定位分支分开,分别进行训练,这无疑会增加参数量。可能的尝试是将 SSD 的 6 个网络头部合并成一个网络头部使得参数量减少,合并成一个网络头部后,如果统一取 4 个 anchor,必然会影响 SSD 中原本 6 个 anchor 的网络头部的特征提取;如果统一取 6 个 anchor,会使得低级特征的 anchor 数量过多,提取得到过多的不必要特征,从而影响检测性能。

本文针对以上的两个问题,提出了一种新颖的 SMHM 模块,如图 3 所示。在该模块中,首先使用一个 Conv:128 * 1 * 1 的卷积层 B 统一通道数,使得各输出层的输出通道为 128;将网络头部中分类分支和定位分支分开,分别提取特征,从而满足分类和定位的特性需求。在这一过程中,为了解决将 6 个网络头部变为一个网络头部带来的特征提取问题,本文提出了以下 3 个策略:

(1)使用 BAM 模块,增强特征提取。

(2)使用了两个分类头部(即 B0 对应的头部,B1 对应的头部),第一个分类头部使用了一个 BAM 模块进行特征的提取,第二个分类头部使用了两个 BAM 模块进行特征提取,从而使得这两个头部能够提取到具有不同表征性的特征。

(3)使用了自适应权重,通过将两个网络头部进行平均池化后得到的数值作为该网络头部的权重,将两个网络头部进行加权和,以提取得到最适合的特征。

另一方面，为了进一步降低参数量，本文在 SMHM 模块中的 B0、B1、B2 实施共享参数。最后，在定位分支中，使用了 4 个 Conv：128 * 3 * 3 的卷积层。

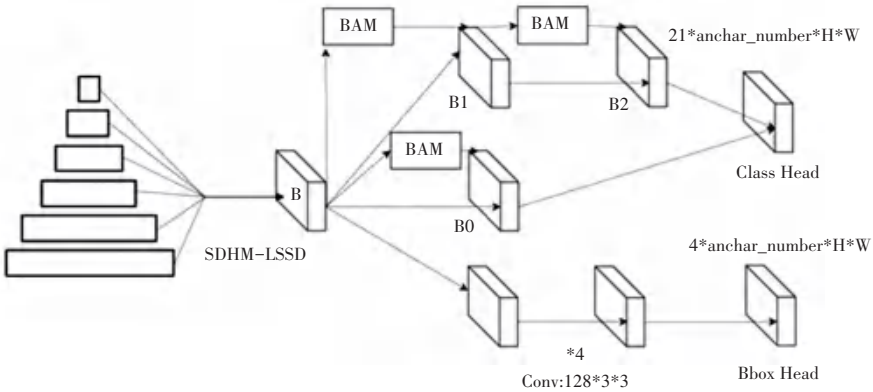

图 3　SMHM 模块

Fig. 3　SMHM module

2　实验结果与分析

2.1　实验数据集和评价指标

本文采用 PASCAL VOC 数据集验证提出的基于共享多头模块的轻量型一阶段网络的有效性。该数据库包含 20 个类别以及一个背景类别，共计 21 个类别。更为具体地，VOC2007 一共包含了9 963张图片，5 011 张为训练图片，4 952 张为测试图片；VOC2012 一共包含了 11 540 张训练图片。本次实验联合 VOC2007 和 VOC2012 训练集进行训练，在 VOC2007 测试集上进行测试。采用了 VOC2007 的评价指标，即采用均值平均精度作为评价标准。

2.2　模型与参数设置

本实验基于 Ubuntu18.04 平台，使用 PyTorch 深度学习框架以及一块 GeForce GTX 1080GPU 进行训练。具体地，主干网络为 VGG16，输入图片的大小为 300×300，每个 batch 大小为 8 张图片。使用 SGD 进行优化，初始学习率为 0.001，总共运行了 24 个 epoch，其中在运行第 16 个和第 20 个 epoch 时，下降了训练的学习率，且下降后的学习率分别为0.000 1和0.000 01。

2.3　实验结果与分析

实验结果见表 1～表 2，损失对比如图 4～图 6 所示。表 1 为当前流行的网络构架的实验结果，可以看出本文提出的 LSSD 以及 SMHM-LSSD 网络构架(表中简写为 SLSSD)有着较高的性能，超过了当前大部分的检测算法；LSSD300 相比于 SSD300 性能增加 0.2%；SLSSD300(4)相比于 SSD300 性能增加0.1%；SLSSD300(6) 相比于 SSD300 性能增加 0.6%。

表 1　Pascal VOC2007 测试结果

Tab. 1　Pascal VOC2007 test results

Method	Backbone	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
SSD300	VGG16	0.783	0.832	0.843	0.757	0.724	0.531	0.870	0.870	0.865	0.639	0.822	0.770	0.843	0.875	0.844	0.807	0.539	0.789	0.805	0.875	0.766
LSSD300	VGG16	0.785	0.818	0.853	0.754	0.714	0.533	0.879	0.867	0.875	0.637	0.835	0.787	0.849	0.869	0.851	0.807	0.520	0.795	0.813	0.863	0.778
SLSSD300(4)	VGG16	0.784	0.813	0.858	0.752	0.719	0.529	0.878	0.866	0.879	0.635	0.835	0.778	0.847	0.872	0.852	0.809	0.510	0.797	0.810	0.866	0.778
SLSSD300(6)	VGG16	0.789	0.826	0.861	0.765	0.728	0.567	0.858	0.871	0.881	0.627	0.842	0.796	0.844	0.878	0.863	0.807	0.540	0.770	0.814	0.881	0.765
ION300	VGG16	0.756	0.792	0.831	0.776	0.656	0.549	0.854	0.851	0.870	0.544	0.806	0.738	0.853	0.822	0.822	0.744	0.471	0.758	0.727	0.842	0.804
Faster	VGG16	0.732	0.765	0.790	0.709	0.655	0.521	0.831	0.847	0.864	0.520	0.819	0.657	0.848	0.846	0.775	0.767	0.388	0.736	0.739	0.830	0.726
Faster	Residual-101	0.764	0.798	0.807	0.762	0.683	0.559	0.851	0.853	0.898	0.567	0.878	0.694	0.883	0.889	0.809	0.784	0.417	0.786	0.798	0.853	0.720
MR-CNN	VGG16	0.782	0.803	0.841	0.785	0.708	0.685	0.880	0.859	0.878	0.603	0.852	0.737	0.872	0.865	0.850	0.764	0.485	0.763	0.755	0.850	0.810
R-FCN	Residual-101	0.805	0.799	0.872	0.815	0.720	0.698	0.868	0.885	0.898	0.670	0.881	0.745	0.898	0.906	0.799	0.812	0.537	0.818	0.815	0.859	0.799
DSSD321	Residual-101	0.786	0.819	0.849	0.805	0.684	0.539	0.856	0.862	0.889	0.611	0.835	0.787	0.867	0.887	0.867	0.797	0.517	0.780	0.809	0.872	0.794
ASSD	VGG16	0.800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R-SSD(4)	VGG16	0.762	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R-SSD(6)	VGG16	0.770	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
YOLOV2	DarkNet-19	0.737	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DSOD	DS/64-192-48-1	0.777	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FSSD	VGG16	0.788	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RefineDet320	VGG16	0.800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PELEE	VGG16	0.709	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 2 每个分类器网络的框数和总框数以及参数和测试速度

Tab. 2 The number of frames and the total number of frames of each classifier network, as well as the parameters and test speed

Network	38 * 38	19 * 19	10 * 10	5 * 5	3 * 3	1 * 1	total bbox	Speed	Params
SSD300	4	6	6	6	4	4	8 732	33	210.3 M
LSSD300	4	6	6	6	4	4	8 732	38	187.2 M
SLSSD300(4)	4	4	4	4	4	4	7 760	41	181.6 M
SLSSD300(6)	6	6	6	6	6	6	11 640	31	182.0 M

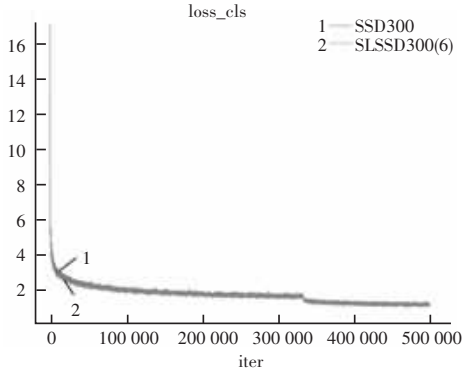

图 4 分类损失图

Fig. 4 Classification loss graph

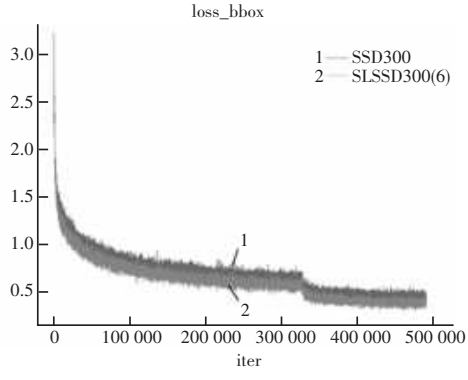

图 5 定位损失图

Fig. 5 Location loss graph

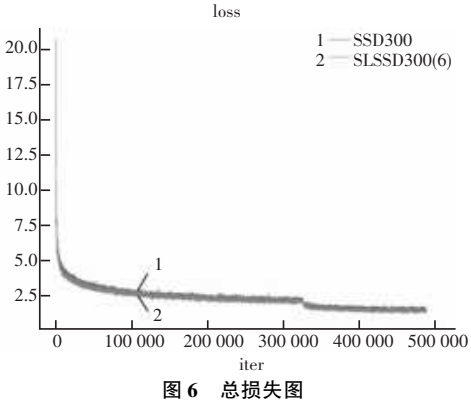

图 6 总损失图

Fig. 6 Total loss graph

表 2 为 SSD、LSSD 以及 SMHM-LSSD 的 anchor 数量、测试速度、参数量的对比,可以看出 LSSD300 相比于 SSD300 参数量减少 23.1 M,速度提升 5 fps/s;SLSSD300(4)相比于 SSD300 参数量减少 28.7 M,anchor 的数量减少了,速度增加 8 fps/s;SLSSD300(6)参数量减少 28.3 M,但由于 anchor 数量的增加,使得在速度上减少 2 fps/s。

图 4~图 6 为 SSD 和 SMHM-LSSD(6)的损失对比图。图 4 为分类损失图,可以看出两者的分类损失值基本上重合,但是相比于 SSD,SMHM-LSSD(6)的分类损失更加的稳定,究其原因是由于双头网络和 BAM 模块的使用使得提取得到的特征更加准确;图 5 为定位损失图,可以明显的看出将分类分支和定位分支进行分开训练,分支各自拥有参数后,定位分支能够得到更好的收敛;图 6 为总的损失值,可知 SMHM-LSSD(6)的收敛性相比于 SSD 更好。

3 结束语

本文针对经典单阶段网络构架 SSD 在拓展层上输出层参数量过多,输出层间特征减少过多的问题,提出了 LSSD 网络构架;针对分类定位卷积层共享特征的问题,在减少参数量的基础上,使用了头部共享 SMHM,提出了两个分类头部、运用了 BAM 模块以及使用参数共享和自适应权重的策略,同时解决了由头部共享后 anchor 数量变动导致特征提取性能下降的问题;本文提出的 SMHM-LSSD 相比于 SSD,在性能、速度、参数量上都得到了改善,并且相比于别的经典网络构架在性能上还是非常的优异。

参考文献

[1] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.

[2] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[3] SONG G, LIU Y, WANG X. Revisiting the sibling head in object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11563-11572.

[4] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

(下转第 100 页)