文章编号：2095-2163(2019)06-0312-04 中图分类号：TP391.41 文献标志码：A

基于改进 Yolov3 的目标检测的研究

晏世武，罗金良，严 庆
（南华大学 机械工程学院，湖南 衡阳 421001）

摘 要：目标检测在视频监控、无人驾驶系统、机械自动化等领域起着重要作用。在如今大数据的背景下，为进一步提高 Yolov3 在不同数据集下的性能，本文以 KITTI 数据集为基础，利用重新调整 anchor 数值和增加尺度融合的方法改进 Yolov3，并通过增加数据的方法平衡类别，进一步提高 Yolov3 性能。实验结果表明，改进的 Yolov3 较原始的框架，其 mAP 提高了近 5.31%，从侧面说明改进的 Yolov3 具有较高的实用价值。

关键词：目标检测；深度学习；尺度融合；平衡类别；mAP

Research on object detection based on improved Yolov3

YAN Shiwu, LUO Jinliang, YANG Qing
(School of Mechanical Engineering, Nanhua University, Hengyang 421001, China)

【Abstract】 Object detection plays an important role in video monitoring, driverless systems, and mechanical automation. In the context of big data, traditional object detection algorithms can no longer meet the high performance requirements of people. With the rapid development of deep learning, researchers have developed high-efficiency object detection frameworks such as Yolov3. In order to further improve the performance of Yolov3 under the different datasets, this paper improves the Yolov3 performance by re-adjusting the values of anchors and increasing the scale fusion method based on the KITTI dataset, and further improving the Yolov3 performance by adding data to balance the categories. The experimental results show that the improved Yolov3 has a mAP increase of nearly 6% compared with the original frame. It shows that the improved Yolov3 has high practical value.

【Key words】 object detection; deep learning; scale fusion; balance categories; mAP

0 引 言

目标检测能够对图像或视频中的物体进行准确分类和定位，在监控、无人驾驶、机械自动化等领域中起着至关重要的作用。早前的目标检测是通过人工提取特征的方法，使用 DPM 模型，并在图像上进行窗口滑动的方法进行目标的定位，这种方法十分耗时且精度不高。随着信息时代的快速发展，如今的数据量成几何式地增长，再使用人工提取特征的方式是十分不明智的。自 2012 年 Alexnet[1] 在 ILSVRC(Large Visual Recognition Challenge)比赛中大放光彩以来，学者们不断地使用卷积神经网络[2-4](Convolution Neural Network，CNN)设计新的目标检测框架，并出现了 Faster RCNN[5-7]、SSD[8]、Yolov3[9-11]等高性能的目标检测框架，并且在实践中展现出强大性能。

在如今较为主流目标检测框架中，Yolov3 在检测速度和精度的平衡性方面表现较好，人们不断在各种领域使用 Yolov3 实现目标检测功能。然而原始的 Yolov3 架构并不能在各种数据集下均表现出色，对于小目标物体会出现定位不准确的和漏检的情况。本文针对 Yolov3 的问题，设计以下改进方法：

(1)针对目标定位不准确的问题，对于不同的数据集，重新调整 anchor 的数值；

(2)针对小目标难检和漏检的情况，增加一个尺度融合；

(3)通过增加较少类别的物体数的方式平衡类别来优化 Yolov3。

1 Yolov3 及其改进方式

1.1 Yolov3 框架

Yolov3 是目标检测算法之一，是基于回归的方式进行特征提取，通过端到端的过程训练网络，最终在多尺度融合的特征层中回归出目标的类别与位置。端到端的训练方式使得分类与定位过程为一体。其两者共同的损失函数参与反向传播计算，在节约特征提取时间的同时又提升了精度，满足了目标检测的实时性需求。Yolov3 目标检测框架如图 1

作者简介：晏世武(1994-)，男，硕士研究生，主要研究方向：深度学习、机器学习、图像算法；罗金良(1968-)，男，博士，教授，主要研究方向：机械设计及理论；严 庆(1993-)，女，硕士研究生，主要研究方向：机械设计及理论。

收稿日期：2019-09-18

所示。

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
	Convolutional	32	1 × 1	
1×	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
	Convolutional	64	1 × 1	
2×	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
	Convolutional	128	1 × 1	
8×	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
	Convolutional	256	1 × 1	
8×	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
	Convolutional	512	1 × 1	
4×	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

图 1　Yolov3 框架[9]

Fig. 1　The framework of Yolov3[9]

图 1 中提取曾为 Darknet−53 的网络结构，该结构以 256×256 图片作为输入，大量使用 1×1 和 3×3 的卷积层进行堆砌，并使用残差网络[12]（如图 2 所示）将浅层信息传递到深层，可在增加网络深度的同时不引起梯度爆炸等问题，图 1 中最左边的数字即代表所重复的残差网络模块的个数；Yolov3 结构在检测方面采用的是多尺度检测策略，使用 32×32、16×16、8×8 三个不同尺寸的特征图进行检测输出。原图进行尺寸映射到检测特征层的每个点上，且每个点有 3 个预测框，因此在三个特征层检测上共有 4 032 个预测框，该预测数极大满足了检测多类物体的需要。最终使用 logistic 回归，对每个预测框进行目标性评分，根据目标性评分来选择满足需求的目标框，并对这些目标框进行预测。

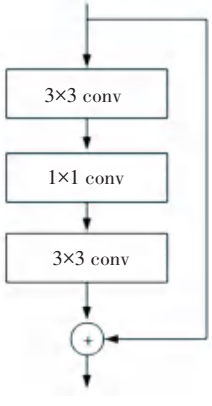

图 2　残差网络

Fig. 2　Residual network

1.2　anchors 的设置

在 Yolov3 目标检测框架中 anchor 十分重要，它是由当前数据集通过 kmeans 聚类算法[13]统计出来的，合适的 anchor 值能够降低网络架构的损失值，加快收敛。在原始的 Yolov3 网络层中，用于检测物体的特征层为 32x32、16x16、8x8 大小的特征提取层，这些特征层可以映射到原始图像，即原始图像被切分为对应特征层的网格大小（grid cell）。如果真实框（ground truth）中某个物体的中心坐标落在 grid cell 里，就由该 grid cell 预测该物体，并且每个 grid cell 预测 3 个边界框，其边界框的大小由 anchor 值决定，然后对预测的边界框与真实框的交互比（IOU）来选出超过 IOU 值的边界框去进行检测，为进一步减少不必要的检测次数，使用设置目标置信度的方法，当预测框的置信度小于该设定值就不再去检测该框。

本文训练使用的 KITTI 数据集，而原始 Yolov3 中的 anchor 值是使用 COCO[14]数据集得到的。因此，为提升本文物体的定位精度，重新使用 kmeans 算法去统计是十分必要的，且由于 KITTI[15]数据集的图片较大，本文将 Yolov3 的初始图片大小设为 608×608，其对应特征层的大小也相应的会改变。COCO 数据集和 KITTI 的 anchor 值分别见表 1、表 2。

表 1　COCO 数据集的 anchor 值

Tab. 1　The value of anchor in COCO dataset

特征层	特征图大小	anchor 值
特征层 1	8×8	(116,90)、(156,198)、(373,326)
特征层 2	16×16	(30,61)、(62,45)、(56,119)
特征层 3	32×32	(10,13)、(16,30)、(33,23)

表 2　KITTI 数据集的 anchor 值

Tab. 2　The value of anchor in KITTI dataset

特征层	特征图大小	anchor 值
特征层 1	19×19	(89,157)、(121,267)、(181,346)
特征层 2	38×38	(23,168)、(55,100)、(44,280)
特征层 3	76×76	(16,41)、(11,98)、(31,66)

1.3　多尺度检测

Yolov3 目标检测框架中使用了多尺度检测，即上文所提到的 19×19、38×38、76×76 三个特征层同时检测图像或视频中的物体，且根据 anchor 中的值预先画出预测边界框。这种方式对中大型物体具有很好的检测效果，但是对于小物体存在难检或漏检的情况。本文针对 KITTI 数据集，增加一个特征尺度以提升检测精度。三尺度与四尺度检测模型如图 3、4 所示，由于增加了一个特征尺度，则 anchor 值也

需要重新调整,见表 3。

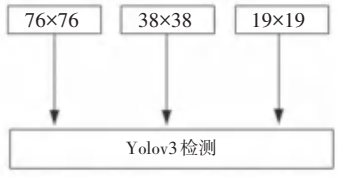

图 3 三尺度检测

Fig. 3 The three scale fusion of detection

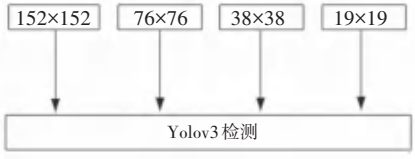

图 4 四尺度检测

Fig. 4 The four scale fusion of detection

表 3 KITTI 数据集的 anchor 值

Tab. 3 The value of anchor in KITTI dataset

特征层	特征图大小	anchor 值
特征层 1	19×19	(115,207)、(136,307)、(193,354)
特征层 2	38×38	(55,126)、(87,139)、(59,295)
特征层 3	76×76	(34,89)、(55,74)、(29,222)
特征层 4	152×152	(14,41)、(26,55)、(13,118)

1.4 平衡数据类别

本文训练的数据集为 KITTI,它是由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办,是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集。标注了九个类别的物体,分别为 Car、Van、Truck、Pedestrian、Person_sitting、CyClist、Tram、Misc、DontCare。由于车辆的数据集较多而其它的数据较小,有结合 CNN 需要大量数据集的特点。本文对 KITTI 数据集中类别进行合并,合并策略如下:

(1) Car、Van、Truck、Tram 合为一类,记为 Vehicle;

(2) Pedestrian、Person_sitting 合为一类,记为 Person;

(3) Cyclist 自成一类;

(4) 忽略 Misc 和 DontCare。

三个类别的数量见表 4。

表 4 每个类别图像数量

Tab. 4 The number of images in every label

类别	数量/张
Vehicle	33 261
Person	4 709
Cyclist	1 627

从表 4 可看出 Vehicle 图像数量充足,而 Person 和 Cyclist 类别的数量过少,不利于 CNN 提取特征。因此本文增加 INRIAPerson 数据集以扩充 Person 数据集,从百度图片中下载 Cyclist 图片并标注以扩充 Cyclist 数据集,扩充后的数据集情况见表 5。

表 5 扩充后每个类别图像数量

Tab. 5 The value of anchor in KITTI dataset after data augumentation

类别	数量/张	增加数量/张
Vehicle	33 261	81
Person	4 709	1 349
Cyclist	1 627	1 717

2 实 验

2.1 实验设备

本文使用的基础 Yolov3 网络架构是 Darknet53,并使用 GPU 对训练网络进行加速计算,其使用的计算机参数见表 6。

表 6 实验用计算机配置参数

Tab. 6 The parameters of the experiment computer

硬件	型号
处理器	AMD Ryzen 5 2600X 六核
内存(16 G)	威刚 DDR4 2 667 MHz
主硬盘	NVMe SSDPEKKW25
显卡	Nvidia GeForce GTX 1060 6 GB

2.2 实验结果与分析

本文实验通用参数见表 7。首先使用 Yolov3 原始结构训练并测试未扩充前 KITTI 数据;重新计算 anchor 值后,重新训练并测试数据;增加一个检测尺度后;训练测试 KITTI 数据集;扩充 KITTI 数据集,重新训练后测试数据。测试结果见表 8。

表 7 通用实验参数

Tab. 7 The parameters of the basic experiment

参数名称	参数值
batch	32
subdivisions	32
momentum	0.9
decay	0.000 5
learning_rate	0.001

表 8 实验结果

Tab. 8 Experiment result

实验	mAP
原始 Yolov3 架构+未扩充前数据	0.738 1
原始 Yolov3 架构+重新计算 anchor 值	0.771 7
增加一个检测尺度+重新计算 anchor 值	0.781 9
四尺度 Yolov3 架构+数据扩充	0.791 2

实验分析：原始 Yolov3 架构的 anchor 值是由 COCO 数据集得到的，当为新的数据集时，使用 kmeans 算法重新计算 anchor 值，能够提升数据集的 mAP 值；对于 KITTI 数据集中的小目标物体，增加一个检测尺度并重新计算 anchor 值，能够提升网络架构的性能；当数据集中类别中图像过少可以进行数据扩充来提升数据集的 mAP 值。其中四尺度 Yolov3 的检测效果如图 5 所示。

图 5　检测效果

Fig. 5　The effect of detection

3　结束语

（1）Yolov3 是目标检测框架中较为优秀的架构，仅使用原始的 Yolov3 架构去训练和测试未扩充前的 KITTI 数据，其 mAP 值约为 0.7381；

（2）重新计算 anchor 值，对于 KITTI 数据集的精度有近 0.0336 的提升。因此，当使用 Yolov3 架构训练不同的数据集时，需要重新计算 anchor 值，它对 Yolov3 性能有较大的提升作用；

（3）增加 Yolov3 网络结构的检测尺度，对于小目标有较为不错的检测效果，其 mAP 值达到 0.7819，提升近 0.0102 的效果；

（4）CNN 训练需要大量的数据，当图像数据较少时，可适当地增加数据，最终其 mAP 值为 0.7912，对于最初的实验结果来说有近 0.0531 的提升，具有较为理想的检测效果。

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]. International Conference on Neural Information Processing Systems. Curran Associates Inc, 12:1097-1105.

[2] 刘健，袁谦，吴广，喻晓.卷积神经网络综述[J].计算机时代，2018(11):19-23.
LIU J, YUAN Q, WU G AND YU X. Summary of Convolutional Neural Networks[J]. Computer Ara, 2018(11):19-23.

[3] 杨真真，匡楠，范露，康彬.基于卷积神经网络的图像分类算法综述[J].信号处理，2018，34(12):1474-1489.
YANG Z Z, KUANG N, FAN L, et al. Summary of Image Classification Algorithms Based on Convolutional Neural Networks[J]. Jouranal Of Signal Processing, 2018,34(12):1474-1489.

[4] 李策，陈海霞，汉语，左胜甲，赵立刚.深度学习算法中卷积神经网络的概念综述[J].电子测试，2018(23):61-62.
LI C, CHEN H X, HAN Y, et al. Summary of Concepts of Convolutional Neural Networks in Deep Learning Algorithms[J]. Electronic Test, 2018(23):61-62.

[5] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] IEEE Conference on Computer Vision and Pattern Recognition, 2014:280-587.

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

[7] 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.

[8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C] European Conference on Computer Vision. 2016.

[9] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.

[10] REDMON J , FARHADI A . [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger[J]. 2017:6517-6525.

[11] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement [J]. 2018.

[12] HE K , ZHANG X , REN S , et al. Deep Residual Learning for Image Recognition [J]. Conference on Computer Vision and Pattern Recognition, 2015:770-778.

[13] ARTHUR D, VASSILVITSKII S. k-means++:the advantages of careful seeding[C] Eighteenth Acm-siam Symposium on Discrete Algorithms. 2007.

[14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[J]. 2014.

[15] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C] IEEE Conference on Computer Vision & Pattern Recognition. 2012.