文章编号：2095-2163(2020)11-0126-06 中图分类号：TP391 文献标志码：A

基于金字塔模型和注意力机制的 YOLO V3 目标检测算法

高建瓴，冯娇娇，王子牛，韩毓璐，孙 健
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：YOLO V3 作为一种多尺度的目标检测算法，结构简单，可快速检测。但在训练过程中，由于底层的卷积层中包含较多的目标细节信息，而通过不断的卷积，底层的信息会被逐渐淡化掉。针对于此，本文利用改进的金字塔结构，将浅层特征图的细节信息和高层的语义特征信息进行融合，从而提高检测效果。在特征提取网络的残差连接中加入注意力机制，使得具有注意力效果的梯度能流入更深的网络中，通过在 YOLO V3 中加入注意力机制提升检测精度。即提出一种基于金字塔模型和注意力机制的 YOLO V3 目标检测算法。将改进后的算法和原有算法在 COCO 2014 数据集上进行对比实验，结果表明改进后的 YOLO V3 算法在 mAP 上有所提升，证明了它的可行性。

关键词：注意力机制；目标检测；YOLO V3；金字塔结构

YOLO V3 target detection based on pyramid model and attention mechanism

GAO Jianling, FENG Jiaojiao, WANG Ziniu, HAN Yulu, SUN Jian
(School of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 As a multi-scale target detection algorithm, YOLO V3 has a simple structure and is relatively fast in detection speed and accuracy. Therefore, a YOLO V3 target detection algorithm based on pyramid model and attention mechanism is proposed. In the training process, the bottom convolutional layer often contains more target detail information, and through continuous convolution, the bottom information will be gradually diluted. Therefore, this article uses an improved pyramid structure to reduce the feature map of the lower layer. Detailed information and high-level semantic feature information are fused to improve the detection effect. The attention mechanism is added to the residual connection of the feature extraction network, so that the gradient with attention effect can flow into the deeper network, and the accuracy is improved by adding the attention mechanism to YOLO V3. Using the improved algorithm and the original algorithm to conduct a comparative experiment on the COCO 2014 data set, the results show that the improved YOLO V3 algorithm has improved on mAP, which proves its feasibility.

【Key words】 Deconvolution; target detection; YOLO V3; pyramid structure

0 引 言

目标检测[1-3]是计算机视觉领域中一个具有挑战性的课题[4]，主要目的是从视频或者一些数据集中检测并定位特定的目标。由于目标所在的环境影响，至今还没有一种比较通用且成熟的检测方法。传统的目标检测算法已经难以满足现有数据的处理效率、速度和精度等各方面的要求。

随着网络的发展，陆陆续续出现很多基于深度学习[5]方法思想的目标检测模型。例如，一种是基于区域提取的 R-CNN 系列[6-8]为代表的目标检测，其次就是基于回归的目标检测算法，包括 YOLO[9]（You Only Look Once）和 SSD[10]（Single Shot Multi-BoxDetector）。在基于回归的方法中，YOLO 第一个采用了回归思想，实现了 one-stage 检测的算法[11]。如今该算法已经从 YOLO 经 YOLO V2 发展到现在 YOLO V3[12]，不论在检测速度方面还是在精度上，已远远超过第一代的 YOLO。相比作为后辈的 SSD 法，性能也得以超越。但是 YOLO 算法依旧存在一些问题，比如定位精度、召回率等较低，且对尺寸较小的物体检测效果欠佳。

YOLO V3 利用特征金字塔[13]（Feature Pyramid Network, FPN）的思想，将相邻尺度的特征图通过串联（concat）操作进行特征融合[14]，通过特征融合可以有效提高目标的检测精度。

本文基于 YOLO 存在的问题，以 YOLO V3 模型为主要架构，提出了一种基于金字塔模型和注意力机制的 YOLO V3 目标检测算法，建立了多尺度的目标检测算法。通过反卷积和金字塔模型，来融合不同尺度的浅层和高层的信息，从而提升对目标的检测精度。利用注意力机制，使得具有注意力效果的

作者简介：高建瓴（1969-），女，硕士，副教授，主要研究方向：数据分析、数据库应用；冯娇娇（1996-），女，硕士研究生，主要研究方向：信息与通信工程；王子牛（1961-），男，硕士，副教授，主要研究方向：计算机应用系统、信息系统；韩毓璐（1996-），女，硕士研究生，主要研究方向：信息与通信工程；孙 健（1996-），男，硕士研究生，主要研究方向：电子与通信工程。

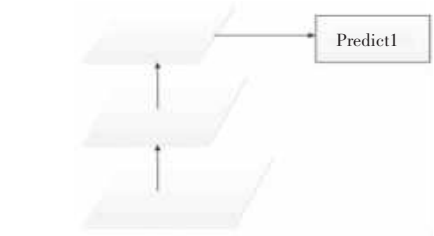
收稿日期：2020-09-08

梯度能流入更深的网络中。实验结果表明，改进后的 YOLO V3 比原始的 YOLO V3 在 COCO 2014 数据集上具有更高的目标检测精度。

1　相关工作

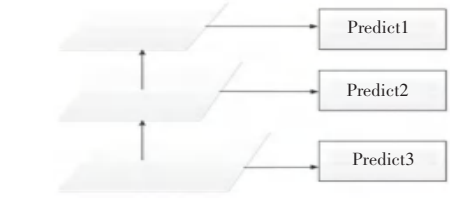
1.1　特征金字塔

通过构建特征金字塔网络，解决了众多尺度检测问题，代替了 Faster R-CNN 之类的检测模型的特征提取器，生成多层特征映射，信息的质量比普通的用于特征检测的特征金字塔更好。如图 1 所示。



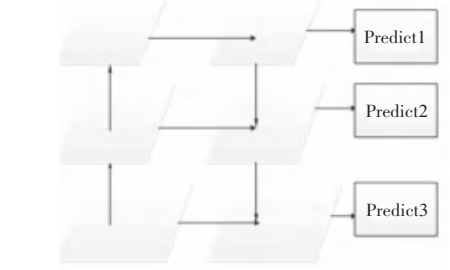
（a）单层特征层预测模型

（a）Single layer characteristic layer prediction model



（b）多层特征层预测模型

（b）Multi-layer feature layer prediction model



（c）特征金字塔模型

（c）Feature pyramid model

（d）跳跃连接金字塔结构

（d）Jump connection pyramid structure

图 1　预测网络

Fig. 1　Predicting network

图 1 中（a）、（b）、（c）是现在常有的金字塔结构，（d）是本文采用的金字塔结构。图 1（a）是检测系统选择使用于更快速检测的单尺度特征，它自底向上卷积，使用最后一层特征图进行预测。如 SPP-Net、Faster R-CNN 就是采用这种方式，即仅采用网络最后一层的特征。该方式虽然检测速度快，但会造成检测小目标时的精度下降。图 1（b）的改进是将图片分成不同尺寸，然后进行不同尺度的预测，有效的改善了图 1（a）中的单一预测。图 1（c）是自底向上然后自顶向下的线路，横向连接，同时融合了相邻特征之间的信息。通过高层特征进行上采样和浅层特征进行自顶向下的连接，而且每一层都会进行预测。其问题是它忽略了高层和浅层的联系。所以，针对于此提出了图 1（d）的改进金字塔结构模型。该模型采用跳跃连接的方式，同时采用 4 倍和 2 倍的反卷积进行上采样，从而得到相同的特征图大小，有效的解决了图 1（c）的问题。改进的金字塔详细结构如图 2 所示。

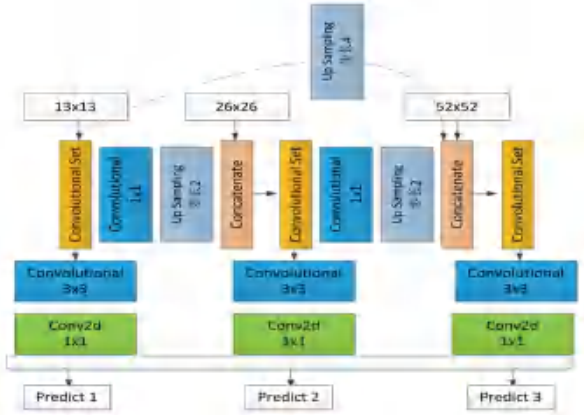


图 2　改进的金字塔的细节结构

Fig. 2　The detailed structure of the improved pyramid

从图 2 中可以清晰看到 3 个预测层分别来自何处。YOLO V3 的融合通过 catenate 进行张量拼接，在进行一个 Convolutional Set 操作时，其结构如图 3 所示。在大小为 13 * 13 和 52 * 52 的特征图之间，采用步长为 4 的反卷积进行上采样，使其不同特征层的大小一致，最后输出检测结果。

1.2　注意力机制

注意力机制[15]（Attention mechanism）可以按不同的形式分类。按照注意力产生的方式可分为两种：一种是自顶向下的注意力或者称聚焦式注意力，还有一种是自底向上的注意力，称为基于显著性的注意力。若按照注意力作用的特征形式，可分为基于项的注意力和基于位置的注意力。在计算机视觉领域中，基于位置的注意力是与任务相关、作用方法直接的注意力机制，其应用比较广泛。

1.2.1 空间注意力机制[16](Spatial Transformer Network,STN)

STN 网络模型如图 3 所示。其中包含 3 个部分:定位网络(Localisation network)、网格生成器(Grid generator)、采样器(Sampler)。

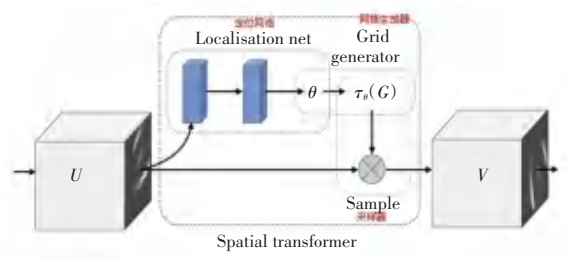


图 3 STN 模型结构

Fig. 3 STN model structure

定位网络用来生成仿射变换的系数,其中输入为 U,输出为 θ,θ 是一个可变换参数,其维度大小根据变换类型而定。

网格生成器是根据上面生成的参数 θ,对输入进行变换。从而得到原始图像或者是经过平移、旋转等变换的特征图结果。

采样器根据网格生成器得到的结果,生成一个新的输出图片或者特征图 V,用于下一步操作。

1.2.2 通道注意力(Channel Attention,CA)

通道注意力可以理解为让神经网络看什么,典型的代表是 SENet[17]。卷积网络的每一层都有很多卷积核,每个卷积核对应一个特征通道。相对于空间注意力机制,通道注意力在于分配各个卷积通道之间的资源。网络结构如图 4 所示。

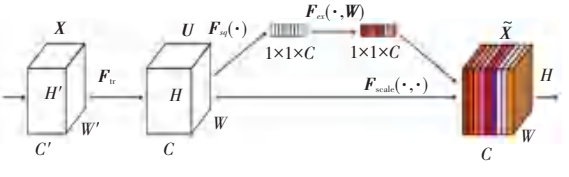


图 4 SE 块

Fig. 4 SE block

1.3 YOLO V3 网络

1.3.1 网络结构

相对 YOLO V1 和 YOLO V2 的特征提取网络,YOLO V3 的网络结构是 Darknet-53。对比同期其它的特征提取网络拥有较好的表现。YOLO V3 网络结构如图 5 所示。

从图 5 中可以看出,YOLO V3 的 Darknet-53 网络不同于 YOLO V2 的 Darknet-19 网络。Darknet-53 网络中连续使用 1x1 和 3x3 的卷积核,一共有 53 个卷积层和 5 个残差块。其中,YOLO V3 通过卷积层,对输入图片进行了 5 次下采样,最后进行了多尺度的预测,处理过程借鉴了残差神经网络的思想[18]。Darknet53 结构单元如图 6 所示。

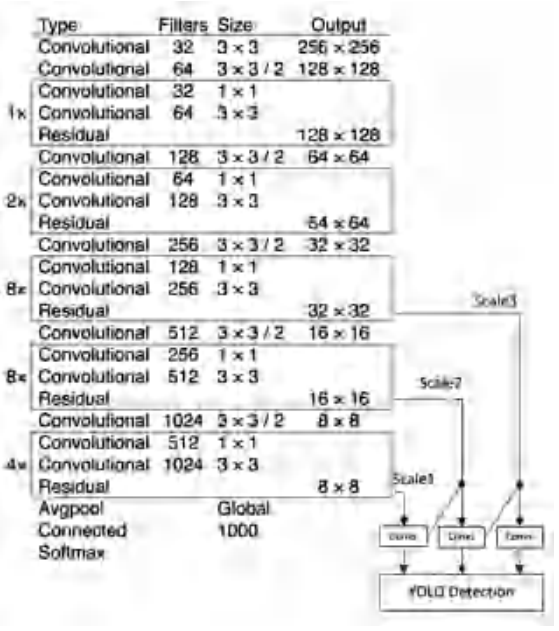
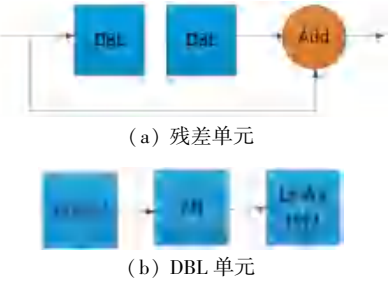


图 5 YOLO V3 网络结构

Fig. 5 YOLO V3 network structure



(a) 残差单元

(b) DBL 单元

图 6 Darknet53 结构单元

Fig. 6 Structural unit of Darknet53

图 6(a)中,每个残差块由多个残差单元组成,通过输入两个 DBL 单元进行残差操作,构建了残差单元。其中,DBL 单元包含卷积、批归一化和 leaky relu 激活函数,如图 6(b)所示。通过引入残差单元使得网络深度可以更深,避免梯度消失。

1.3.2 k-means 维度聚类算法

通过 k-means 聚类算法得到更好的边界框,对网络来说,则能学习到准确的预测方法。所以 YOLO V3 继续延用了 YOLO V2 的思想,采用 k-means 的方式对训练集的边界框做聚类,找到合适的边界框。YOLO V3 在 VOC 和 COCO 数据集上聚类,边界框的数量为 9。因是对边框做相应的聚类,所以聚类方法中的距离公式定义为:

$$d(box,centroid) = 1 - IOU(box,centroid). \quad (1)$$

1.3.3 检测过程

YOLO V3 是基于回归的目标检测,不需要预先

生产感兴趣区域。如图 7 所示，首先输入尺寸为 416x416 的图片，经过 Darknet53 网络进行特征提取，可以获得 3 个不同尺寸的网格区域，其分别是 13×13、26×26、52×52。经过特征金字塔，在网络后两个特征图上采样，与网络前期相应尺寸的特征图聚合，最后再预测输出结果。

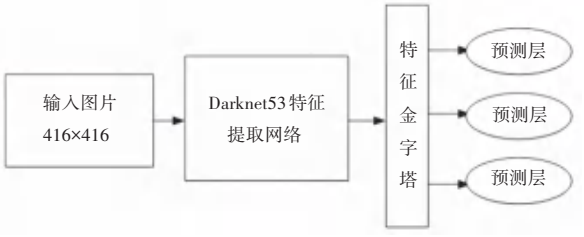


图 7　检测流程图

Fig. 7　Detection flow chart

2　本文算法

2.1　加入通道注意力机制

通过对神经网络中传递的特征通道加以不同的权重，网络可以更加重视权重较大的通道进行参数更新。由于全局平均池化[19]不同于平均池化和最大池化，它是一种特殊的池化，通常用来聚集空间信息。它是对特征图全局平均后输出一个值，从而使其具有全局的感受野，使得网络浅层也能利用全局信息。通道注意力机制的用途是通过对特征图的各个通道之间的依赖性进行建模，从而提高对于重要特征的表征能力。

在 YOLOv3 的网络结构中含有很多的残差连接，所以在残差块中加入通道注意力机制，它具有挤压(Squeeze)和激活(Excitation)两部分操作。设输入注意力模块的卷积核集合为 $Y=[y_1,y_2\cdots,y_c]$，卷积操作为 F_{conv}，前一层的卷积核集为 $X=[x_1,x_2\cdots,x_c]$，H、W、C 分别为特征图的长度、宽度、通道数。挤压操作为F_{sq}，也就是全局池化过程。设输出结果为长度为 c 的一维数组，具体操作公式如式(2)：

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}y_c(i,j)=F_{sq}(y_c). \qquad (2)$$

然而激励过程就是对各通道间的相关度进行建模，公式如下：

$$s=Sigmoid(FC_2\times ReLU(FC_1\times z_c)). \qquad (3)$$

式中，FC_1 和 FC_2 是全连接层，s 是最终激活操作的输出参数为向量。通过上述 2 个公式，在经过特征加权，可以得出最终的输出结果为：$Y+y_c\times s_c$。

相对之前的 YOLO V3，由于引入了注意力机制，增加了两个全连接层和一个全局平均池化以及特征加权，这大大的添加了算法的计算量，在速度上会比原 YOLO V3 低。其算法流程如图 8 所示。

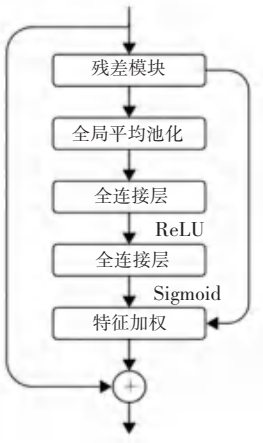


图 8　改进算法流程图

Fig. 8　Improved algorithm flow chart

2.2　特征金字塔融合的改进

在 YOLO V3 中，引入了 FPN 网络，将浅层的信息和高层的信息通过使用上采样融合，从而得到 3 个不同尺度的特征图，再进行检测输出结果。由于原 YOLO V3 中忽略了浅层信息和高层信息二者之间的联系，所以本文对网络进行了相应的改进，如图 9 所示。

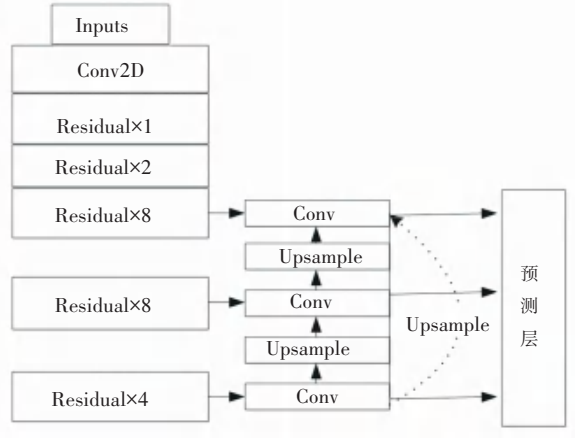


图 9　改进的 YOLO V3 检测模型图

Fig. 9　Improved YOLO V3 detection model diagram

在原 YOLO V3 的 3 个尺度检测上，把浅层的信息和高层的信息进行反卷积。改进的具体细节是：输入图片经过 Darknet-53 得到特征图，在输出尺度为 13x13 时执行上采样操作，再和 26x26 融合，得到新的特征信息，再把融合后的信息与 52x52 的输出融合。之后，对 13x13 做反卷积与 52x52 进行融合，因此实现了浅层和高层二者信息的联系。3 个尺度在经过上采样和反卷积以及特征融合，得到了新的 3 组不同尺度的 YOLO 特征层，并使用这 3 组特征

层进行位置和类别的预测。

3 实验结果分析

3.1 实验环境和数据集

实验环境配置见表 1。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

名称	配置
显卡	GeForce RTX2080Ti,11GB
GPU	CUDA11.0,CUDNN 7.0
深度学习框架	Tenorflow,pytorch
操作系统	Linux: openSUSE Leap 42.3

本实验的数据采用 COCO_2014 数据集,使用该数据进行预训练得到该数据集的预测模型。数据集大约包含 16 万张图片,图片包括了动物图片、自然风景图片、生活街景图片和人物图片等。图片中背景都较复杂,目标数量较多,大量图片样本中具有多目标物体。其中用于训练为 117 263 张图片,测试集共有 5 000 张图片。

3.2 实验结果分析与对比

在 COCO 2014 上对已经训练好的网络进行测试,分别计算了 YOLO-tiny、YOLO V3 和改进的 YOLO V3 模型的平均准确率均值(mAP),作为检测效果评价指标。测试结果见表 2。YOLO-tiny 和 YOLO V3 模型的 mAP 分别是 16.3%和 31.1%,改进后的 YOLO V3 模型的平均精准度为 32.4%,提升了 1.3%。

表 2 COCO 2014 测试集中平均准确率(mAP%)测试结果

Tab. 2 COCO 2014 test set average accuracy rate (mAP%) test results

Iou 取值	0.5	$0.5:0.95^{-1}$
YOLO-tiny	34.2	16.3
YOLO V3	55.2	31.1
Ours	55.5	32.4

本文对特征金字塔融合的改进以及添加的注意力机制,使其对数据集中的小目标有了一定的提升。在 80 类中选择了 7 类进行做对比,结果见表 3。

表 3 COCO 2014 测试集中不同目标的平均准确率(mAP%)

Tab. 3 The average accuracy of different targets in the COCO 2014 test focuses (mAP%)

算法	bicycle	cat	bird	bottle	cow	bed	horse
YOLO-tiny	27.1	60.4	23.3	16.1	38.4	50.9	57.7
YOLO V3	44.3	80.5	45.2	41.5	59.1	70.7	77.7
Ours	49.1	81.2	46.6	44.2	62.3	71.8	81.0

从表 3 可知,改进后的模型对目标的精准度有了相应的提高,对于 cat、bird、cow、bed 等目标的精准度均比原来的有所提高。其中每个小目标分别提升了 5.1%、0.7%、1.4%、2.7%、3.2%、1.1%、3.3%。相对于 YOLO-tiny 来说,提升的较大,因为 YOLO-tiny 简化了网络架构。

4 结束语

本文提出一种基于金字塔模型和注意力机制的 YOLO V3 算法,保留了多尺度检测的特点,同时还把浅层特征图的细节信息和高层的语义特征信息进行融合,从而提高检测效果。外加在残差连接中加入通道注意力机制,整个网络能更好地筛选出有利于后续检测的特征向量。实验结果表明,改进后的 YOLO V3 在小目标和精准度上都有提高,但在速率上还不够理想。究其原因,是在特征融合阶段加入了反卷积以及对残差网络的改进所致。下一步的工作是继续研究如何在保持目标精准度的情况下,减轻网络层,降低计算量,提升检测速度。

参考文献

[1] 张鹤,孙瑜. 基于双摄像头下的活体人脸检测方法[J]. 软件,2020,41(7):51-56.

[2] 更藏卓玛,安见才让.基于深度学习的青藏高原畜牧业多目标动物图像检索研究[J]. 软件,2020,41(7):126-131.

[3] 高建瓴,孙健,王子牛,等. 基于注意力机制和特征融合的 SSD 目标检测算法[J]. 软件,2020,41(2):205-210.

[4] ZHANG X Y, DING Q H, LUO H B, et al. Infrared dim target detection algorithm based on improved LCM [J]. Infrared and Laser Engineering, 2017, 46(7): 0726002.

[5] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition.2014, 580-587.

[7] GIRSHICK R. Fast R-CNN [C]//Proceedings of the IEEE international conference on computer vision.2015: 1440-1448.

[8] REN S, HE K, GIRSHICK R, et al.Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017 (6): 1137-1149.

[9] 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.

(下转第 136 页)