计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700122

基于区域激活策略的 Tiny YOLOv3 目标检测算法

余晗青　杨　贞　殷志坚
江西科技师范大学通信与电子学院　南昌 330000
(yuhanqing_1996@163.com)

摘　要　针对 Tiny YOLOv3 模型检测精度低的问题,提出一种将分割信息引入深度卷积神经网络结构中的方法。模型训练期间,将目标真实的位置信息加入网络层中,并手动激活这些目标区域,激励的大小随着训练的进行逐渐减小直至降为零。测试结果表明,在 VOC2007 数据集上,改进后的 Tiny YOLOv3 模型的平均准确率提升至 58.9%,并且在检测速度上与原模型保持一致,满足实时检测的需要。

关键词:Tiny YOLOv3;分割信息;深度卷积神经网络;位置信息
中图法分类号　TP391

Tiny YOLOv3 Target Detection Algorithm Based on Region Activation Strategy

YU Han-qing, YANG Zhen and YIN Zhi-jian
School of Communication and Electronics, Jiangxi Science and Technology Normal University, Nanchang 330000, China

Abstract　Aiming at the problem of low detection accuracy of Tiny YOLOv3 model, a method to introduce segmentation information into deep convolutional neural network structure is proposed. During the model training, the real position information of the target is added to the network layer, and these target areas are manually activated. The size of the excitation gradually decreases as the training proceeds until it drops to zero. The test results show that on the VOC2007 data set, the average accuracy of the improved Tiny YOLOv3 model is increased to 58.9%, and the detection speed is consistent with the original model to meet the needs of real-time detection.

Keywords　Tiny YOLOv3, Segmentation information, Deep convolutional neural network, Location information

1　引言

目标检测是计算机视觉领域的重要研究方向,经历了两个阶段的发展:1)传统目标检测方法;2)深度学习目标检测方法。传统的目标检测方法用滑动窗口技术筛选出物体可能出现的区域,再用人工设计的算法进行特征提取,最后利用分类器判定人工设计的特征以完成检测功能。上述算法的缺点是运行速度慢,又由于提取到的特征是人工设计的,因此,模型的性能很大程度上取决于特征的选取。随着深度学习方法在目标检测方向上的广泛应用,传统的人工设计特征渐渐退出历史舞台。当前,基于深度学习的目标检测方法主要分为两大类:一类是以 Faster R-CNN[1] 为代表的两阶段目标检测算法,其核心思想是先提出多个候选区域,再通过卷积神经网络[2]进行分类;另一类是以 YOLO 为典型的一阶段目标检测算法[3],其方法不再需要候选区域,而是直接将目标框定位的问题转化为回归问题处理。

从准确度、实时性和可移植性等方面综合考虑,YOLO 是目前最佳的目标检测算法模型之一,被广泛应用于工业界。YOLOv1 的主干网络借鉴了 GoogLeNet,先将输入图片均分成多个小格,每个小格负责中心点在自身区域范围内的目标,经过一系列卷积池化操作,在输出层直接输出目标物的位置坐标和所属类别[3]。该算法检测速度很快,但是检测精度和召回率较低,识别小物体性能也较差。YOLOv2 在 v1 的基础上改进了主干网络[4],并加入了一系列有效策略来提升模型的精度和召回率,然而其对小物体的检测效果无明显改善。YOLOv3 调整了网络结构,在 3 个不同尺度的输出特征图上进行预测,加强了对小物体的识别能力[5],在检测速度不变的情况下,提升了预测精度。YOLOv4[6] 在 v3 的基础上融合了最新的一些算法和训练技巧来提升神经网络的准确率,在速度上和 YOLOv3 几乎持平。Tiny YOLOv3 由 YOLOv3 精简网络结构而来,检测速度更快,满足实时检测的要求,但精度稍低。

常见的目标检测算法仅仅利用目标的真实位置信息做边框回归,却忽略了位置信息中包含的弱分割信息[7]。这些弱

基金项目:国家自然科学基金(61866016);江西科技师范大学青年拔尖项目(2018QNBJRC002);江西省教育厅一般项目(GJJ190587);江西省自然科学基金面上项目(20202BABL202014);甲骨文信息处理教育部重点实验室开放课题(OIP2019E008)
This work was supported by the National Natural Science Foundation of China(61866016), Youth Top-notch Project of Jiangxi Science and Technology Normal University(2018QNBJRC002), General Project of Jiangxi Provincial Department of Education (GJJ190587), General Project of Natural Science Foundation of Jiangxi Province(20202BABL202014) and Oracle Information Processing Ministry of Education Key Laboratory Open Project Funding Project(OIP2019E008).
通信作者:杨贞(yangzhen@jxstnu.edu.cn)

分割信息能帮助卷积神经网络更好地完成目标检测的任务。在深度卷积神经网络中引入分割信息的方法主要分两类：第一类是在检测的同时加入分割信息；第二类是将目标检测过程分为检测和分割两个独立分支，两个分支联合输出结果。后者修改了模型原有的网络结构，引入额外的参数，增加了计算量。本文提出的方法属于第一类，在网络训练时将分割信息引入检测过程，既不改变原有网络结构，也不增加额外的计算量。将此方法应用在 Tiny YOLOv3 上，在保持其速度优势的同时，提高了检测精度。

2 相关工作

YOLO 算法自提出以来，就在目标检测方向展现出了独特的速度优势。为了提升算法的精度，YOLOv2 借鉴了 Faster R-CNN 中先验框的思想[1]，使定位更加准确，在不损失速度的前提下提升了网络性能。YOLOv3 采用多尺度预测的方法[5]，在 3 个不同尺度大小的特征图上进行预测，提升了模型精度，对小目标的识别也更加准确。YOLOv4 融合了最新的算法和训练技巧提升了模型精度，但在速度上没有提升。为了满足模型的实时性要求，YOLOv3 的作者又提出了检测速度更快的 Tiny YOLOv3[5]，其速度可达 YOLOv3 的十多倍，但是精度有所下降。

一阶段算法与二阶段算法最大的不同在于前者缺少建议区域。Tiny YOLOv3 中密集的候选框不仅给目标定位带来难度，还造成正负样本严重失衡的问题，阻碍其精度进一步提升。实验验证得出，同时学习对象检测和语义分割可以分别改善检测和分割的实验结果[8-9]，但这种做法会给目标检测带来额外的计算量。基于此，研究员提出在网络中引入分割信息而不是构造额外的分割分支来提升目标检测的效果[10-11]。Gidaris 等[10]利用分割信息优化了目标定位。Zhang 等[11]在 SSD[12]上添加了语义分割信息，将目标边框信息引入卷积神经网络，证明了添加分割信息在 SSD 上的有效性。由于目标边框位置信息是弱分割信息，因而无需额外的标注，但是上述两种引入分割信息的方法都添加了额外的 loss 函数。Derakhshani 等[13]提出了一种更加简单高效的方法，将分割特征直接加入卷积神经网络结构中，用手动激活目标区域的方法替代构建新的网络分支，既将真实的目标位置信息引入卷积网络，又不增加额外计算量，并在 YOLOv2 和 YOLOv3 上证明了该方法的有效性。本文提出的方法借鉴了 Derakhshani 等[13]的方法，做了适当改进后将此方法运用到 Tiny YOLOv3 上。

3 本文方法

本文提出将目标边框信息加入卷积神经网络特征层上的策略来提高模型的检测精度，目标边框信息由检测数据集提供，无需额外的注释信息。在网络训练过程中，将目标真实的边框信息加入特征层中，并对这些目标区域做激活处理，流程如图 1 所示。从图 1 可以看出，激活操作是对特征图中的目标区域做了激活处理，并没有对网络结构进行改动。

激活过程的具体计算公式如下：

$$\alpha^{l+1}_{(c,i,j)}=\alpha^{l}_{(c,i,j)}+\alpha(t)e_{(c,i,j)} \tag{1}$$

其中，α^l 表示卷积神经网络第 l 层的特征层；c 为通道数；α 是控制激活水平的影响因子，由当前 epoch 数 t 决定；e 是激活张量。由式(1)可得，激活后的特征是上一层特征加上激活张量所得，激励的大小由影响因子决定。

$$g(i,j)=\begin{cases}1, & \text{if some bbox exists at } cell(i,j)\\ 0, & \text{if no bbox exists at } cell(i,j)\end{cases} \tag{2}$$

首先定义一个与输入特征长宽相同的掩膜 $g(i,j)$，根据有无目标区域对掩膜进行赋值。目标区域是需要激活的区域，像素点置为 1；对其余区域做抑制处理，像素点置 0。

$$e_{(c,i,j)}=\frac{g(i,j)}{d}\sum_{c=1}^{d}\alpha_{(c,i,j)} \tag{3}$$

激活张量是由输入特征图经通道压缩后与掩膜点乘所得，式(3)中，d 是特征的通道数。

$$\alpha(t)=\frac{0.5(1+\cos\pi t)}{\lambda\cdot Max\_Iteration} \tag{4}$$

激活因子的设定参照学习率的设计思路[14]，λ 是手动设计的一个参数，无需参与反向传播。$Max\_Itreation$ 是实验的最大迭代次数，在本实验中是 41 400。训练初期，激活因子最大，随着迭代次数的增多，激活因子逐渐减小至零。

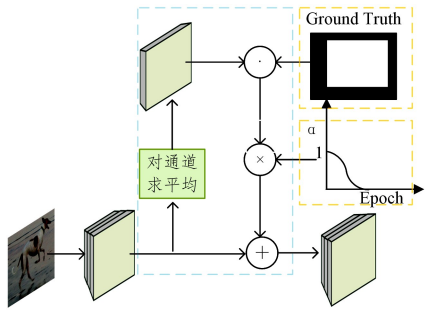

图 1 激活流程图

Fig. 1 Activation flowchart

4 实验分析

4.1 实验环境配置

本实验的硬件配置为 Intel(R)Core(TM) i7-8700，GPU 型号为 NVIDIA GeForce GTX 1050Ti，显存为 4 GB，内存 8 GB，操作系统为 ubuntu16.04。全局迭代次数为 40，每个 batch 训练 16 个样本，设置动量常数 β 为 0.9，权重衰减为 0.0005，初始学习率为 0.001。

本实验采用 YOLO 官方提供的 Tiny YOLOv3 预训练权重。在对特征层进行激活处理后，进行归一化处理，之后，模型的收敛速度更快，训练得到的结果曲线也更为平稳。

4.2 实验数据集

本文将 VOC2007 的验证集和 VOC2012 的验证集一起作为实验的训练集，VOC2007 的测试集用来测试[15]。训练集中共有 16551 张图片，测试集中有 4952 张，涵盖了 20 类物体。这 20 类物体中既包含飞机、火车等大物体，又有盆栽、鸟等小物体。

4.3 单层和多层激活实验

在选择激活的特征层上有多种方案：1)激活神经网络中的单层特征方案；2)激活多层特征方案。先对激活单层特征方案进行实验，实验结果如表 1 所列。总体来看，将激活方法加在网络模型中后部分的特征层上效果更佳，对于激活网络前面的特征层，得到的实验结果反而低于 baseline。在激活单层的实验中，激活原网络模型的第 21 层特征层结果最佳，此时 mAP 比 baseline 高 0.4%。

表 1 单层激活效果

Table 1 Single layer activation effect

Method(λ=2)	Stage	*mAP*/%
Tiny YOLOv3(512)	—	57.1
Tiny YOLOv3+(512)	Stage 6	42.1
Tiny YOLOv3+(512)	Stage 8	40.0
Tiny YOLOv3+(512)	Stage 12	49.5
Tiny YOLOv3+(512)	Stage 18	57.4
Tiny YOLOv3+(512)	Stage 19	57.1
Tiny YOLOv3+(512)	Stage 20	57.2
Tiny YOLOv3+(512)	Stage 21	**57.5**
Tiny YOLOv3+(512)	Stage 22	57.2

其次，对激活多层特征方案进行实验，实验结果如表 2 所列。多层激活有多种层数组合方式，本实验选取在单层激活实验中表现优异的特征层进行组合激活。从表 2 结果得出，激活多层特征的总体表现比激活单层特征更好。在 Case5 条件下，实验结果最为理想，比激活单层实验中最佳 *mAP* 高 0.3%。同时，由表 2 的推理速度比较得出：不论是单层还是多层激活，添加激活方法的模型的推理速度与原模型持平。

表 2 多层激活效果

Table 2 Multi-layer activation effect

Case	Stage(λ=2)	*mAP*/%	Inference Speed per image /ms
1	不加 AE	57.1	5.8
2	Stage 21	57.5	5.7
3	Stage 18+21	57.6	5.8
4	Stage 18+20+21	57.5	5.8
5	Stage 18+20+21+22	**57.8**	5.7
6	Stage 18+19+20+21+22	57.7	5.7

最后总结得出：Tiny YOLOv3 模型前面的特征层中包含了大量的浅层信息，缺少语义信息，中后部分富含深层语义信息，因而激活深层语义信息对目标检测起到了促进作用。

Tiny YOLOv3 改进前后的对比如表 3 所列。

表 3 Tiny YOLOv3 改进前后的对比

Table 3 Comparison of Tiny YOLOv3

Stage	Tiny Yolov3	Improved Tiny YOLOV3
1	Conv	Conv
2	Maxpool	Maxpool
3	Conv	Conv
4	Maxpool	Maxpool
5	Conv	Conv
6	Maxpool	Maxpool
7	Conv	Conv
8	Maxpool	Maxpool
9	Conv	Conv
10	Maxpool	Maxpool
11	Conv	Conv
12	Maxpool	Maxpool
13	Conv	Conv
14	Conv	Conv
15	Conv	Conv
16	Conv	Conv
17	Detection	Detection
18	Route	Route
19	Conv	Activation
20	Upsample	Conv
21	Route	Upsample
22	Conv	Activation
23	Conv	Route
24	Detection	Activation
25	—	Conv
26	—	Activation
27	—	Conv
28	—	Detection

4.4 λ 参数设置

探究完激活层数对结果的影响，接下来验证激活因子中手动设计的参数 λ 对实验结果的影响，结果如表 4 所列。由于实验变量取值范围太广，本实验选取了几个具有代表性的值进行测试。控制激活层数相同的前提下，预设 λ 的取值，分别进行实验。实验得出在训练 40 个 epoch，*max_Iteration* 为 41400 的情况下，λ=3 时得到的 *mAP* 最高，最好和最差取值的实验结果相差 0.6%。

表 4 λ 参数设置

Table 4 λ parameter setting

λ	*mAP*/%
1/2	57.6
1	57.4
2	57.8
3	**58.0**
4	57.5
5	57.4
10	57.6

4.5 学习策略的改进

实验预设的学习策略参照了课程学习[16]的思想：模拟人在学习过程中，首先需要大量的外部信息作为参考，然后随着学习的深入，慢慢降低对外部信息的需求，直至最后能自主学习。在激活层数相同、λ 相同的前提下，本文实验了 4 种训练策略，分别是：1)训练开始时 $\alpha=1$，随着训练的进行，α 沿着学习曲线下降，训练截止时，α 恰好降为 0；2)训练开始时 $\alpha=1$，训练尚未截止，α 降为 0 并保持不变；3)训练开始时 $\alpha=1$，训练进行到 2/3 处，α 降为 0 并保持不变；4)训练开始时 $\alpha=1$，训练进行到 1/2 处，α 降为 0 并保持不变。实验结果如表 5 所列。可以看出，策略 1)最佳。

表 5 不同学习策略对模型的影响

Table 5 Effects of different learning strategies

Strategy(λ=3)	*mAP*/%
1)	**58.0**
2)	57.3
3)	57.9
4)	57.2

4.6 通道压缩方法的改进

前文所述的实验中，激活张量是由输入特征图经通道压缩后与掩膜点乘所得。通道压缩的实现方法是直接对通道取平均，此时各个通道的特征层对融合后的单层特征贡献程度相同。本实验假设各个通道对融合后特征的贡献程度不同，用 1*1 卷积取代对通道求平均的方法对通道进行压缩。表 6 的实验结果表明，对通道加权融合进行通道压缩的方法比初始的直接对通道取平均的方法更好。如表 6 所列，改进了通道压缩方式后，*mAP* 从原来的 58%提升到 58.9%。

表 6 通道压缩方法改进

Table 6 Channel compression method improvement

通道压缩方法	*mAP*/%
通道求平均	58.0
1*1 卷积	**58.9**

4.7 结果对比

由上述实验对比可知：激活网络前半部分的浅层信息效果不佳，而激活中后部分的深层语义信息能辅助网络更好地训练。多层激活比单层激活能给网络精度带来更大的提升，且不论采用单层还是多层激活的方法，均不会对推理速度产生影响。在激活因子参数的选择上，选取 λ=3 比 λ=1 时的

mAP 高 0.6%。最后是对激活策略的研究，实验结果表明运用课程学习的思想[13]，从训练开始至结束，激励水平由最大值逐渐降为零的策略对网络精度提升最大。在通道压缩的方法选择上，对特征通道加权后再融合比直接对通道取平均的效果更好。实验得出的最优策略使 Tiny YOLOv3 的精确度从 57.1% 提升至 58.9%，推理速度为 175 帧/s，与原模型持平。

结束语 本文以 Tiny YOLOv3 算法为基础，提出了一种将目标位置信息引入深度卷积神经网络结构中的方法。与 Tiny YOLOv3 原网络相比，改进后网络的平均准确率更高，且检测速度与原网络保持一致，满足实时目标检测的需求。但改进后的 Tiny YOLOv3 算法还存在对一些小目标检测效果不好的问题，且仍有精度提升空间。下一步将引入改进小目标检测的方法来提升网络对小目标的检测性能。

参 考 文 献

[1] 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 & Machince Intelligence, 2017, 39(6): 1137-1149.

[2] CHEN C, QI F. The development of convolutional neuralnetwork and its application in the field of computer vision[J]. Computer Science, 2019, 46(3): 69-79.

[3] 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.

[4] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[5] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement[J]. arXiv: 1804.02767, 2018.

[6] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. arXiv: 2004.10934, 2020.

[7] LIU Z H. Structural model learning based on local feat-ures and its application in target detection and location[D]. Shanghai: Shanghai Jiaotong University, 2012.

[8] DAI J F, HE K M, SUN J. Instance-aware semantic segmentation via multi-task network cascades[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3150-3158.

[9] HARIHARAN B, ARBELAEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]// European Conferenc on Computer Vision. Springer, 2014: 297-312.

[10] GIDARIS S, KOMODAKIS N. Object detection via a multi-region and semantic segmentation-aware cnn model[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1134-1142.

[11] ZHANG Z S, QIAO S Y, XIE C H, et al. Single-shot object detection with enriched semantics[R]. Technical report, Center for Brains, Minds and Machines (CBMM), 2018.

[12] HE P, HUANG W L, HE T, et al. Single Shot Text Detector with Regional Attention[C]// 2017 IEEE International Conference on Computer Vision(ICCV). IEEE, 2017.

[13] DERAKHSHANI M M, MASOUDNIA S, SHAKER A H, et al. Assisted Excitation of Activations: A Learning Technique to Improve Object Detectors[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[14] HE T, ZHANG Z, ZHANG H, et al. Bag of Tricks for Image Classification with Convolutional Neural Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2019.

[15] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[16] BENGIO Y S, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 41-48.

YU Han-qing, born in 1996, postgraduate. Her main research interests include object detection and so on.

YANG Zhen, born in 1985, Ph.D. His main research interests include pattern recognition and intelligent systems, machine learning and image processing.