文章编号: 1671－9824(2020) 05－0128－05

基于 yolo 的上课状态检测方法

常思远

(许昌学院 信息化管理中心,河南 许昌 461000)

摘　要: 目标检测的一个重要的应用场景就是对于室内人员进行检测,包括人员的流动,人员的状态等等. 提出了一种基于 yolov 3 的教室人员上课状态的检测方法. 首先进行数据的标记,标记之后使用 vgg 进行初次预分类,预分类之后进行人工筛选,然后将数据交给 yolo 进行训练,最后再将该训练好的模型,放入视频进行实时的检测. 结果表明可以较好的检测出来学生的上课状态.

关键词: 目标检测; 上课状态; 深度学习; 实时检测
中图分类号: TP389. 1　　文献标识码: A

随着近年来深度学习的蓬勃发展,它影响到了人们的日常生活的很多方面. 例如刷脸支付,闯红绿灯检测等. 目标检测是深度学习的一个研究重要方向,目标检测是用来检测一张图像之中的某个既定目标,例如行人,车辆等. 目标检测的应用场景十分的广泛,除了人脸识别还有自动驾驶,人数统计,姿态识别等等. 但是随着应用范围的扩大,却很少有一些针对教育行业的应用. 基于目标检测和教育相结合的思路,尝试提出一种对于学生在课堂上课状态进行检测的方法.

1　目标检测方法简介

在深度学习发展起来之后,现阶段的目标检测大体分为两种: two-stage 和 one-stage. 其中 two-stage 系列的代表有 R-CNN[1] ,FAST R-CNN[2] ,FASTER R-CNN[3] . One-stage 系列的代表有 yolo[4] ,yolov2[5] ,yolov3[6] .

two-stage 的主要思路是以 R-CNN 为例如图 1 所示. 首先从原始图片中随机提取约 2 000 个候选区域(Region proposal) ,然后将每一个候选区域进行缩放到固定的大小(227* 227) 并使用 CNN 进行特征提取. 再将提取得到的向量使用一个支持向量机(SVM) 分类器进行打分,最后使用一个非极大值抑制来确定候选框的类别. 从上面可以看出 two-stage 的思路是非常简单粗暴的,这样就造成 two-stage 系列的检测方法在检测速度上是比较缓慢的. 后续的 FAST R-CNN 和 FASTER R-CNN 在检测效率上做了很大的提升,例如 FAST R-CNN 对输入的图片进行特征提取,ROI 池化; FASTER R-CNN 通过使用滑动窗口对图片的特征进行提取等等.

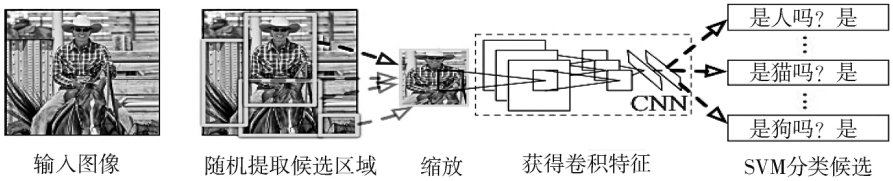

图 1　two-stage 检测过程

收稿日期: 2019－12－05
基金项目: 许昌学院校级科研项目(2020YB44)
作者简介: 常思远(1988—) ,男,河南许昌人,助教实验师,硕士,研究方向: 机器学习,深度学习.

one-stage 的主要思路则是将候选框这个因素拿掉,他的大体思路如图 2 所示(以 yolo 为例) . 首先将图片划分为 $N*N$ 的网络(一般为 7* 7) ,然后根据你的分类将图片输入 CNN 得到一个 $N*N*$ 分类数量的张量,对这个张量进行非极大值抑制. 从上面的说明可以直观看出 one-stage 比 two-stage 快很多. 当然它也有自己的问题,例如检测小目标效果不好等. 后续的 yolov2,yolov3 的主要改进是在检测的精度上面,yolov2 引入了锚框的概念对提高了检测的精度,而 yolov3 则是添加了特征提取器 FPN 来进一步提高精度.

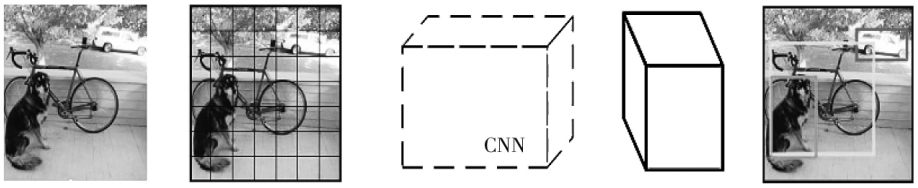

图 2　one-stage 检测过程

对于我们的教室内学生的上课状态检测这个问题[7] ,首先这是一个在固定摄影位置对学生的状态进行录像(目前大部分上课监控都是如此) ,其次,不会有太大的人员流动情况,最后对于学生的实时性有一定的要求. 因为是检测的目的就是要实时判断学生的上课状态,提供一些可以量化的数据. 从而可以用来帮助学校做一些决策或者提供一些用于分析的数据. 所以综合考虑以上这些因素,使用 one-stage 的检测方法效果会更好一些. 本文使用了 yolov3 的方法.

2　yolov3 原理

yolov3 是在 yolo 和 yolov2 的基础上通过吸收 faster r-cnn 和 RetinaNet[8] 的先进之处,引入了 FPN 架构(Feature Pyramid Networks) ,来实现多尺度的检测. 整体的架构如图 3 所示. 除了传统的 CNN 之外,还引入了 Residual[9] (深度残差网络) 层. Res 在 2015 年由何凯明等人提出,主要在网络中引入了恒等映射的概念,从而缓解了一直困扰着深度学习业界的由于深度增加带来的梯度消失问题,增加了特征信息传递的路径,从而把网络深度从几十层一跃提升到上百层. Res 极大的提高了系统的准确率. Res 的架构如图 4 所示. Res 的核心思想就是引入了一个快捷连接(short connect) ,这个快捷连接可以把经过卷积丢失的特征信息重新加入到卷积之后的数据里面,从而减少整个网络梯度消失的风险. 该方法对于教室内的上课状态监测具有一定的意义,因为目前的教室监控大多是固定单机位的,一般位于教室的最前面,监控录像拍到后排的学生必然是分辨率比较小的情况,这样使用 Res 正好可以检测这种小型的目标[10] .

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

图 3　yolov3 网络结构图

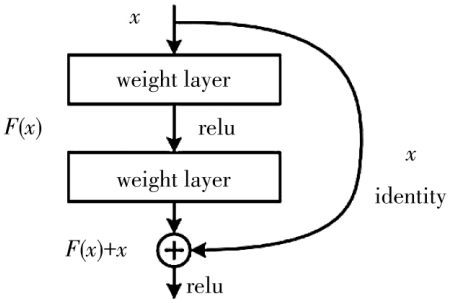

图 4　res 网络结构

FPN 的结构如图 5 所示，可以看出在使用了 416* 416 的输入大小时，图像被卷积到 52* 52，26* 26，13* 13 的时候会有三个先验框来进行预测，可以理解为 yolov3 采用了三个不同尺度的特征图来进行对象的检测，这样可以检测到更加细粒度的特征大小. 因为对于一个图片来说，分辨率越高，那么他对其他检测对象的表达就越是丰富，这样就加大了对于检测对象进行预测的难度，因为除了检测对象之外的东西对于 yolo 来说都是背景; 那么如果使用了不同尺度的特征图来检测，就可以减少背景对象的干扰，从而达到较好的检测效果. 正如上面所提到的一般而言教室监控视频中位于后排的学生必然分辨率较小，使用这种 FPN 结构就可以更好的酱这些小分辨率的对象检测出来.

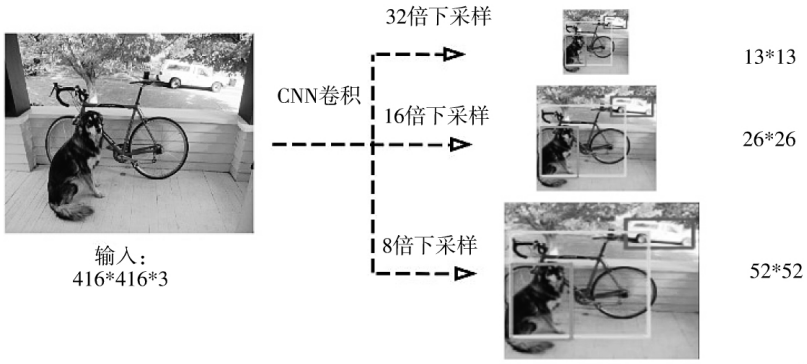

图 5 fpn 结构

3 实验及分析

3.1 实验数据准备

对于训练样本标注问题，因为是研究的学生上课的状态这一问题. 目前并没有相应的开放数据集可以供直接使用. 所以采用了手工标注的方法，前期收集了大概 100 张图片，进行手工标注，标注之后先使用一个简单的 CNN(VGG16) 网络进行训练. 训练之后再使用该网络去获取更多的训练样本. 获取之后再进行一次人工排错. 这样得到较好的训练样本数据，如图 6 所示.

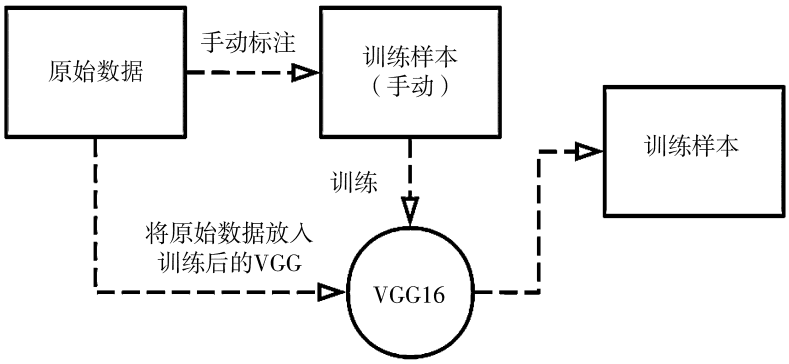

图 6 数据标注流程

通过标注得到了大约 800 张的标注数据，因为各种的原因人会有不同的倾斜，例如因为摄像头角度问题侧脸，或者光线不好造成的明暗不同等等. 所以需要进行一定的数据增强，来增加神经网络的鲁棒性. 所谓的数据增强就是人为的做一些例如角度变换，旋转，镜面，明暗度变换等等操作，之后再进行训练. 大体上将数据分为两个类别，一个是抬头，一个是非抬头. 其他的一些因素暂不考虑.

3.2 模型训练

模型训练主要如表 1 所示. 使用 yolo 默认的输入视频大小，基于动态学习率. 在进行了 2 000 次迭代之后达到了收敛. 因为收实验电脑的性能限制(显存 4 G) ，batch-size 只能设置为 8. 在经过大约 2 000 次的迭代后，达到收敛.

表 1　模型参数

参数	值
输入大小(input)	416* 416* 3
迭代次数(epochs)	2 000
Batch-size	8
学习率(lr)	0.000 1

3.3　结果分析

通过实验结果对比如图 7 所示. 在前排的学生都可以较好的检测出来,但是对于侧面以及后面的一些学生无法检测出来. 这个主要有几方面的原因: 1 这些学生在画面中的分辨率太小了,对于 CNN 来说将这么小的图像进行卷积压缩之后可能就无法提取有效的画面特征信息了 2 由于摄像头的角度所限制,拍摄出来的侧面学生的角度和在中间的学生角度是不一样的,那么 cnn 提取出来的特征就和之前手动标记的数据提取出来的特征是不一样的.

图 7　实验截图

3.4　实验环境及性能分析

使用的实验环境电脑配置为: i7-7 700,16 G 内存,2 t 硬盘,NVIDIA 1650 显卡(4 G 显存) . 软件环境为 python 3.7.3,tensorflow 1.13.1,keras 2.2.4.

性能分析见表 2. 在实时监测的时候,由于实验的电脑性能有限,所以只能达到 7 fps 的效果,但是横向对比其他的目标检测方法可以看出. 使用 yolo 是相对来说速度最快的方法了. Retinanet 因为实验电脑的显存不足,只能使用 CPU 版本进行对比分析.

表 2　性能分析

检测方法	内存占用值/ MB	显存占用/ MB	FPS
Yolov3	2 852	3 800	7
Faster R-cnn	3 521	3 850	4
Retinanet	6 351	无	1

4　结语

提出了一种基于 yolo 的学生状态实时监测的方法,从结果来看可以较好的检测出来学生的上课状态. 但是也存在一定的问题,比如一些坐在后排的学生没有被检测出来,这个和当前的摄像头的摆放及监控视频的分辨率有一定的关系,还有由于实验条件所限,无法与其他的常见目标检测算法进行对比. 下一步的工作可以把这些学生的上课状态信息进行统计分析,用来分析教师的上课质量或者学生的上课状态与学生的考试成绩之间的关系等等.

参考文献:

[1] 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. Columbus，OH，USA：IEEE，2014：580－587.

[2] GIRSHICK R. Fast R-CNN [C]. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago，Chile：IEEE，2015：1440－1448.

[3] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards real-time object detection with region proposal networks [C]. Advances in Neural Information Processing Systems. Montreal，Canada：NIPS2015，QC，2015：91－99.

[4] 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. Las Vegas，Nevada，USA：IEEE Computer Society Press，2016：779－788.

[5] REDMON J，FARHADI A. Yolo9000：better，faster，stronger [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii，USA：SIEEE Computer Society Press，2017：7263－7271.

[6] REDMON J，FARHADI A. YOLOv3：an incremental improvement [EB/OL]. (2018－04－08) [2019－12－2]. https：//arxiv. org/abs/1804. 02767.

[7] 邓柏聪. 智慧教室学生状态检测系统研究与设计 [J]. 信息技术与信息化，2019(1)：111－112.

[8] LIN T Y，GOYAL P. Focal Loss for Dense Object Detection [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017 (99)：2999－3007.

[9] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition [C]. CVPR 2016：2016 IEEE Conference on Computer Vision and Pattern Recognition，Las Vegas：IEEE，2016：770－778.

[10] 陈久红，张海玉. 基于深度学习的教室人数统计系统设计 [J]. 软件导刊，2019，18(10)：27－29.

A Detection Method for Class Status Based on Yolo

CHANG Siyuan

(*Information Management Center*，*Xuchang University*，*Xuchang* 461000，*China*)

Abstract: An important application scenario of target detection is to detect the status of the personnel，including the flow and status of personnel and so on. This paper proposes a detection method based on yolov3 to detect the state of personnel in class. The data is manually marked first. Then vgg is used for pre-classification. After the pre-classification，the data are trained by yolov 3. Finally，the trained model is sued for real-time detection by the video. The result shows that the class status of students can be well detected.

Keywords: target detection；class status；deep learning；real-time detection

责任编辑: 赵秋雨