文章编号：2095-2163(2022)01-0143-04 中图分类号：TP391.41;TP18 文献标志码：A

基于注意力机制的足球视频目标检测

亓 淼，郑凯东
（西安石油大学 计算机学院，西安 710065）

摘 要：足球运动在全世界范围内拥有广泛的受众和巨大的市场，利用计算机视觉技术对足球比赛视频进行目标检测，自动识别出球员、足球的位置，可以为进一步的跟踪提供良好的基础，对球队、对转播都有很大的帮助。本文提出了基于注意力机制的神经网络足球视频目标检测方案，通过搭建使用注意力机制的深度神经网络，并在足球相关数据集上进行训练，实现了对目标较为完整、准确的检测，为进一步的跟踪提供了帮助。

关键词：注意力机制；transformer；深度学习；目标检测

Soccer video object detection based on deep learning with attention mechanism

QI Miao, ZHENG Kaidong
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 Football has a wide audience and a huge market all over the world. Computer vision technology is used to detect the objects in football game videos and automatically identify the positions of players and football, which can provide a good foundation for further tracking task. It is of great help to broadcasting. This paper proposes a football video object detection scheme based on neural network using the attention mechanism: by building a deep neural network using the attention mechanism and training on football-related dataset, a more complete and accurate detection of the object is achieved and may help further tracking tasks.

【Key words】 attention mechanism; transformer; deep learning; object detection

0 引 言

在足球比赛视频中，通常存在球员在特定情况下分布密集，在剧烈运动中形变较大，以及足球本身目标较小等问题。因此，对目标的检测提出了挑战。Kamble[1]等人提出，先使用中值滤波进行背景消除，然后使用迁移学习的方式，离线训练 VGG[2]网络进行球员和足球检测，并设计了一种可以自动发现足球位置的算法，取消了人为干预过程，具有良好的扩展性。Şah[3]等人使用滑动窗口提取固定大小的图像块，经过手工方法滤波后，使用 CNN 进行检测，并且评估了卷积神经网络对不同图像表示类型（如 RGB、灰度图像、SIM、PSIM 等）在球员检测任务上的表现。Lu 等人[4]使用级联神经网络，通过分层的方式训练网络，使用网格搜索设置级联阈值，并使用膨胀策略提升整张图像上的性能，轻量化了网络并提高了精度，但背景相对单一，仅适用于运动场馆场景。文献[5]基于 SSD[6]网络进行球员检测，设计了一个反向连接模块，将高层次的语义信息传回底层，使低层次与高层次特征融合，提高了精度，特别是对小尺度目标的检测。文献[7]基于 HIS 颜色模型的视频图像，使用主颜色分割算法，对足球场地进行分割，并从形态学的角度提取场地区域，标记候选识别与跟踪的目标，使用 Hough 变换对场地线进行擦除，标记出球员和足球目标。文献[8]使用 SSD 网络对相机拍摄的球场图像进行目标检测，并基于 CN 颜色特征统计直方图进行球员队伍的分类。

注意力机制的 transformer[9]架构自从诞生以来，首先在 NLP 问题上取得了瞩目的成绩，之后人们尝试将其应用到计算机视觉领域。在目标检测方面，其表现甚至超过了常见的 CNN 检测器。本文通过搭建使用注意力机制的神经网络，使用迁移学习的方法，在足球相关数据集上进行训练，取得了不错的效果。

1 基于注意力机制的神经网络

深度学习方法一般将目标检测问题视为分类问题或者回归问题，或者两者皆有。例如 RCNN[10]，通过在输入图像上确定感兴趣区域，然后将其分类为背景或目标，最后使用回归模型生成目标的检测框。但是这种方法会受到图像近似拷贝（near-duplicate），即假正例问题的干扰。Carion[11]等第一

作者简介：亓 淼（1997-），男，硕士研究生，主要研究方向：计算机视觉、深度学习；郑凯东（1964-），男，硕士，副教授，主要研究方向：图形学与虚拟现实、智能计算与可视化、程序设计语言及应用开发。

收稿日期：2021-09-26

次将 transformer 应用到目标检测任务当中,构造了目标检测新框架 DEtection TRansformer(DETR)。DETR 是一个基于 transformer 的编码-解码架构,通过二分匹配(bipartite matching)进行独一预测的全局损失,进行固定大小的集合预测,从而可以一次性预测所有目标。其网络结构如图 1 所示:

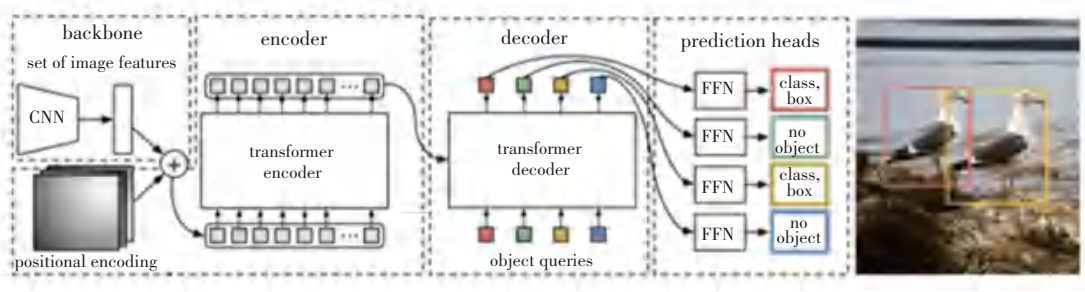

图 1 DETR 网络结构

Fig. 1 Network architecture of DETR

DETR 网络主要由 CNN 骨干网络、transformer 编码器-解码器、前馈网络 FFN 3 部分组成。首先,使用 CNN 对输入的图像进行特征提取,学习到图像的二维表示,被展开成为一维序列后,加上位置编码传递到编码器,与编码器的输出一起作为对象查询(object queries)输送到解码器。由解码器输出嵌入(output embedding)传递给前馈网络,被解码成归一化的边界框(bounding box)坐标和类别标签,完成预测。DETR 取消了卷积神经网络所使用的手动设计锚框生成和非极大值抑制等方法,实现了端到端的训练过程。

DETR 通过两步将检测转化成集合预测问题:第一步将预测框与实际框进行独一匹配的集合损失函数;第二步预测所有目标和目标间的关系。为了更好拟合预测目标与真实标签之间的匹配,需要计算一组参数 σ,使预测目标与真实标签的二分匹配损失最小化:

$$\hat{\sigma} = \underset{\sigma \in S_N}{\operatorname{argmin}} \sum_{i}^{N} L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \tag{1}$$

其中,y_i 表示真实数据,$\hat{y}_i$ 表示得到的预测结果。

损失函数的定义为:

$$L_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} [-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + 1_{\{c_i \neq \varnothing\}} L_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i))] \tag{2}$$

其中,L_{box} 定义如式(3),用来给标签框打分,预测其位置。

$$\lambda_{\text{iou}} L_{\text{iou}}(b_i, \hat{b}_{\sigma}(i)) + \lambda_{L1} \| b_i - \hat{b}_{\sigma}(i) \|_1 \tag{3}$$

2 实验验证

2.1 训练数据

本文实验中使用的数据集,一部分来自 soccerdb[12],可以用来进行目标检测、动作识别等多项视觉任务;此外,作者从网络上爬取了24 475张图像,并确保这些图像覆盖了各种足球活动中的场景。利用这些图像训练出一个检测器,加速标注过程。再利用这个检测器,标注一些从足球比赛视频中抽取的图像,选出预测效果最差的45 732帧图像,作为数据集。数据集一共有 70 万个带标签的边界框,具体类别数量见表 1:

表 1 soccerdb 数据集标签框分布

Tab. 1 Distribution of bounding boxes in the dataset

Bbox 类型	数量
Player	760 K
Ball	64 K
Goal	20 K

2.2 数据预处理与模型训练

本文实验环境为:操作系统为 Ubuntu 16.04,处理器为 Intel ® Core™i7-9750H CPU @ 2.60 GHz,显卡为 NVIDIA GeForce RTX 2080 Ti,显存为 11 GB。采用的深度学习框架为 pytorch。

首先,将数据集中的图像按照 6∶2∶2 的比例分为训练集、验证集和测试集。由于足球比赛转播视角有限,为了增加样本数据的多样性,提升神经网络模型的鲁棒性,将训练集分为两部分,一部分保留不做处理,另一部分使用在线数据增强的方法,包括缩放、翻转等等。

由于原数据集提供的是文本文件格式的标签边界框数据信息,所以需要手动转换成 COCO[13] 数据集所使用的 json 格式的 annotation 文件,以便后续训练。将数据集中每一张图像的 bounding box 信息由一个同名的文本文件存储,文件里每一行存放一个 bounding box 信息,每行共 5 列分别为:类别、中

心点横坐标、中心点纵坐标、标签框宽度与图像宽度的比例，以及标签框高度与图像高度的比例。遍历边界框文件，按行读取标签框信息，将其转化为以左上点的横、纵坐标和图像长宽表示的形式，写入 annotation 文件，用于训练神经网络。

本文使用迁移学习的思想，先加载预训练模型，然后利用处理好的数据集对模型进行微调。相比于随机初始化参数，这样做既保留了网络模型的能力，又减少了前向传播的计算量，从而避免了过长的训练时间。根据实验结果显示，当数据样本量小于一万时，微调比较有效。当数据量大于一万时，从头开始训练准确度会更高，但相应的训练时间也长。为了提升 GPU 的运算速度，本文实验将图像的最大宽度设置为 800 像素点，并将注意力头和骨干网络的学习率分别设置为 1e-5 和 1e-6，权重优化算法选择 AdamW[14]，权值衰减设置为 1e-4。Backbone 骨干网络选择 ResNet[15]，在 ImageNet 数据集上预训练好，加载权重参数，并将批归一化层的参数冻结，保持了提取图像目标信息的能力。图像通过 CNN 骨干网络的处理，得到低分辨率的激活特征图，将其展开成一维的激活特征图加入位置编码后，经过 transformer 编码-解码器，最后经过前向传播网络得到标签框的坐标以及类别标签。

2.3　实验结果与分析

网络训练完成后，得到的检测效果如图 2 所示。图 2 中的矩形框为训练好的模型对图像数据进行预测后的检测框，检测框的左上方标注了预测出的检测框的种类及置信度。可以看到，预测出的检测框与物体实际大小、位置基本一致，模型具有较为良好的检测效果。

图 2　足球比赛视频目标检测效果图

Fig. 2　An example prediction result picture

实验使用目标检测任务的评价指标 *AP* 和 *mAP*（mean average precision），对本文的实验效果进行评估。单类别精确度（*AP*）指标由真正例（*TP*）、真反例（*TN*）、假正例（*FP*）和假反例（*FN*）构成。其计算公式如下：

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

AP 的几何意义是指以召回率为横轴，精确度为纵轴所建立的坐标系中，*PR* 曲线与坐标轴包围成的面积。*AP* 值越高，代表对当前类别的检测效果越好。计算出 *AP* 后，对所有类别的 *AP* 求平均值，就得到整个数据集上的 *mAP*，可以从整体上反映模型的预测效果。本文的 *mAP* 由 3 类构成：球员（player）类别 *AP*、足球（ball）类别 *AP* 和球门（goal）类别 *AP*。

训练结束，将得到的 *mAP* 与原数据集上的 *baseline* 进行比较。表 2 所示的是针对不同尺寸的目标时，本文所使用的方法与基线的对比结果，表 3 展示的是针对不同类别的目标，本文所用方法与基线的对比结果。

表 2　不同尺寸的目标对比结果（*IOU*=[0.5：0.95]）

Tab. 2　Different bounding box scales

Method	Small	Medium	Large	All
DETR	29.6	63.4	75.8	65.1
BASELINE	30.2	63.1	75.4	64.8

表 3　不同类别的目标

Tab. 3　Different object classes

Method	mAP	Player	Ball	Goal
DETR	65.0	60.1	61.2	73.3
BASELINE	64.7	59.9	61.4	72.9

3　结束语

本文通过搭建使用注意力机制的深度神经网络，并且在大型足球数据集上进行训练。实验结果表明使用注意力机制的 DETR 网络提高了检测的准确率，特别是对大目标的检测性能。但另一方面，模型对于小目标的检测效果不是很理想，也存在网络收敛较慢的问题。在以后的工作中，将会改进网络模型，提高对足球这样的小目标的检测精度和加快网络收敛速度。

参考文献

[1] KAMBLE P R, KESKAR A G, BHURCHANDI K M. A deep learning ball tracking system in soccer videos [J]. Opto-Electronics Review, 2019, 27(1): 58-69.

（下转第 154 页）