文章编号:1007-1423(2020)17-0025-04 DOI:10.3969/j.issn.1007-1423.2020.17.005

# 基于深度强化学习的道路目标检测

吴志鹏,董超俊

(五邑大学智能制造学部,江门 529000)

**摘要:**

道路目标检测是高级驾驶员辅助系统(Advanced Driver Assistant System,ADAS)道路场景分析中的重要一环。提高目标检测的效率和速度,对图像目标的检测和识别具有重要的现实意义。深度强化学习是近年来兴起的一种多层结构的神经网络与具有决策能力的强化学习相结合的一种算法,通过端对端的学习方式来直接控制输入和输出。从提出此方法至今,深度强化学习已经有了实质性的突破。但是仍然有不足之处,因此提出一种改进的基于深度强化学习的检测算法,并将此算法用于道路目标检测。测试结果表明,该算法具有良好的目标检测性能。

**关键词:**

深度强化学习;道路目标检测;改进算法

## 0 引言

如今,随着社会经济的快速发展,汽车已成为几乎家家户户的便捷交通工具之一。这使得道路交通环境越来越复杂,人们期望有一个智能视觉辅助应用,为驾驶员提供交通标志信息,道路车辆信息,道路行人信息,以及协助车辆控制,来确保道路安全。道路目标检测与识别作为驾驶员辅助系统的重要功能之一,已经成为国内外研究人员的一个热点研究方向。它主要是利用车辆摄像头采集实时的道路图像,然后对道路上遇到的目标进行检测和识别,从而为驾驶系统提供准确的信息。

自卷积神经网络(Convolutional Neural Networks,CNN)被提出以来,目标检测的准确度有了较为明显的提高,其中比较经典的算法有 R-CNN、Faster R-CNN 等。R-CNN[1]是一种结合区域提名(Region Proposal)和卷积神经网络(CNN)的目标检测方法,采用的是选择性搜索(Selective Search),所以目标候选区的重叠使得 CNN 特征提取的计算中有着很大的冗余,在很大程度上限制了检测速度。而之后提出的 Faster R-CNN[2]抛弃了选择性搜索(Selective Search),引入了区域候选网络(Region Proposal Networks,RPN),使得区域提名、分类、回归一起共用卷积特征,从而加速了目标检测的速度。但是 Faster R-CNN 需要先进行目标判定,然后再进行目标识别。所以两种算法在检测速度和稳定性上仍然有提升的空间。

深度强化学习,顾名思义是将深度学习的感知能力和强化学习的决策能力相结合,目的是让两种算法的优势得到互补,输入如果是图像,深度强化学习也可以直接进行控制。近年来,深度强化学习的热度一直很高。其中 Mnih 等人[3]结合卷积神经网络(CNN)和 Q-learning 算法,提出一种深度 Q 网络模型(Deep Q-Network,DQN),并且在雅达利 2600 游戏中表现出色。由于 Q 学习存在过高估计的现象,Hasselt 等人[4]提出了深度双 Q 网络(Deep Double Q- Network,DDQN),证明了 DDQN 可以减小过高估计带来的误差。之后,Schaul 等人[5]在 DQN 中加入了优先级经验重放系统,可以更高效的使用样本。Hara 等人[6]提出了一种深度增强学习,用于检测视觉目标。本文通过调整折扣因子γ和学习率α,可以使 DQN 模型更加稳定,学习的质量也有所提升,从而提高目标检测的精准的。

## 1 相关工作

### 1.1 强化学习

强化学习(Reinforcement Learning,RL),是机器学习分支。与传统机器学习不同的是,强化学习是通过奖励值来训练模型,而机器学习是通过标签和数据特征来训练模型的。强化学习一般用于描述和解决智能体(agent)在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题[7]。强化学习主要包含以下几个个元素:环境的状态S、个体的动作A、环境的奖励R、个体的策略π、奖励衰减因子γ和状态转化模型P。强化学习中通常会引入马尔可夫决策过程(Markov Decision Process,MDP)。一般的,会将马尔科夫决策过程定义为一个四元组。其中:

(1)状态S,有限集合 $\{s^1,s^2,\cdots,s^N\}$,即 $|S|=N$。对于建模的问题来说,状态是所有信息中唯一的特征。

(2)动作A,有限集合 $\{a^1,a^2,\cdots,a^N\}$,即 $|A|=N$。能够用于某个状态 $s\in S$ 的集合表示为A(s),其中 $A(s)\subseteq A$。

(3)转换函数P,可以通过如下方式定义:$S\times A\times S\rightarrow[0,1]$,即它是从 $(S,A,S)$ 三元组映射到一个概率的函数,其概率表示为 $P(s,a,s')$,表示,从状态 $s$ 转换到状态 $s'$ 的概率,其值需要满足 $0\leqslant P(s,a,s')\leqslant 1$ 且 $\sum_{s\in s'}P(s,a,s')=1$,即概率必须满足实际,否则无意义。

(4)奖励函数R,可以定义为 $S\times A\rightarrow R$,在某状态执行某动作获得奖励。

马尔科夫决策过程与环境交互如图1所示。

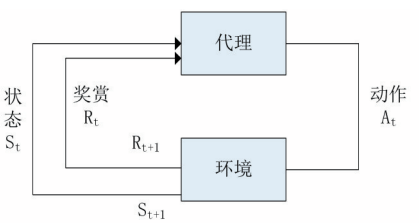


图1 马尔可夫决策过程

这里假设未来每个时间步获得的即时奖励都必须乘以一个折扣因子 $\gamma$,则从 $t$ 时刻开始到 $T$ 时刻情节结束时,奖励之和定义为:

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{1}$$

其中,$R_t$ 称之为回报或者累计奖赏,$\gamma\in(0,1]$ 称之为折扣因子。Agent的目标是通过最大化每个状态 $s_t$ 下的期望累未来回报的方式来选择操作。

状态-动作值函数:在状态下执行动作后获得的期望回报。

$$Q^{\pi}(s,a)=E\left[R_t|s_t=s,a_t=a,\pi\right] \tag{2}$$

对于所有的动作状态,假如一个策略 $\pi^*$ 的期望回报大于等于其他策略的期望回报,那么策略 $\pi^*$ 即为最优策略。

$$Q^*(s,a)=\max_{\pi}E[R_t|s_t=s,a_t=a,\pi] \tag{3}$$

公式(3)为最优状态动作值函数,即当处于状态s,执行了动作a,然后再按照π执行下去到最后,能获得的最大累计回报与期望。并且此值函数遵循贝尔曼最优方程(Bellman Optimality Equation)。即:

$$Q^*(s,a)=E_{s'\sim S}[r+\gamma\max_{a'}Q(s',a')|s,a] \tag{4}$$

强化学习算法的基本思想是通过使用贝尔曼方程作为迭代更新来估计动作值函数:

$$Q_{i+1}(s,a)=E_{s'\sim S}[r+\gamma\max_{a'}Q_i(s',a')|s,a] \tag{5}$$

当 $i\rightarrow\infty$ 时,$Q_i\rightarrow Q^*$。这种值迭代算法收敛于最优动作值函数。但是实际上,这种基本方法是完全不切实际的,因为每个序列的作用值函数是单独估计的,没有任何概括。相反,使用函数逼近器来估计动作值函数是常见的,即 $Q(s,a|\theta)\approx Q^*(s,a)$。

### 1.2 深度Q网络

深度Q网络(Deep Q-Network,DQN)是DeepMind团队提出来的深度强化学习算法,它是将卷积神经网络与强化学习中的Q-learning算法相结合,这里卷积神经网络的作用是对在高维且连续状态下的Q-Table做函数拟合,DQN相比于Q-learning有三大改进:①加入了卷积神经网络;②引入了目标网络(Target Network);③训练过程中应用了经验回放机制(Experience Replay)。图2表示了DQN的训练流程。

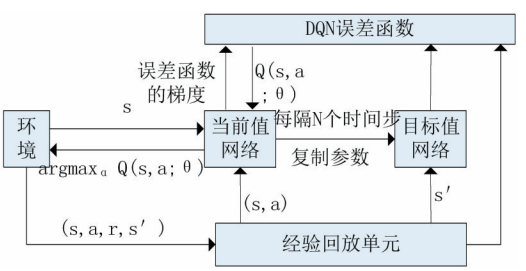


图2 DQN的训练流程

经验回放机制[8],把每个时间步中个体和环境交互

所得到的经验样本数据存储到经验池中，当模型在进行训练时，就会从经验池中随机抽取小批量的样本进行训练。引入经验回放机制后，不仅可以较为容易的对奖励数据进行备份，并且随机的从经验池中抽取小批量的数据也减小了样本之间的关联性，提高了系统的稳定性。其中，经验样本是以五元组(s,a,r,s',T)的形式进行存储的。具体表示为个体在状态 s 下执行动作 a，到达下一个状态 s'，就可以获得相应的奖励 r。其中 T 表示下一个状态 s'是否为终止状态。

在经典 Q-learning 算法中，目标 Q 值会随着预测 Q 值得增大而增大，这会是模型有震荡或者发散的可能性。所以 DQN 使用了两个神经网络模型：一个是用卷积神经网络来近似表示当前值函数，另一个神经网络则用来产生目标 Q 值。

目标函数为：

$$TargetQ = r + \gamma \max_a Q^*(s',a';\theta^-) \tag{6}$$

当前状态下估计值和目标值之间的误差计算公式(损失函数)：

$$L(\theta) = E_{(s,a,r,s')}[(r + \gamma \max_a Q^*(s',a';\theta^-) - Q(s,a;\theta))^2] \tag{7}$$

DQN 算法根据损失函数的公式来更新神经网络中的参数，通过引入目标函数，使得一段时间里目标 Q 值是不变的，在一定的程度上降低了两个 Q 值得相关性，使得训练时损失震荡甚至是发散的概率降低，提高了算法的稳定性。

### 1.3 自适应学习率

通过实验表明，高度复杂的任务，DQN 可以很好地训练，但存在过度拟合的风险。相反，复杂度较低的模型不会过度拟合，但可能无法捕获重要的特性。这时候，折扣因子γ在 DQN 的训练过程中起到了作用，当折扣因子γ在训练过程中越来越逼近其最终值，则可以加快魔性的收敛，从而降低了过拟合的现象，增加了系统的稳定性。

$$\gamma_{k+1} = 1 - 0.96(1 - \gamma_k) \tag{8}$$

随着折扣因子的增加，学习率随之降低，最终可以得到一个稳定的 DQN 训练模型。

$$\alpha_{k+1} = 0.96\alpha_k \tag{9}$$

## 2 实验与分析

### 2.1 数据准备

本文主要采用的是伯克利大学 AI 实验室(BAIR)发布的 bdd100k 数据集，数据集中的 GT 框标签共有 10 个类别，分别为：Bus、Light、Sign、Person、Bike、Truck、Motor、Car、Train、Rider。其中包含了 10 万段高清视频，每个视频大约约 40 秒，分辨率为 720p，帧数为 30fps。每个视频从第 10 秒对关键帧进行采样，从中获得了 10 万张图片，并进行标注。在 10 万张图片中，包含了不同天气、场景、时间的图片，包括晴天、阴天和雨天，以及白天和晚上的不同时间。并且数据集中都是真实的驾驶场景。

### 2.2 评价标准

由于 bdd100k 数据集中有多个图像标签，所以本文采用计算平均精度的方式来衡量目标检测模型的性能。下面是查准率(precision)和查全率(recall)的定义：

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

由此可以得到查准率-查全率曲线，简称“P-R 曲线”。由于 P-R 曲线不方便比较不同模型的性能，所以将 P-R 曲线换算为 mAP 值进行比较。

### 2.3 实验

将 bdd100k 数据集中的 100000 张图像导入深度 Q 网络模型，本文的实验采用了 Python 编程语言，是 Python 3.7。深度学习框架采用了 TensorFlow1.0.1。将样本图片分为 Bus、Light、Sign、Person、Bike、Truck、Motor、Car、Train、Rider 等 10 大类，不同的图片类别被用作 1,2,…,9,10 个标记。将数据库中 70000 张图像作为训练数据，30000 张图像作为测试数据。根据类别标识设置每组信号的期望输出值。实验结果如表 1。

表 1 不同方法下的相同训练域的 mAP 值比较

| 方法 | 训练域 | mAP |
|---|---|---|
| Faster-RCNN | clear | 36.6 |
| | daytime | 36.6 |
| | city | 42.0 |
| ours | clear | 37.1 |
| | daytime | 37.0 |
| | city | 43.2 |

由表 1 可以看出，本文在晴天(clear)、白天(daytime)、城市街道(city)三种不同的环境下，用两种不同的方法进行对比。实验结果表明，本文提出的带自适应学习率的深度 Q 网络在目标检测的精准度上，有一定的提升；并且在城市街道(city)的环境下表现的

最好。

## 3 结语

本文应用了带自适应学习率的深度 Q 网络，并将此方法建立模型用于道路目标的检测。实验证明了本文提出的方法优于以前的经典算法，确实提高了模型在复杂环境下对目标的检测性能。希望在将来，本文提出的方法能够得到更深层次的研究，并能够不断地优化对于不同对象的检测性能。

**参考文献：**

[1]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，2014:580-587.

[2]Ren Shaoqing，He Kaiming，Girshick Ross，et al. Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis&Machine Intelligence，2015，39(6):1137-1149.

[3]Mnih V，Kavukcuoglu K，Silver D，et al. Playing Atari with Deep Reinforcement Learning//Proceedings of the Workshops at the 26th Neural Information Processing Systems 2013. Lake Tahoe，USA，2013:201-220.

[4]Van Hasselt H，Guez A，Silver D. Deep Reinforcement Learning with Double Q-learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix，USA，2016:2094-2100.

[5]Schaul T，Quan J，Antonoglou I，Silver D. Prioritized Experience Replay//Proceedings of the 4th International Conference on Learning Representations. San Juan，Puerto Rico，2016:322-355.

[6]Hara K，LIU M Y，Tuzel O，et al. Attentional Network for Visual Object Detection[J]. arXiv Preprint arXiv:1702.01478，2017.

[7]Sutton R S，Barto A G. Reinforcement Learning:An Introduction. Cambridge，USA:MIT Press，1998.

[8]Lin L J. Reinforcement Learning for Robots Using Neural Networks. Defense Technical Information Center，USA:DTIC Technical Report: ADA261434，1993.

**作者简介：**

吴志鹏(1993-)，男，湖北荆门人，硕士，研究方向是目标检测

收稿日期：2020-02-26　　修稿日期：2020-04-20

# Road Target Detection Based on Deep Reinforcement Learning

WU Zhi-peng，DONG Chao-jun

(Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529000)

**Abstract:**

Road target detection is an important part of road scene analysis in Advanced Driver Assistant System (ADAS). To improve the efficiency and speed of target detection is of great practical significance to the detection and recognition of image targets. In recent years, deep reinforcement learning is a kind of algorithm which combines multi-layer neural network with reinforcement learning with decision-making ability. It directly controls input and output through end-to-end learning. So far, there has been a substantial breakthrough in deep reinforcement learning. However, there are still some shortcomings, so this paper proposes an improved detection algorithm based on deep reinforcement learning, and applies this algorithm to road target detection. The test results show that the algorithm has good performance in target detection.

**Keywords:**

Deep Reinforcement Learning; Road Target Detection; Improved Algorithm