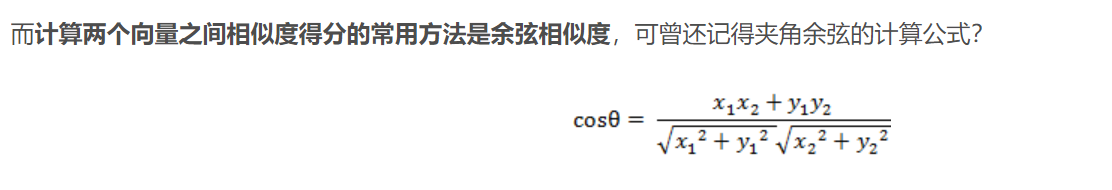
论文随笔



Word2vec 训练流程：不断缩小error（target(ground\_truth model)-sigmoid\_scores）

Embedding矩阵/ 上下文context矩阵

维度：词典大小

Embedding\_size

1.对于输入词，查看embedding矩阵

2.对于上下文单词，查看context矩阵

3.计算输入嵌入和每个上下文嵌入的点积，经sigmoid处理后，代表相似程度

4.loss = ground\_truth – sigmoid score 进行训练，

Seq2Seq 到Seq2Seq with Attention

**<DRL hands on>**

Discrete space:

Cross-Entropy method

DQN

Improved DQN

Policy gradient:

REINFORCE

Actor-critic

Asynchronous Advantage Actor-Critic(A3C)

Continuous space:

A2C

**<基于强化学习的路径规划技术综述>**

强化学习方法分类：

1. 基于值的强化学习方法
2. TD算法

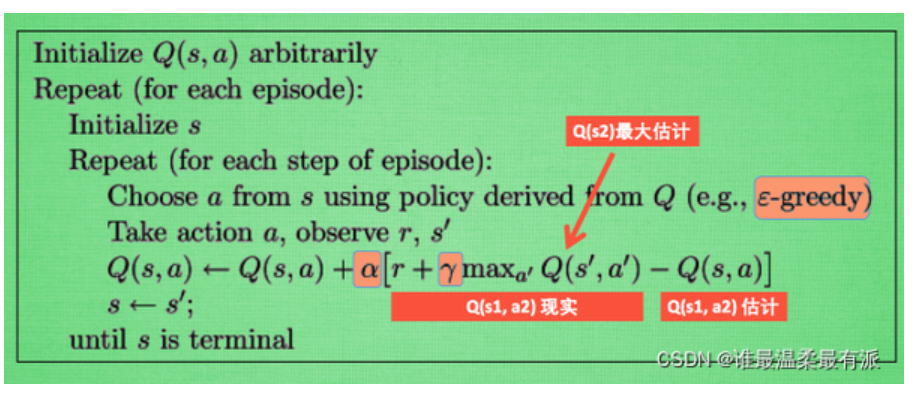


NAIR针对静态障碍物路径规划和避障的修正TD算法，降低了其复杂度

MARTIN 将TD的更新过程简化为高斯回归过程，提高效率

2）Q learning

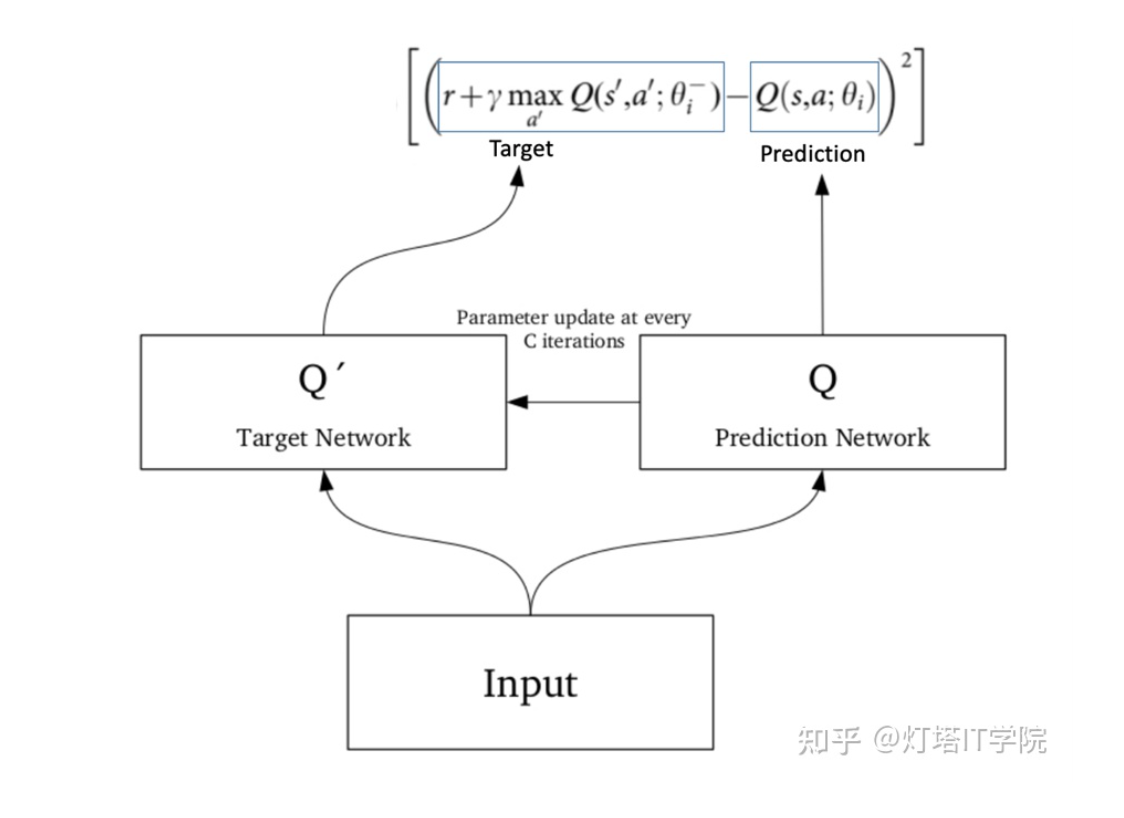
伪代码：



DQN: main idea : approximation

1.replay buffer : solve iid problem

2.target network:

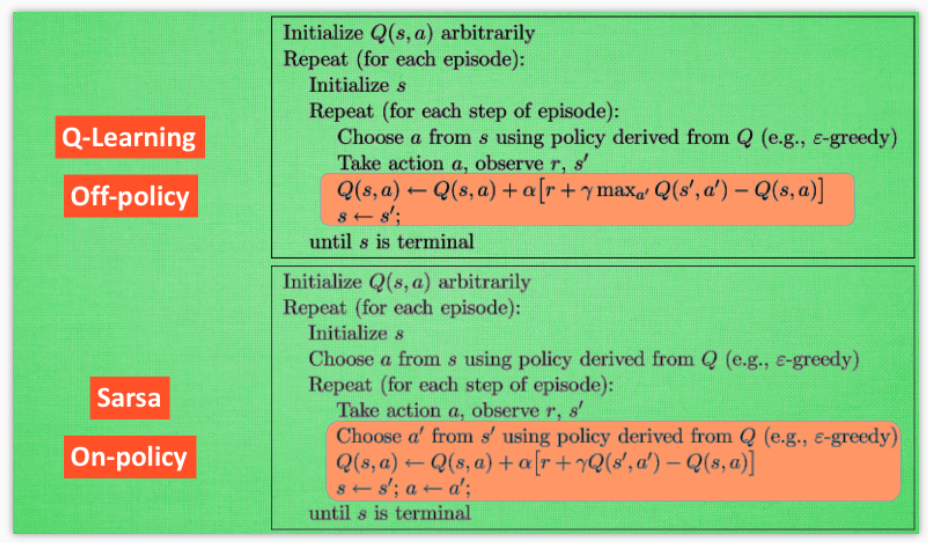


使用MSE规划得到loss, 对prediction进行反向传播，更新prediction参数，n batches之后target <- prediction network

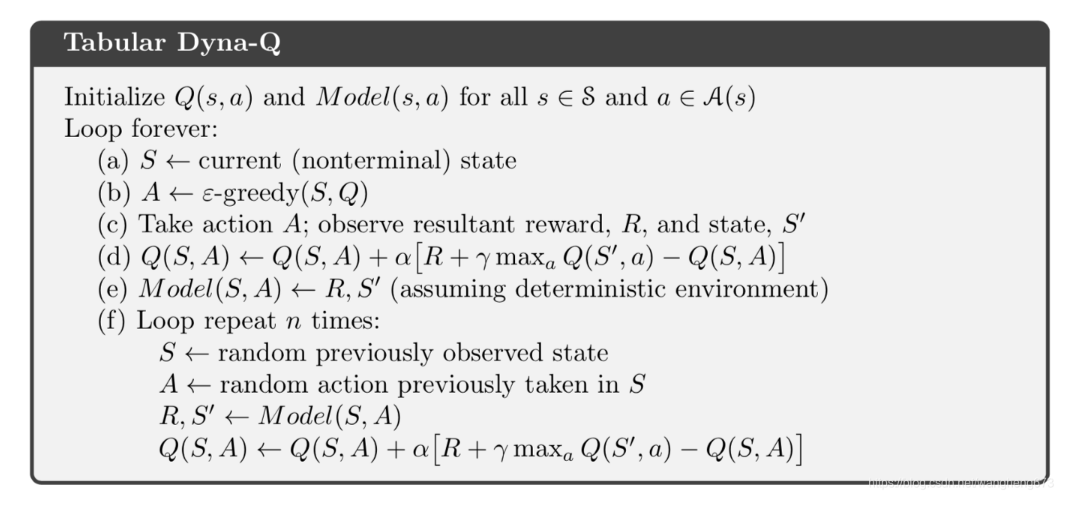
Double DQN:

经常会发现DQN网络会出现高估的情况，高估来自于对action value的估计不准确，有的时候这种估计是好的，有的时候是坏的

* 1. SARSA



* 1. dyna-q learning



1. 基于策略的强化学习方法
2. policy gradient

deterministic policy gradient 确定性策略梯度法 argmax，性能在连续动作空间中更加优秀

stochastic policy gradient 随机性策略梯度法 动作以某概率被执行

Monte-Carlo method:

如果要预测，更行的还是Q function

2) imitation learning

跟强化学习相结合

3)值与策略相结合的强化学习方法

Actor-critic

1. 局限性
2. 值方法：1. 难以应用于连续空间

2. 得到的是一个确定性策略，无法得到随即策略

3.微小扰动可能会导致极大结果的变化

2）policy gradient：

1.需要完全序列样本，训练慢，方差高

2.容易收敛到局部最优解

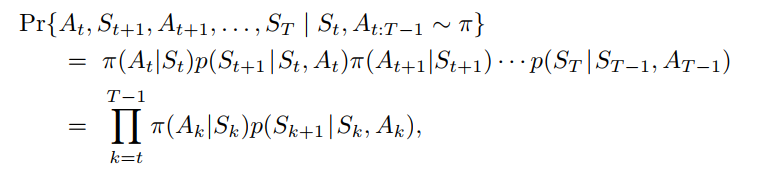
3.移动机器人通常在离散的动作空间中

前瞻：

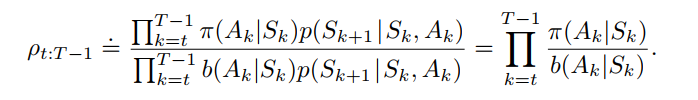
1. 设计有效的奖励函数
2. 解决强化学习的探索-利用困境
3. 研究强化学习方法和常规方法的结合方法
4. 将强化学习算法应用于多智能体协作的路径规划研究中

**Off policy prediction via importance sampling:**

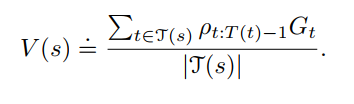
The probability of the subsequent state-action :



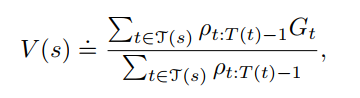
Importance-sampling ratio:



Ordinary importance sampling:



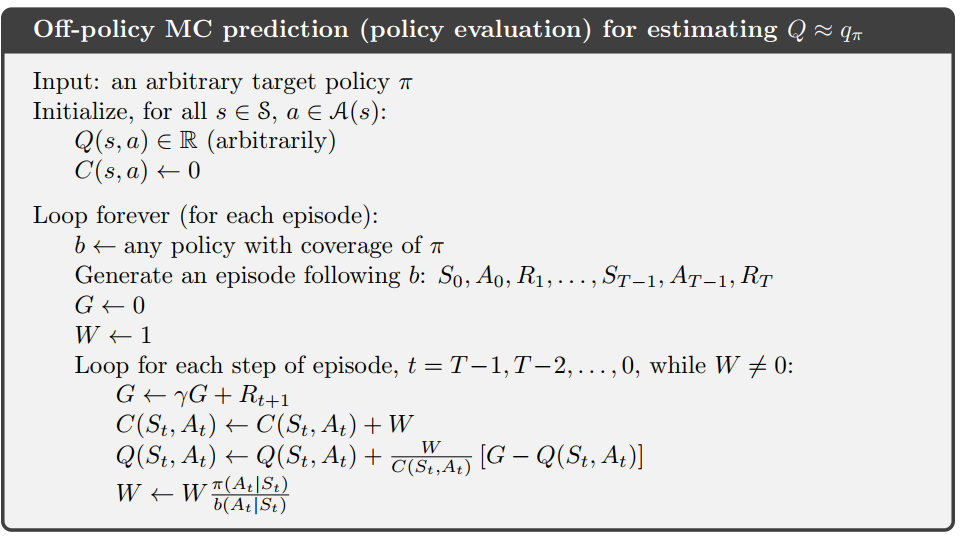
Weighted average sampling:

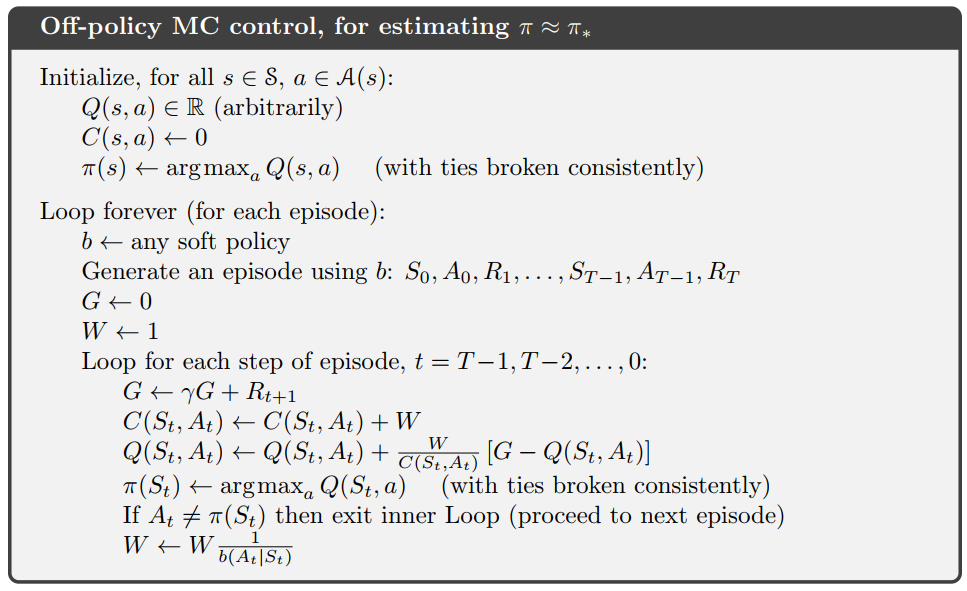


Ordinary importance sampling is unbiased, and weighted is biased . it is more likely an evaluation. But as the samples increases, it converges to zero.

But weighted estimator usually has dramatically lower variance and is strongly preferred.

Combined the off-policy mc / weighted importance sampling / incremental implementation

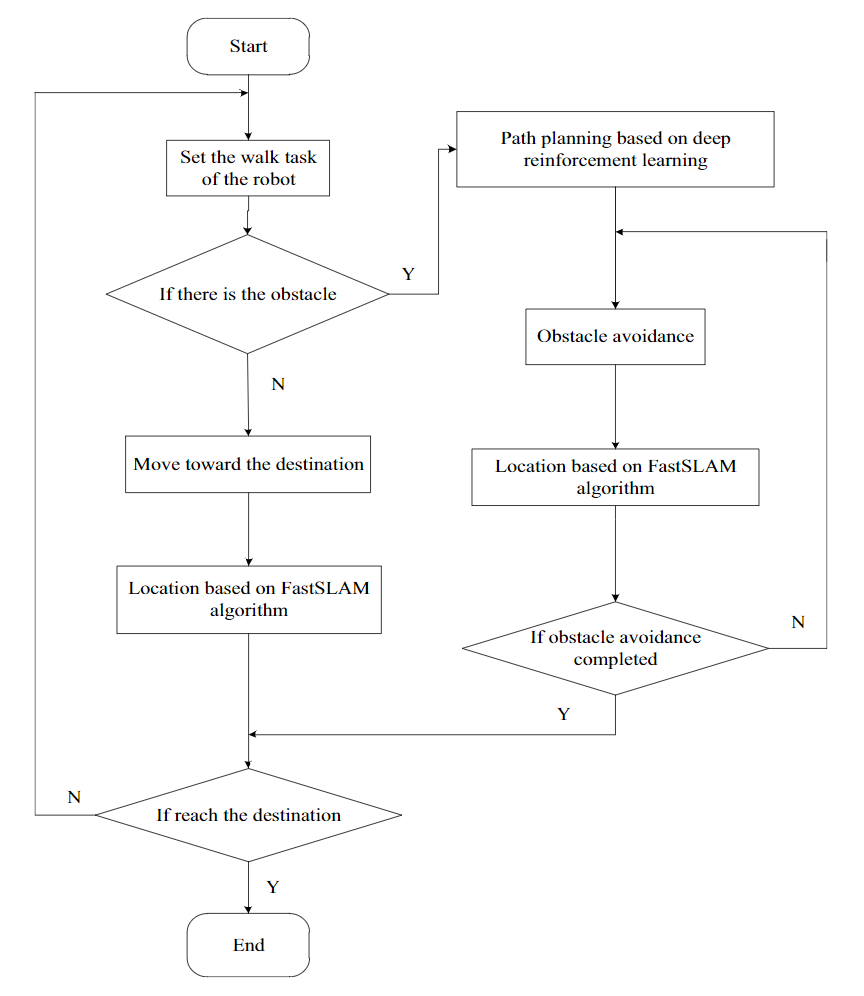




<Path planning for active SLAM based on deep reinforcement learning under unknown environments>

in this paper, fully convolutional residual network (FCRN)is used to recognize the obstacles to get depth image and build the actual environment 2D maps. The avoidance path of the obstacle is planned by Dueling DQN algorithm in the robot’s navigation; at the same time, the 2D map of the environment is built based on FastSLAM.

Action space: linear velocity is set to 0.4 or 0.2 . Angular velocity is pi/6,…(5 options) . The instantaneous reward function is r = v\*cos(w) \*dt and plus a somehow bonus function to make the robot run as fast as possible.



Deterministic Policy Gradient Algorithm

此外，我们表明确定性策略梯度是随机策略梯度的极限情况，因为策略方差趋于零。

the policy gradient integrates over both state and action spaces, whereas in the deterministic case it only integrates over the state space

we introduce an off-policy learning algorithm to explore.

Thebasic idea is to choose actions according to a stochasticbehaviour policy (to ensure adequate exploration), but tolearn about a deterministic target policy

<Target-driven Visual Navigation in indoor scenes using deep reinforcement learning>

Needs to address two issues, 1.lack of generalization capability to new goals. Needs to train new models for every new goals 2.data inefficiency. Models requires several episodes of trial and error to converge. Methods:1.

Mapless / do not need 3d reconstruction/do not require supervised training for landmarks

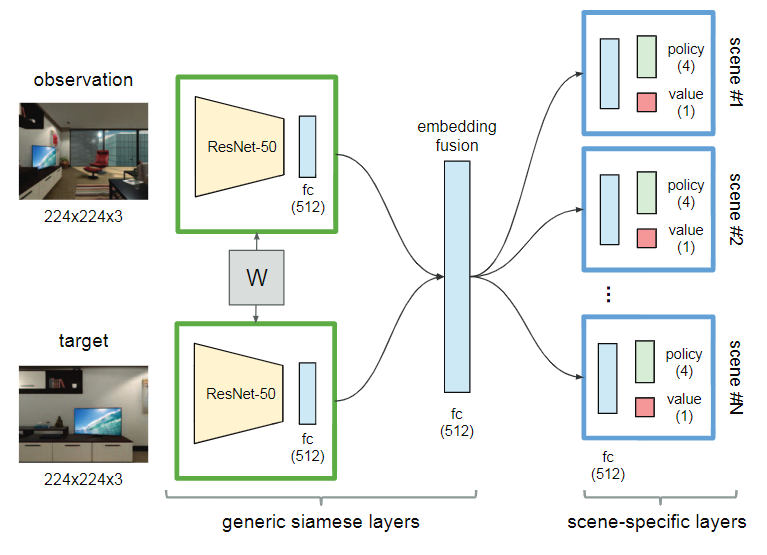
a representation of current statestand a representation oftargetgand produces a probability distribution over theaction spaceπ(st,g).

action space: constant step length and turning angle

observations and goals: current RGB images and target RGB images

reward design: goal-reaching reward 10 and small time penalty -0.01

model:



Navigational decisions demand an understanding of the relative spatial positions between the current locations and the target locations, as well as a holistic sense of scene layout. Our approach to reasoning about the spatial arrangement between the current location and the target is to project them into the same embedding space, Information from both embeddings is fused to form a joint representation. The intention to have scene-specific layers is to capture the special characteristics

<DDPG>

DDPG is a method that is a actor-critic, model-free algorithm based on the deterministic policy gradient that can operate over continuous action space.（have a good performance in high dimensional space）

Use replay buffer to address independently and identically distributed issue.

To solve converge issue, using target network and current network. Target values are constrained to change slowly θ′←τθ+ (1−τ)θ′ with τ<<1.

To manually scale the features in similar ranges across environments and units, the technique normalizes each dimension across the samples in minibatch to have unit mean and variance.

Construct the actor policy with a noise sampled from a noise process N : μ′(st) =μ(st|θμt) +N

