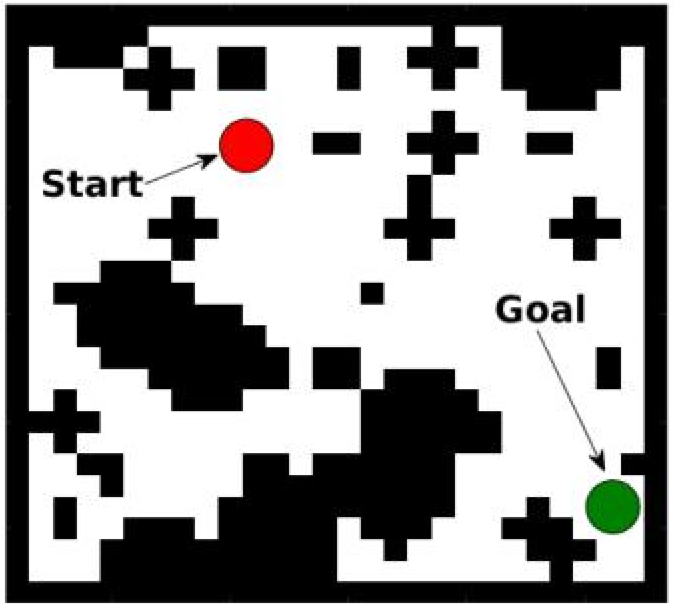
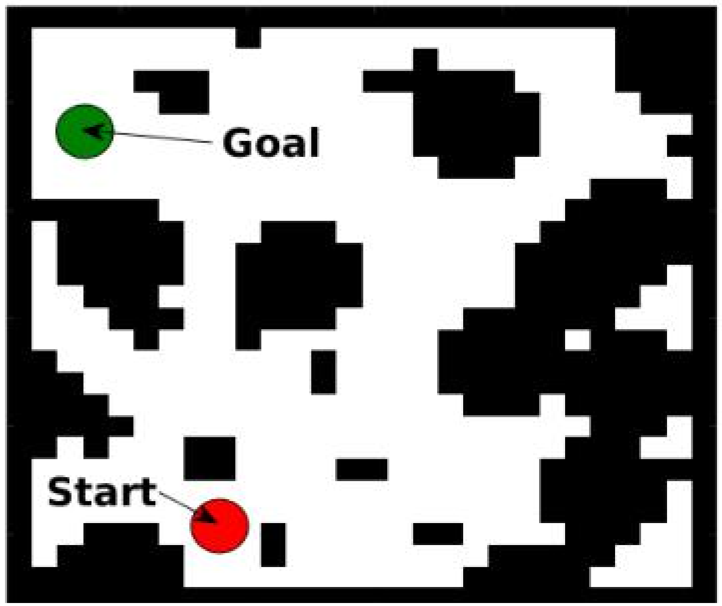
Value Iteration Networks

presented by Jason TOKO

背景与动机

- RL要解决的是序列决策问题，一般需要一定的planning
- 纯learning方法学习的策略泛化能力差，例如：



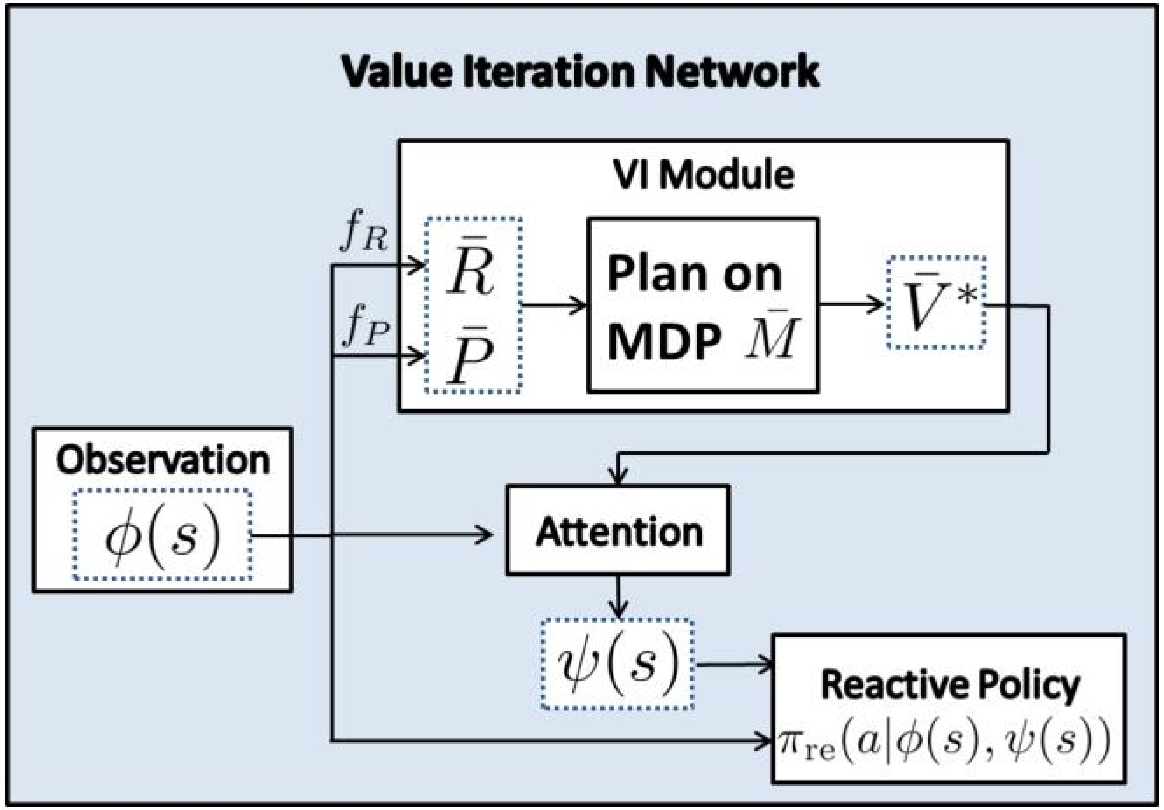


背景与动机

- VIN涉及到的背景知识
  - Value Iteration
  - CNN
  - RL与IL
- VIN的算法思想——learn to plan：
  - 构造一个可微的planning模块，作为VI的逼近
  - end-to-end的训练，契合RL或IL算法
  - 训练后，可根据observation得到相关的planning computation，然后再根据planning得到预测的动作

价值迭代网络

- 整体架构



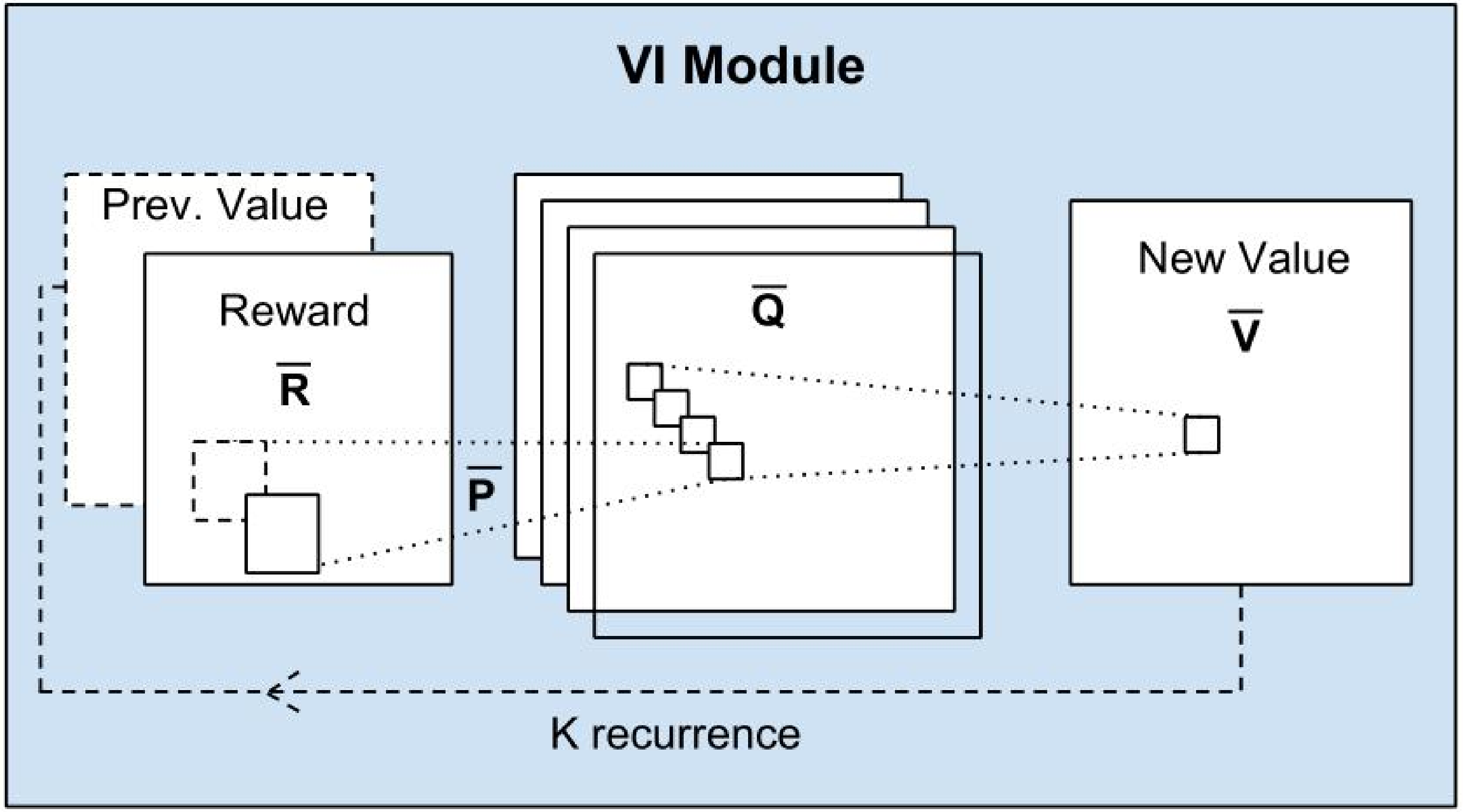
Value Iteration Network
VI Module
f_R
$\bar{R}$
f_P
$\bar{P}$
Plan on MDP $\bar{M}$
$\bar{V}^*$
Observation
$\phi(s)$
Attention
$\psi(s)$
Reactive Policy
$\pi_{\mathrm{re}}(a|\phi(s), \psi(s))$

价值迭代网络

- 定义：M与$\bar{M}$, $\bar{S}$, $\bar{A}$, $\bar{R}(\bar{s},\bar{a})$, $\bar{P}(\bar{s}'|\bar{s},\bar{a})$
- M与$\bar{M}$的联系：$\bar{R} = f_R(\phi(s)), \bar{P} = f_P(\phi(s))$
- VI模块：输入f_R、f_P，输出$\bar{V}^*$
- attention：$\psi(s)$
  - 一个状态的最优策略只与一部分状态有关
  $\bar{\pi}^*(\bar{s}) = \arg\max_{\bar{a}} \bar{R}(\bar{s},\bar{a}) + \gamma \sum_{\bar{s}'} \bar{P}(\bar{s}'|\bar{s},\bar{a})\bar{V}^*(\bar{s}')$
  - 通过减少学习过程中的有效网络参数，可提高学习效果

价值迭代网络

- VI模块



VI Module
Prev. Value
Reward
$\overline{\mathbf{R}}$
$\overline{\mathbf{P}}$
$\overline{\mathbf{Q}}$
New Value
$\overline{\mathbf{V}}$
K recurrence

价值迭代网络

- VI模块：利用CNN实现VI算法迭代过程
- 卷积层：
  - 输入奖励图：$\bar{R}$，维度l, m, n
  - 转移概率卷积核：$\bar{P}$
  - 输出Q值图：$\bar{Q}_{\bar{a},i',j'} = \sum_{l,i,j} W_{l,i,j}^{\bar{a}} \bar{R}_{l,i'-i,j'-j}$
- 池化层：
  - 沿着channel最大池化：$\bar{V}_{i,j} = \max_{\bar{a}} \bar{Q}_{\bar{a},i,j}$
- $\bar{V}$与$\bar{R}$ 堆叠，作为卷积层的输入，反复迭代K次

价值迭代网络

VI模块和VI算法对比：

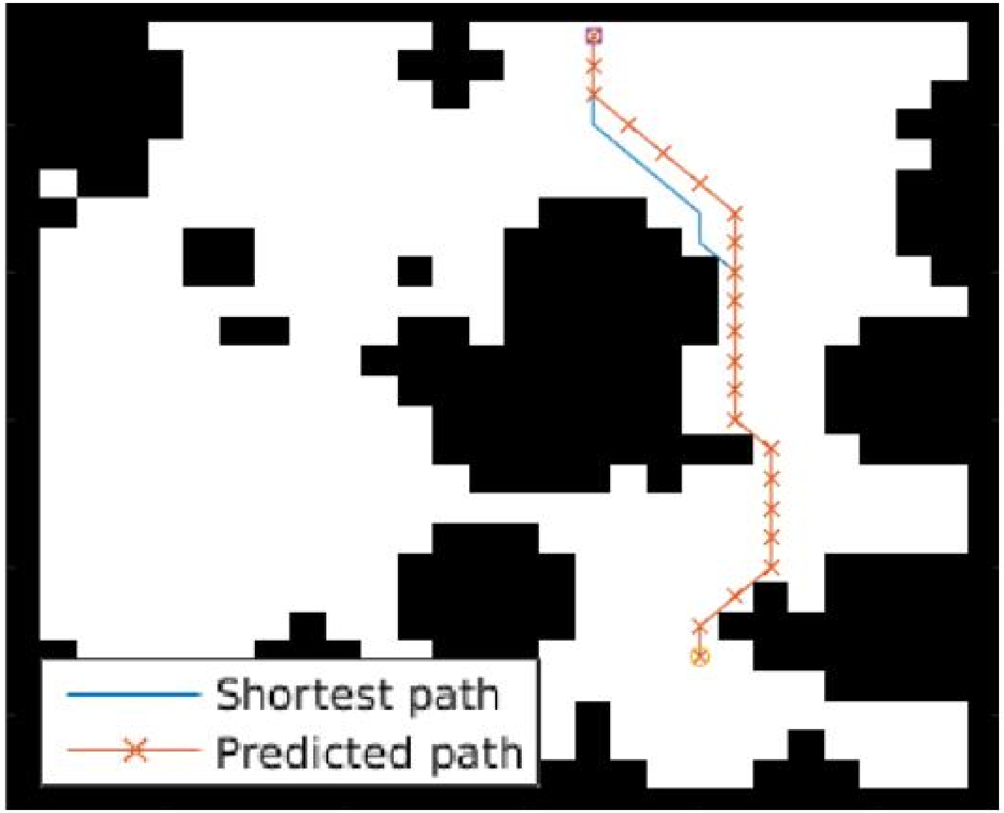
VI模块		VI算法
$\bar{Q}_{\bar{a},i',j'} = \sum_{l,i,j} W^{\bar{a}}_{l,i,j} \bar{R}_{l,i'-i,j'-j}$	$\longleftrightarrow$	$Q_n(s,a) = R(s,a) + \gamma \sum_{s'} P(s'\|s,a) V_n(s')$
$\bar{V}_{i,j} = \max_{\bar{a}} \bar{Q}_{\bar{a},i,j}$	$\longleftrightarrow$	$V_{n+1}(s) = \max_a Q_n(s,a)$
输出:$\bar{V}^*$	$\longleftrightarrow$	输出:$\pi^*(s) = \arg\max_a Q_\infty(s,a)$

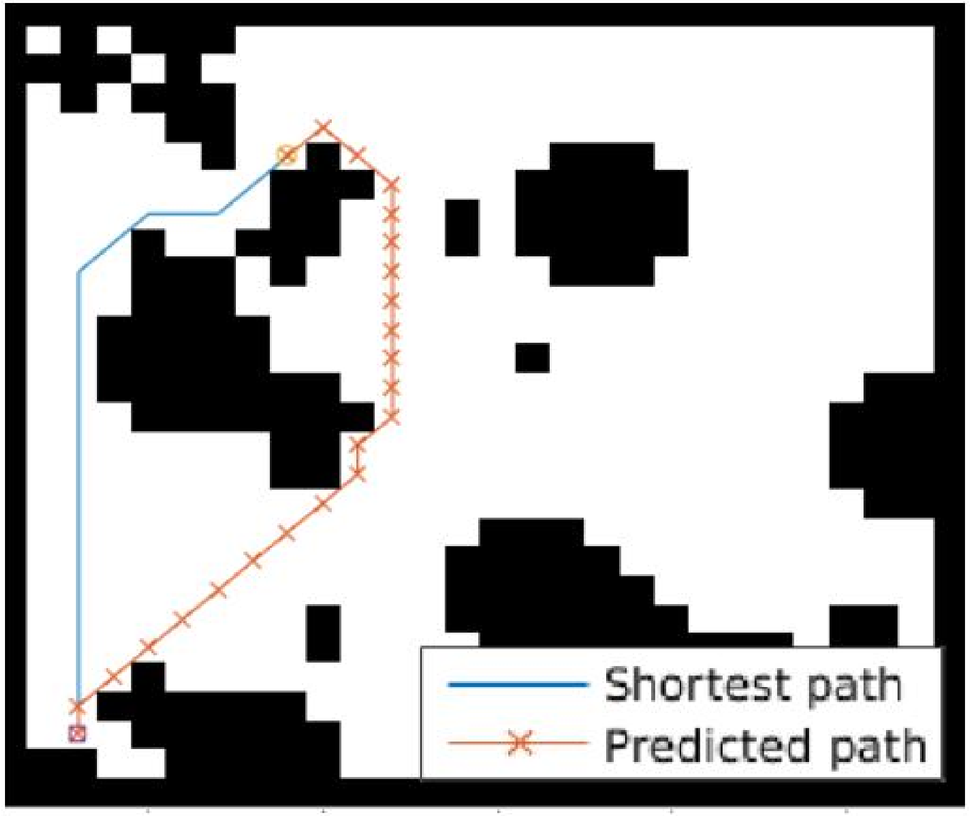
实验

- VIN源代码：https://github.com/avivt/VIN
- Grid-World Domain
- Mars Rover Navigation
- Continuous Control
- WebNav Challenge

实验

- Grid-World Domain：随机起始点、目标点、障碍物





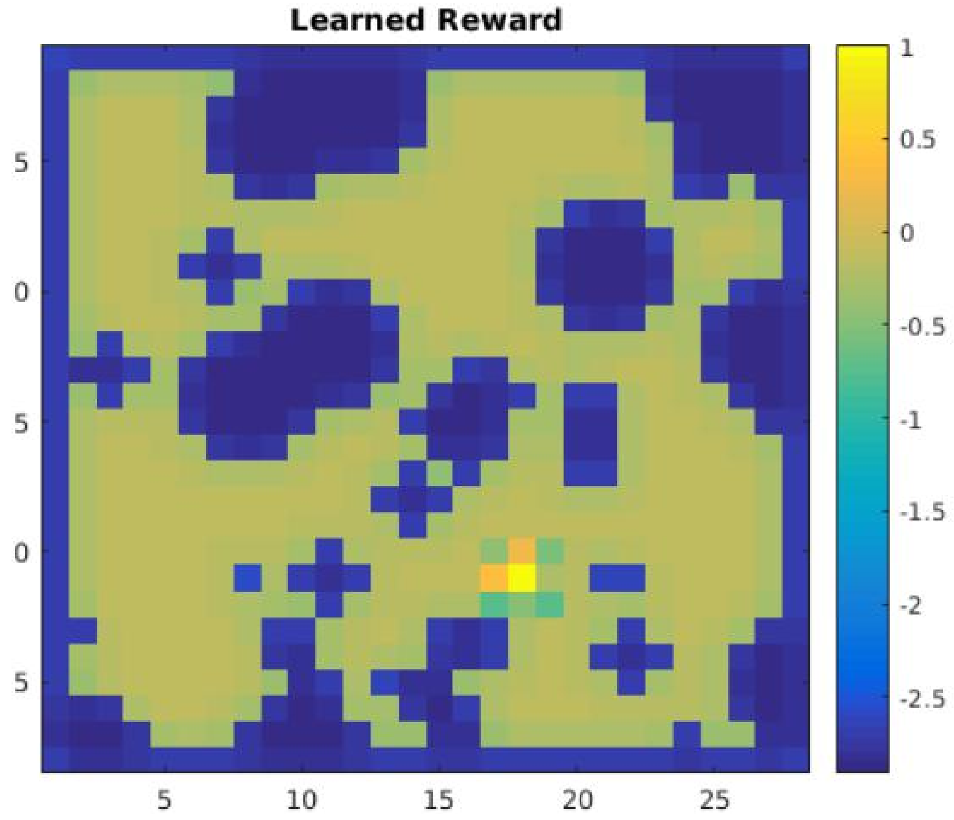
实验

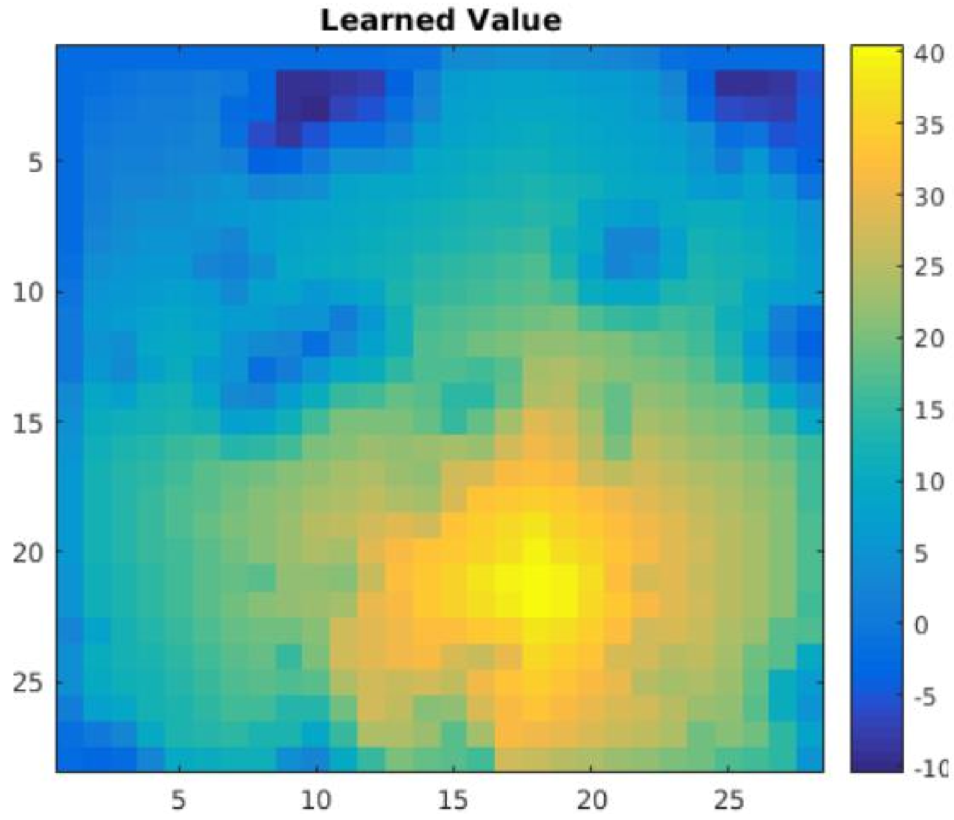
- 评估指标：prediction loss、success rate、trajectory difference
- VIN与CNN、FCN对比

Domain	VIN			CNN			FCN		
	Prediction loss	Success rate	Traj. diff.	Pred. loss	Succ. rate	Traj. diff.	Pred. loss	Succ. rate	Traj. diff.
8×8	0.004	**99.6%**	0.001	0.02	97.9%	0.006	0.01	97.3%	0.004
16×16	0.05	**99.3%**	0.089	0.10	87.6%	0.06	0.07	88.3%	0.05
28×28	0.11	**97%**	0.086	0.13	74.2%	0.078	0.09	76.6%	0.08

实验

- VIN可视化f_R与$\bar{V}^*$





附录E 分层VI模块

- 问题：VI迭代次数K取决于问题的规模，若K过小，会导致奖励信息无法传递到所有状态上。
- 解决方案：应用分层VI模块（Hierarchical VI Modules）来加速奖励信息的传递。

附录E 分层VI模块

- HVIN