

VALUE ITERATION NETWORKS

Aviv Tamar

Joint work with Pieter Abbeel, Sergey Levine, Garrett Thomas, Yi Wu

June 23, 2016

UC Berkeley

Berkeley

Artificial Intelligence Research Laboratory

INTRODUCTION

MOTIVATION

- Goal: autonomous robots

Robot, bring me the milk bottle!



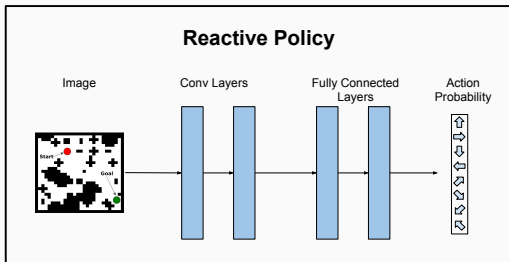
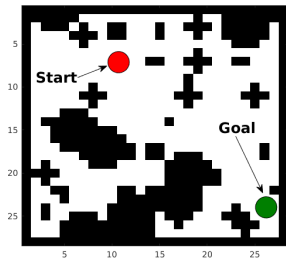
- Solution: RL?

- Deep RL learns policies from high-dimensional visual input^{1,2}
- Learns to act, but does it **understand**?
- A simple test: generalization on grid worlds

¹Mnih et al. Nature 2015

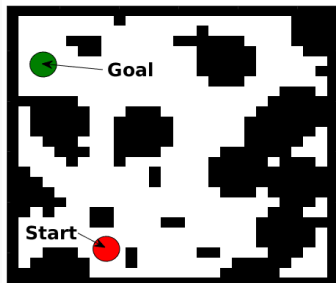
²Levine et al. JMLR 2016

INTRODUCTION



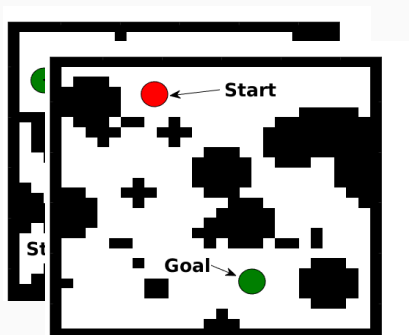
INTRODUCTION

Train



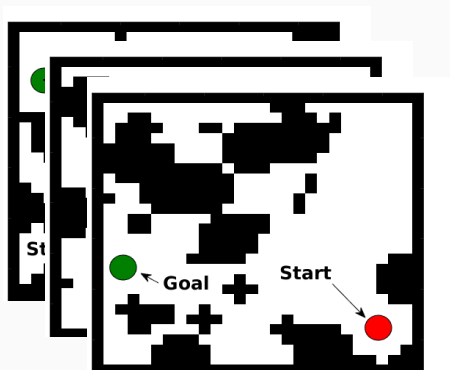
INTRODUCTION

Train



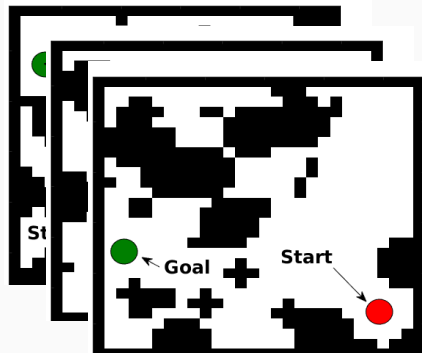
INTRODUCTION

Train

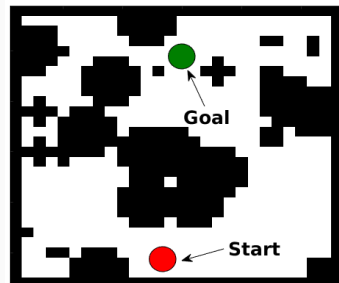


INTRODUCTION

Train



Test



Observation: reactive policies do not generalize well

Why don't reactive policies generalize?

- A sequential task requires a **planning computation**
- RL gets around that – learns a mapping
 - State \rightarrow Q-value
 - State \rightarrow action with high return
 - State \rightarrow action with high advantage
 - State \rightarrow expert action
 - [State] \rightarrow [planning-based term]
- Q/return/advantage: planning **on training domains**
- New task – need to **re-plan**

In this work:

- Learn to plan
- Policies that generalize to unseen tasks

BACKGROUND

Planning in MDPs

- States $s \in \mathcal{S}$, actions $a \in \mathcal{A}$
- Reward $R(s, a)$
- Transitions $P(s'|s, a)$
- Policy $\pi(a|s)$
- Value function $V^\pi(s) \doteq \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right]$
- Value iteration (VI)

$$V_{n+1}(s) = \max_a Q_n(s, a) \quad \forall s,$$

$$Q_n(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) V_n(s').$$

- Converges to $V^* = \max_{\pi} V^\pi$
- Optimal policy $\pi^*(a|s) = \arg \max_a Q^*(s, a)$

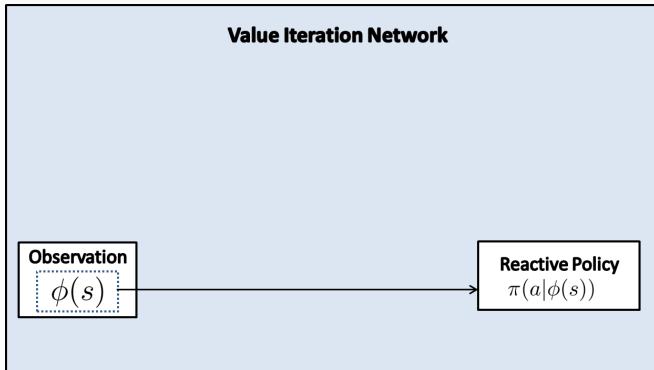
Policies in RL / imitation learning

- State observation $\phi(s)$
- Policy: $\pi_{\theta}(a|\phi(s))$
 - Neural network
 - Greedy w.r.t. Q (DQN)
- Algorithms perform SGD, require $\nabla_{\theta}\pi_{\theta}(a|\phi(s))$
- Only loss function varies
 - Q-learning (DQN)
 - Trust region policy optimization (TRPO)
 - Guided policy search (GPS)
 - Imitation Learning (supervised learning, DAgger)
- Focus on policy representation
- Applies to model-free RL / imitation learning

A MODEL FOR POLICIES THAT PLAN

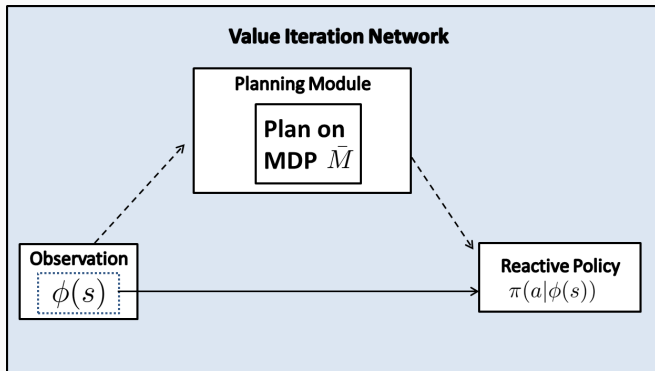
A PLANNING-BASED POLICY MODEL

- Start from a reactive policy



A PLANNING-BASED POLICY MODEL

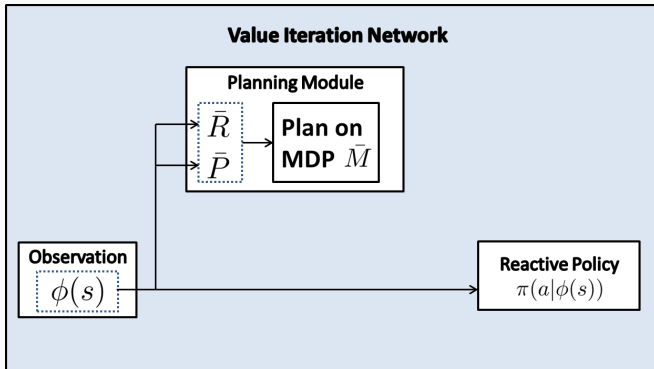
- Add an explicit planning computation
- Map observation to planning MDP \bar{M}



- Assumption: observation can be mapped to a useful (but **unknown**) planning computation

A PLANNING-BASED POLICY MODEL

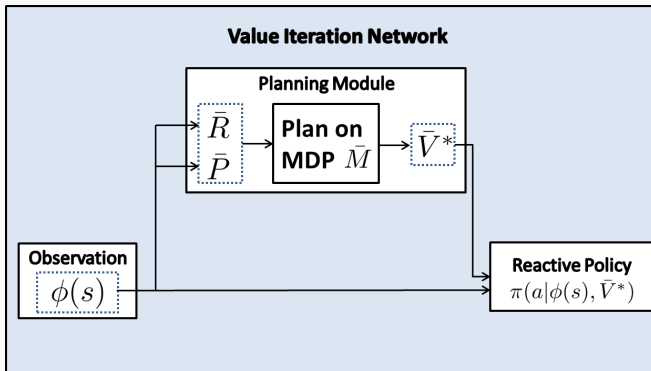
- NNs map observation to reward and transitions
- Later - learn these



How to use the planning computation?

A PLANNING-BASED POLICY MODEL

- Fact 1: value function = sufficient information about plan
- Idea 1: add as features vector to reactive policy

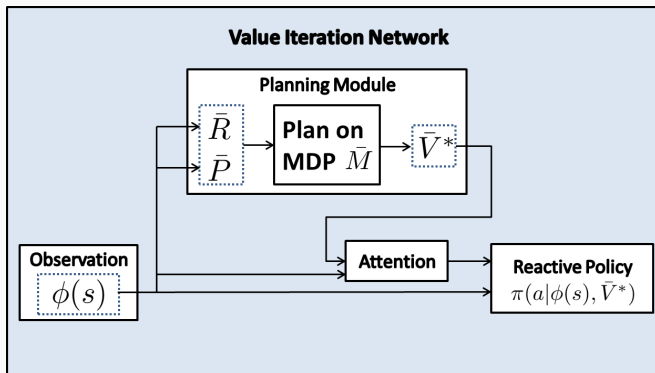


A PLANNING-BASED POLICY MODEL

- Fact 2: action prediction can require only subset of \bar{V}^*

$$\pi^*(a|s) = \arg \max_a R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^*(s')$$

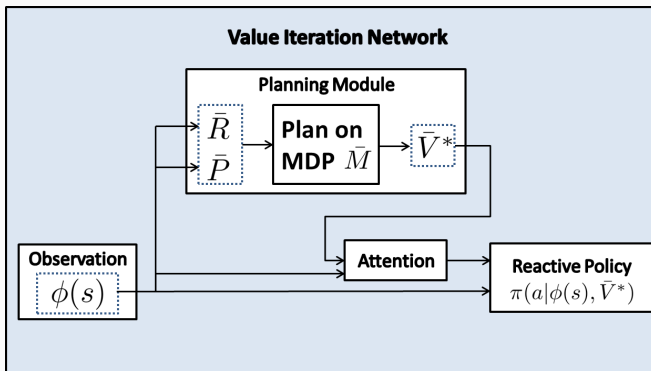
- Similar to **attention** models, effective for learning¹



¹Xu et al. ICML 2015

A PLANNING-BASED POLICY MODEL

- Policy is still a mapping $\phi(s) \rightarrow \text{Prob}(a)$
- Parameters θ for mappings \bar{R} , \bar{P} , attention
- Can we backprop?



How to backprop through planning computation?

VALUE ITERATION = CONVNET

VALUE ITERATION = CONVNET

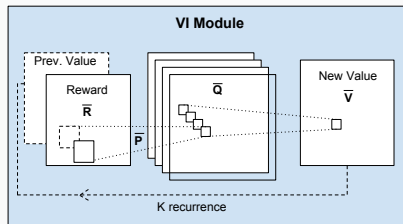
Value iteration

K iterations of:

$$\bar{Q}_n(\bar{s}, \bar{a}) = \bar{R}(\bar{s}, \bar{a}) + \sum_{\bar{s}'} \gamma \bar{P}(\bar{s}' | \bar{s}, \bar{a}) \bar{V}_n(\bar{s}')$$

$$\bar{V}_{n+1}(\bar{s}) = \max_{\bar{a}} \bar{Q}_n(\bar{s}, \bar{a}) \quad \forall \bar{s}$$

Convnet



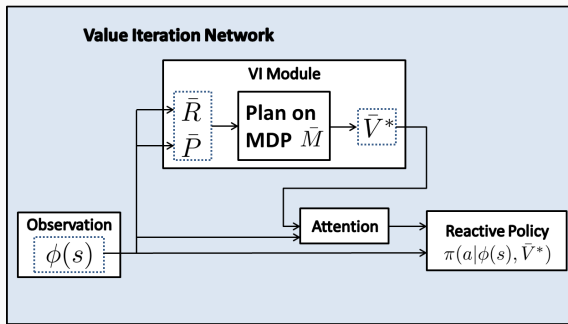
- \bar{A} channels in \bar{Q} layer
- Linear filters $\iff \gamma \bar{P}$
- Tied weights
- Channel-wise max-pooling

- Best for locally connected dynamics (grids, graphs)
- Extension – input-dependent filters

VALUE ITERATION NETWORKS

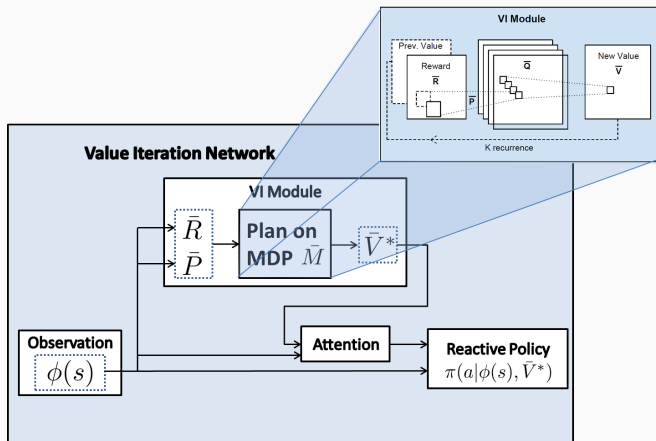
VALUE ITERATION NETWORK

- Use VI module for planning



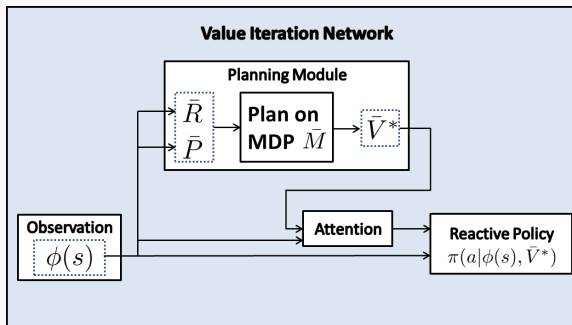
VALUE ITERATION NETWORK

- Value iteration network (VIN)



VALUE ITERATION NETWORK

- Just another policy representation $\pi_{\theta}(a|\phi(s))$
- That can **learn to plan**
- **Train like any other policy!**
- Backprop – just like a convnet
- Implementation – few lines of Theano code



EXPERIMENTS

Questions

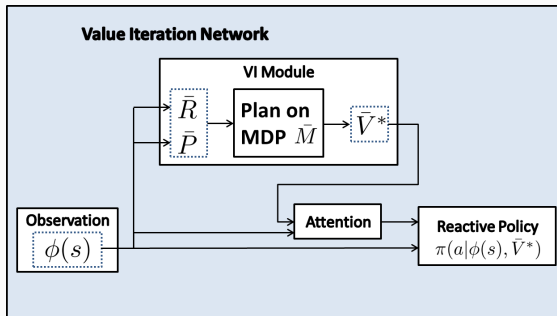
1. Can VINs learn a planning computation?
2. Do VINs generalize better than reactive policies?

GRID-WORLD DOMAIN

- Supervised learning from expert (shortest path)
- Observation: image of obstacles + goal, current state
- Compare VINs with reactive policies

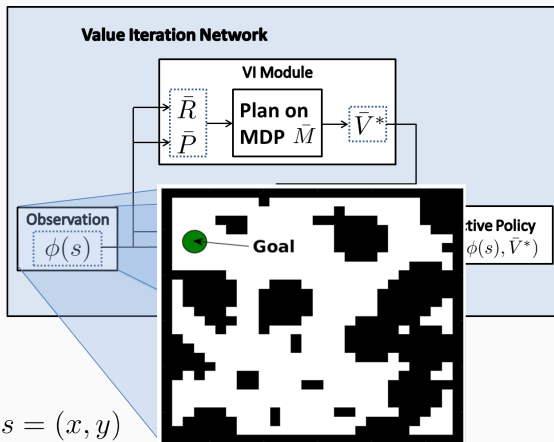
GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



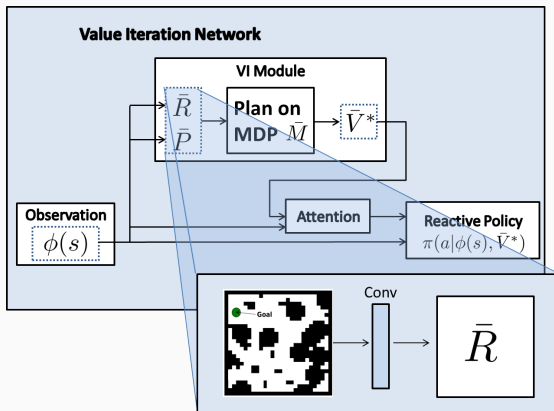
GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



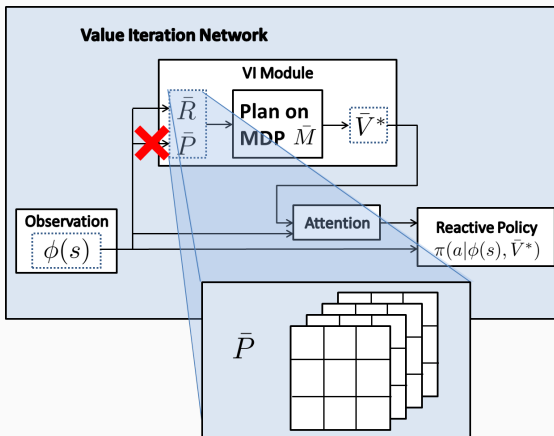
GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



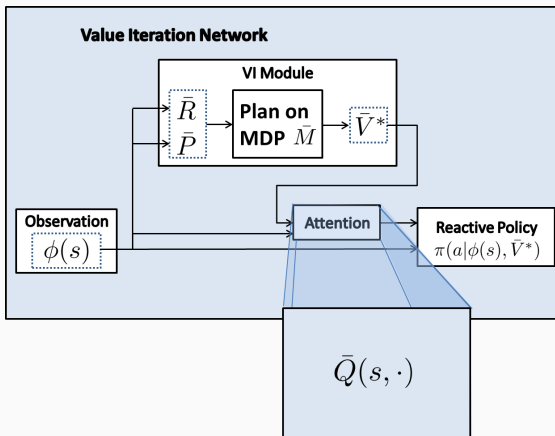
GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



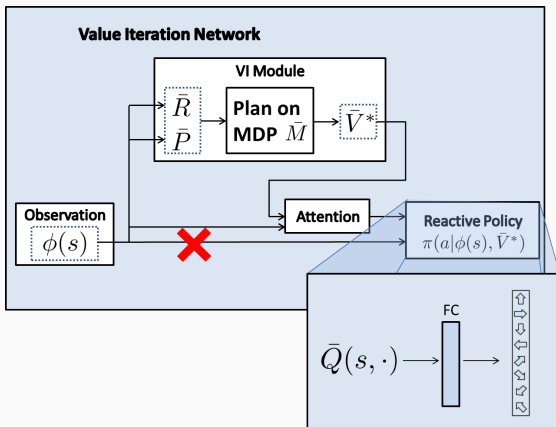
GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



GRID-WORLD DOMAIN

- VI state space: grid-world
- VI Reward map: convnet
- VI Transitions: 3×3 kernel
- Attention: choose \bar{Q} values for current state
- Reactive policy: FC, softmax



Compare with:

- CNN inspired by DQN architecture¹
 - 5 layers
 - Current state as additional input channel
- Fully convolutional net (FCN)²
 - Pixel-wise semantic segmentation (labels=actions)
 - Similar to our attention mechanism
 - 3 layers
 - Full-sized kernel – receptive field always includes goal

Training:

- 5000 random maps, 7 trajectories in each
- Supervised learning from shortest path

¹Mnih et al. Nature 2015

²Long et al. CVPR 2015

Evaluation:

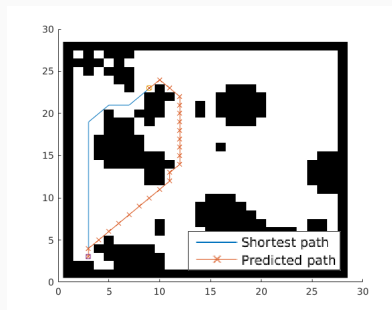
- Action prediction error (on test set)
- Success rate – reach target without hitting obstacles

Results:

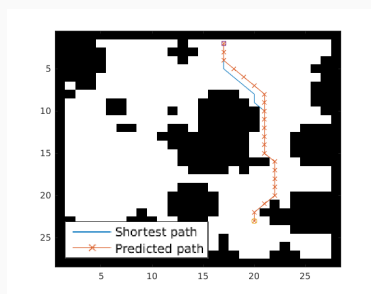
| Domain | VIN | | CNN | | FCN | |
|---------|-----------------|--------------|------------|------------|------------|------------|
| | Prediction loss | Success rate | Pred. loss | Succ. rate | Pred. loss | Succ. rate |
| 8 × 8 | 0.004 | 99.6% | 0.02 | 97.9% | 0.01 | 97.3% |
| 16 × 16 | 0.05 | 99.3% | 0.10 | 87.6% | 0.07 | 88.3% |
| 28 × 28 | 0.11 | 97% | 0.13 | 74.2% | 0.09 | 76.6% |

VINs learn to plan!

Results:



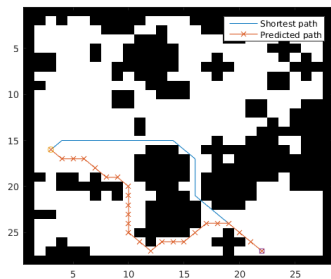
Results:



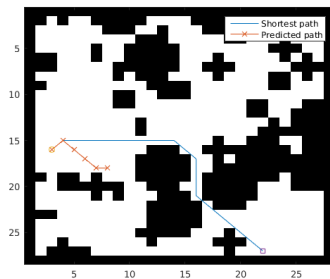
GRID-WORLD DOMAIN

Results:

VIN



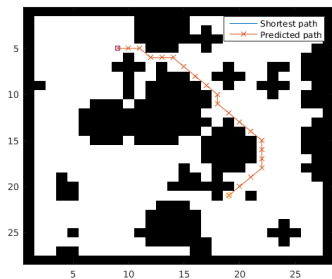
FCN



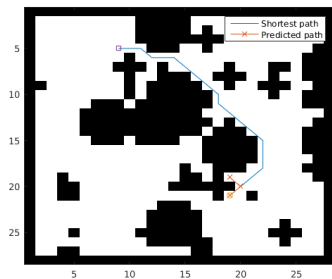
GRID-WORLD DOMAIN

Results:

VIN

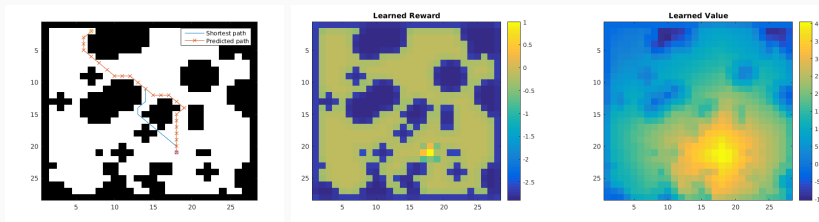


FCN



GRID-WORLD DOMAIN

Results:

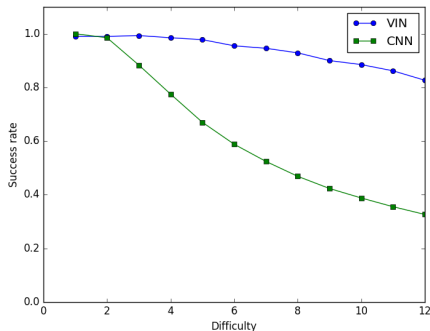


Depth vs. Planning

- Planning requires **depth** – why not just add more layers?
- Experiment: untie weights in VNs
 - Degrades performance
 - Especially with less data
- **The VI structure is important**

Training using RL

- Q-learning, TRPO¹
- Same network structure
- Curriculum learning for exploration
- Similar findings as supervised case

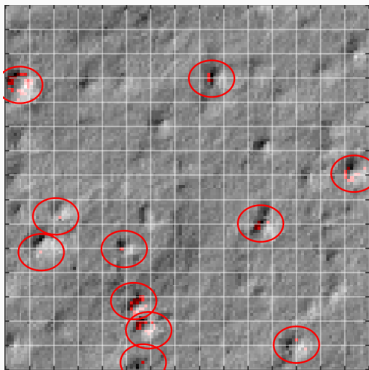


¹Schulman et al. ICML 2015

MARS-NAVIGATION DOMAIN

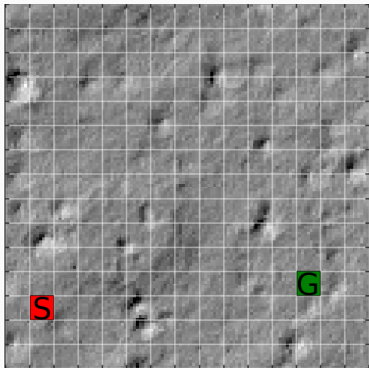
MARS-NAVIGATION DOMAIN

- Grid-world with **natural image** observations
- Overhead images of Mars terrain
- Obstacle = slope of 10° or more
- Elevation data **not part of input**



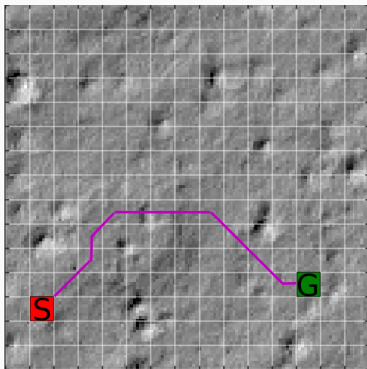
MARS-NAVIGATION DOMAIN

- Grid-world with **natural image** observations
- Overhead images of Mars terrain
- Obstacle = slope of 10° or more
- Elevation data **not part of input**



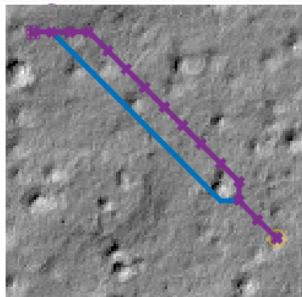
MARS-NAVIGATION DOMAIN

- Grid-world with **natural image** observations
- Overhead images of Mars terrain
- Obstacle = slope of 10° or more
- Elevation data **not part of input**



Same grid-world VIN, 3 layers in \bar{R} convnet

| | Pred. loss | Succ. rate |
|-----------------|------------|------------|
| VIN | 0.089 | 84.8% |
| Best achievable | - | 90.3% |

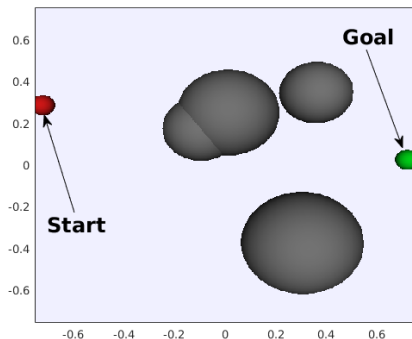


- Best achievable: train classifier with **obstacle labels**, predict map and plan
- VIN **did not** observe any labeled obstacle data
- Conclusion: can handle non-trivial **perception**

CONTINUOUS CONTROL DOMAIN

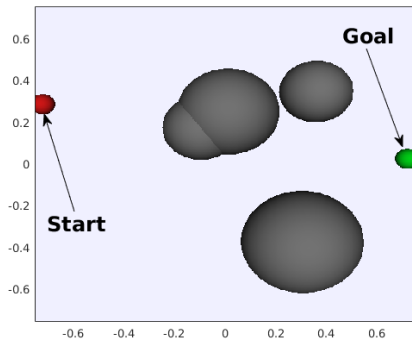
CONTINUOUS CONTROL DOMAIN

- Move particle between obstacles, stop at goal
- 4d state (position, velocity), 2d action (force)
- Input: state + low-res (16×16) map

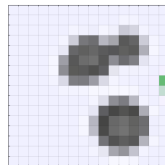


CONTINUOUS CONTROL DOMAIN

- Move particle between obstacles, stop at goal
- 4d state (position, velocity), 2d action (force)
- Input: state + low-res (16 × 16) map

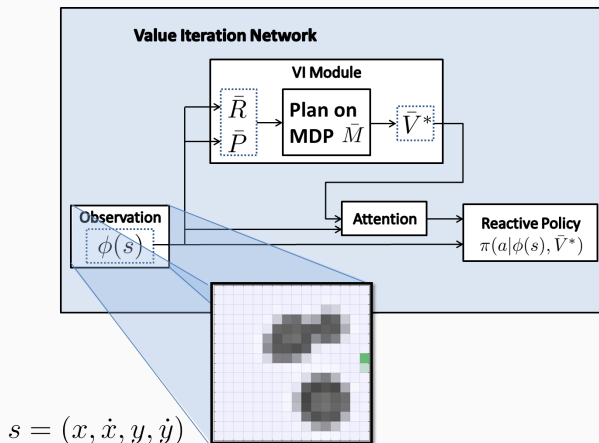


Input map



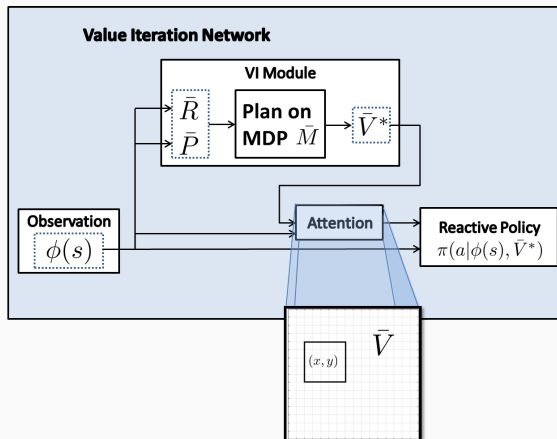
CONTINUOUS CONTROL DOMAIN

- VI state space: grid-world
- Attention: 5×5 patch around current state
- Reactive policy: FC, Gaussian mean output



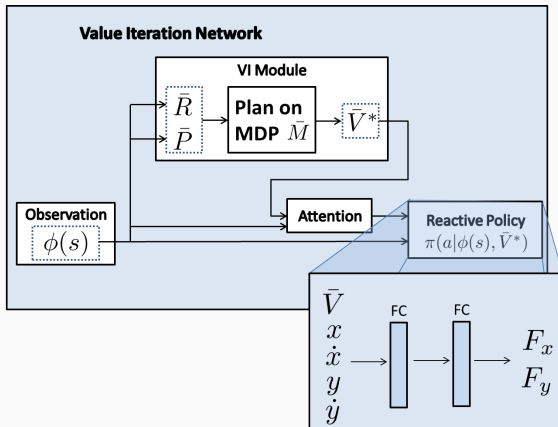
CONTINUOUS CONTROL DOMAIN

- VI state space: grid-world
- Attention: 5×5 patch around current state
- Reactive policy: FC, Gaussian mean output



CONTINUOUS CONTROL DOMAIN

- VI state space: grid-world
- Attention: 5×5 patch around current state
- Reactive policy: FC, Gaussian mean output



Compare with:

- CNN inspired by DQN architecture^{1,2}
 - 2 conv layers + 2 × 2 pooling + 3 FC layers

Training:

- 200 random maps
- iLQG with unknown dynamics³
- Supervised learning (equiv. 1 iteration of guided policy search)

¹Mnih et al. Nature 2015

²Lillicrap et al. ICLR 2016

³Levine & Abbeel, NIPS 2014

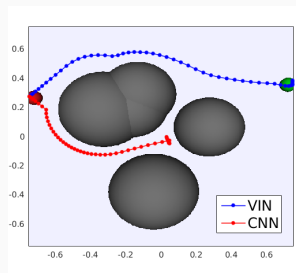
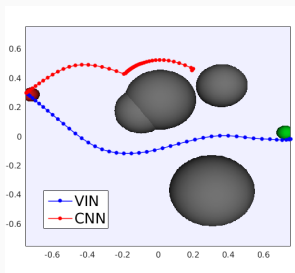
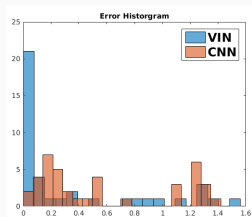
CONTINUOUS CONTROL DOMAIN

Evaluation:

- Distance to goal on final time

Results:

| Network | Train Error | Test Error |
|---------|-------------|------------|
| VIN | 0.30 | 0.35 |
| CNN | 0.39 | 0.58 |



WEB-NAV DOMAIN – LANGUAGE-BASED SEARCH

- "End-to-End Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query

The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II.

The mechanical parts act in such a way as to form a varying electrical circuit—the actual encipherment of a letter is performed electrically. When a key is pressed, the circuit is completed; current flows through the various components and ultimately lights one of many different lamps, indicating the output letter.

The screenshot shows the 'Subject Index' page of Wikipedia for Schools. On the left is a navigation menu with categories like Art, Business Studies, Citizenship, Countries, Design and Technology, Geography, History, Information Technology, Language and Literature, Mathematics, Music, People, Politics, Religion, and Science. The main content area is titled 'Subject Index' and contains a 'Did you know...' section with a note about the 800 Cookies project. Below this is a grid of 10 thumbnail images, each with a subject label: Art (The Starry Night), Business Studies (a factory), Citizenship (a person on a horse), Countries (Euro coins), Design and Technology (a baseball), and others. A 'Title Word Index' link is at the bottom left.

- "End-to-End Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query

The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II.

The mechanical parts act in such a way as to form a varying electrical circuit—the actual encipherment of a letter is performed electrically. When a key is pressed, the circuit is completed; current flows through the various components and ultimately lights one of many different lamps, indicating the output letter.

WIKIPEDIA FOR SCHOOLS

Subject Index











Did you know:
506 Children produced this website for schools as well as this online website about Africa. All received an award for their achievement from 2013 October on every other a family home to the charity. Please here...

The contents of this Schools Wikipedia has been organised by subject along the lines of subjects in the UK's articles, some of which have been listed under several subjects. The general principle has been that this site helpful rather than categorising. So for example we have still listed Pluto in the category of planets even though gained any more and Russia as a country has been included in both Europe and Asia. We have often listed biographies (of which there are 856) but also under the area of study for which they were famous. If you still article we suggest you try the title index which includes every word of the title of every article (including alternate listed under Pythagoras, Food and Goods. Most topics related to the nature world, including a lot of information on medicine, insects and so on are under Science / Biology. This classification was done by hand so please find version in 2005 we found a helpful volunteer had classified Brussels as a vegetable!

Subjects

- Art
- Business Studies
- Civics
- Counting
- Countries
- Design and Technology
- Geography
- History
- Information Technology
- Language and Literature
- Mathematics
- Music
- People
- Plants
- Religion
- Science

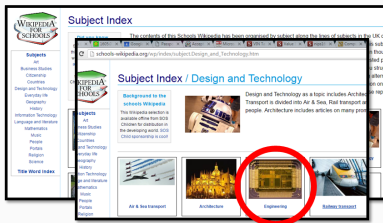
Title Word Index

| | | | | |
|---|---|--|--|---|
|  Art |  Business Studies |  Civics |  Countries |  Design and Technology |
|  Counting |  Countries |  Design and Technology |  History |  Information Technology |

- "End-to-End Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query

The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II.

The mechanical parts act in such a way as to form a varying electrical circuit—the actual encipherment of a letter is performed electrically. When a key is pressed, the circuit is completed; current flows through the various components and ultimately lights one of many different lamps, indicating the output letter.



WEB-NAV DOMAIN

- "End-to-End Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query

The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II. The mechanical parts act in such a way as to form a varying electrical circuit—the actual encipherment of a letter is performed electrically. When a key is pressed, the circuit is completed; current flows through the various components and ultimately lights one of many different lamps, indicating the output letter.

The screenshot shows the 'Subject Index / Design and Technology / Engineering' page on the Wikipedia for Schools website. The page is organized into several sections:

- Background to the subject Wikipedia:** A text box explaining that the content of this subject Wikipedia has been organized by subject along the lines of subjects in the UK curriculum.
- Subjects:** A list of subjects including Art, Business Studies, Citizenship, Countries, Maths, People, Physics, Religion, and Science. 'Engineering' is highlighted in blue.
- Background to the subject Wikipedia:** A text box stating that the content of this subject Wikipedia has been organized by subject along the lines of subjects in the UK curriculum.
- Engineering:** A text box defining engineering as the application of science, technology, in design and production of objects, tools or processes. It lists various types of engineers and their work.
- Automated letter machine:** A list of related terms including Amplifier, Broilers, Bluetooth, Canal, Chemical disaster, Clock, Crash test dummy, Eifel Aqueduct, Electronics, Eurostar, and Forth Road Bridge.
- Enigma machine:** This term is circled in red in the original image. It is listed under the 'Automated letter machine' section.

WEB-NAV DOMAIN

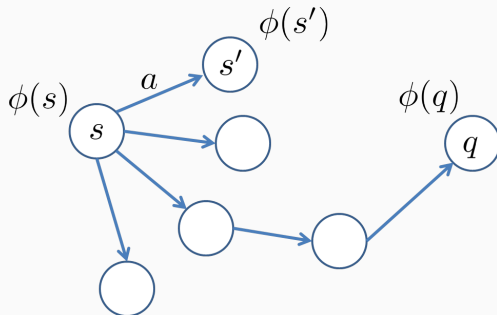
- "End-to-Ends Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query

The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II.

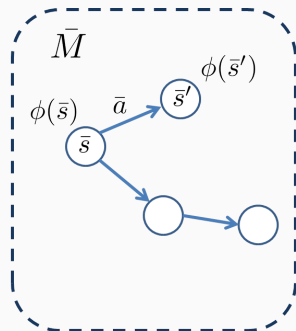
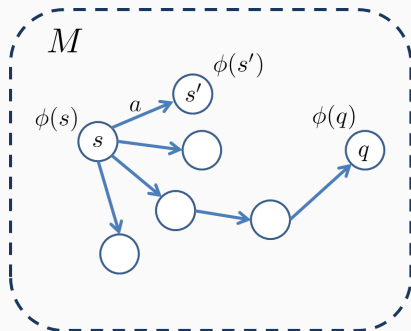
The mechanical parts act in such a way as to form a varying electrical circuit—the actual encipherment of a letter is performed electrically. When a key is pressed, the circuit is completed; current flows through the various components and ultimately lights one of many different lamps, indicating the output letter.

The image displays a sequence of overlapping screenshots illustrating a web navigation path. The top-most screenshot shows the 'Subject Index' page of the 'Wikipedia for Schools' website, with a list of subjects on the left and a search bar at the top. The middle screenshot shows the 'Subject Index / Design and Technology / Engineering' page, with a list of subjects on the left and a search bar at the top. The bottom-most screenshot shows the 'Enigma machine' article page, with a list of subjects on the left and a search bar at the top. A red dashed box highlights the text in the article: 'The Enigma was used commercially from the early 1920s on, and was also adopted by the military and governmental services of a number of nations—most famously by Nazi Germany before and during World War II.' The article also includes a photograph of the machine and a caption: 'The plaintext keyboard, lamps, and rotor wheels of the rotor emerging from the inner lid of a three rotor German military Enigma machine.' The article title is 'Enigma machine' and the page has a 'Checked Content' badge.

- "End-to-End Goal-Driven Web Navigation" Nogueira & Cho, arXiv 2016
- Navigate website links to find a query
- Observe: $\phi(s), \phi(q), \phi(s'|s, a)$
- Features: average word embeddings
- Baseline policy: $h = \text{NN}(\phi(s), \phi(q)), \quad \pi(a|s) \propto \exp(\langle h, \phi(s') \rangle)$

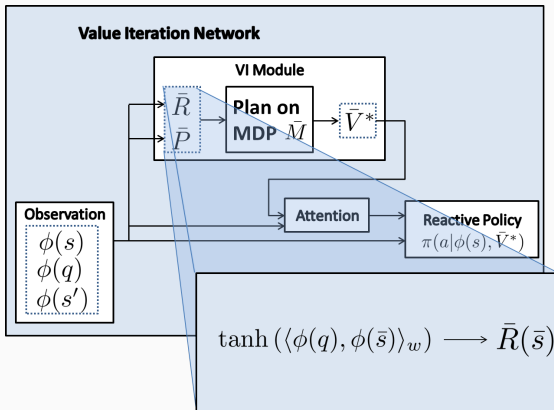


- Idea: use an approximate graph for planning
- Wikipedia for Schools website (6K pages)
- Approximate graph: 1st+2nd level categories (3%)



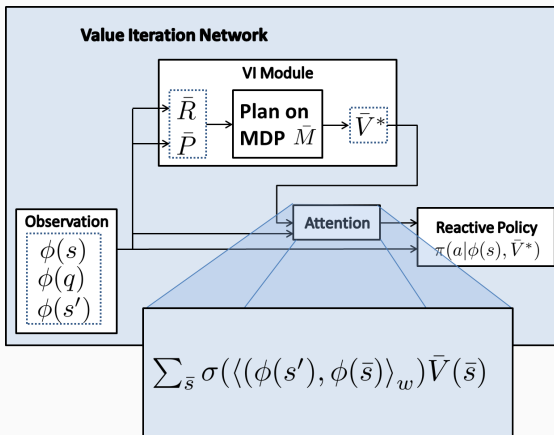
WEB-NAV DOMAIN

- VI state space + transitions : approx. graph
- VI Reward map: weighted similarity to $\phi(q)$
- Attention: average weighted by similarity to $\phi(s')$
- Reactive policy: add feature to $\phi(s')$



WEB-NAV DOMAIN

- VI state space + transitions : approx. graph
- VI Reward map: weighted similarity to $\phi(q)$
- Attention: average weighted by similarity to $\phi(s')$
- Reactive policy: add feature to $\phi(s')$



Evaluation:

- Success – all correct actions within top-4 predictions
- Test set 1: start from index page

Results:

| | | |
|----------|---------------|--|
| Network | Success set 1 | |
| Baseline | 1025/2000 | |
| VIN | 1030/2000 | |



Evaluation:

- Success – all correct actions within top-4 predictions
- Test set 1: start from index page
- Test set 2: start from random page

Results:

| Network | Success set 1 | Success set 2 |
|----------|---------------|---------------|
| Baseline | 1025/2000 | 304/4000 |
| VIN | 1030/2000 | 346/4000 |



Evaluation:

- Success – all correct actions within top-4 predictions
- Test set 1: start from index page
- Test set 2: start from random page

Results:

| Network | Success set 1 | Success set 2 |
|----------|---------------|---------------|
| Baseline | 1025/2000 | 304/4000 |
| VIN | 1030/2000 | 346/4000 |

● Preliminary results: full English Wikipedia website, using wiki-school as approximate graph

SUMMARY & OUTLOOK

- Learn to plan → generalization
- Framework for planning based NN policies
 - Motivated by dynamic programming theory
 - Differentiable planner (VI = CNN)
 - Compositionality of NNs – perception & control
 - Exploits flexible prior knowledge
 - Simple to use

- Different planning algorithms
 - MCTS
 - Optimal control¹
 - Inverse RL²
- How to obtain approximate planning problem
 - Game manual in Atari
- Generalization in RL³
 - theory?
 - benchmarks?
 - Algorithms?
- Generalization \neq lifelong RL, transfer learning⁴
- Hierarchical policies, but not options/skills/etc.

¹Watter et al. NIPS 2015

²Zucker & Bagnell, ICRA 2011

³Oh et al. ICML 2016, Barreto et al. arXiv 2016

⁴Taylor & Stone, JMLR 2009

THANK YOU!
