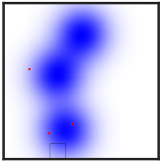
Learning to Search for Targets

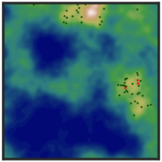
with Deep Reinforcement Learning

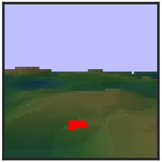
Oskar Lundin

Linköping University

June 7, 2022







Outline

Description

Autonomous search for a set of targets a visual environment with a camera.

- Camera views limited region of environment.
- Moving camera changes visible region.
- Detect when targets are visible.
- Locate targets in minimum time.
- Learn control from a set of sample scenarios.
- Use deep reinforcement learning.

Motivation

- Autonomous systems may reduce risks and cost.
- Applications in search and rescue, surveillance, home assistance, etc.
- Handcrafted systems difficult to design.
- Learning system applicable as long as data is available.
- Subtle patterns and difficult planning decisions.

Aspects

- Random or exhaustive search sufficient in small or random environments.
- Most real-world search tasks exhibit structure.
- Visual cues can be used to find targets quicker.
  - Books are in bookshelves.
  - Cars can be found on roads.
  - Some targets spread out/close together.
- Patterns and cues may be subtle and difficult to pick up.
- Manually engineering a searching system with domain knowledge be difficult and costly.
- **Can a system learn to search intelligently from a set of samples and generalize to similar search tasks?**

Challenges

- Prioritize regions with high probability of targets based on previous experience.
- Learn correlations between scene appearance and target probability.
- Search exhaustively while avoiding searching the same region twice.
- Remember features of searched regions.
- Real-world tasks have limited number of training samples.

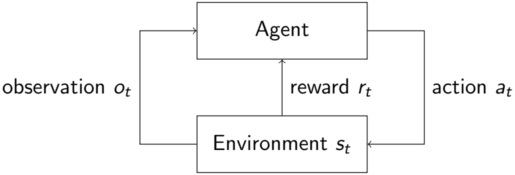
Research Questions

1. How can an agent that learns to intelligently search for targets be implemented with deep reinforcement learning?
2. How does the learning agent compare to random walk, exhaustive search, and a human searcher with prior knowledge of the searched scenes?
3. How well does the learning agent generalize from a limited number of training samples to unseen in-distribution search scenarios?

Markov Decision Process (MDP)

Framework for modeling decision making in partly random processes.
In our case, *partially observable* MDP [1]:

- *Agent* interacts with *environment* over discrete time steps $t = 0, 1, 2 \ldots, T$.
- Takes *action* a_t in state s_t.
- Perceives (partial) *observation* of state o_t.
- Receives scalar reward r_t that indicates whether action is good or bad.
- New state s_{t+1} depends only on history of interactions.
- Agent usually maintains some internal state depending on history $\rightarrow$ memory.



Reinforcement Learning (RL)

Paradigm for learning from interactions how to achieve a goal.

- Tasks usually formalized as (partially observable) MDPs.
- Policy $\pi(a|s)$ is a mapping from states to action probabilities.
- Find π that maximizes cumulative reward $\mathbb{E}\left[\sum_{k=0}^{T} \gamma^{k-t-1} r_k\right]$.
- Often involves estimating the value $v_\pi(s)$ of a state under policy π (useful for training).

Deep RL: Approximate π (and v_π) with deep neural networks. Has been used to play Atari [2], Go [3], StarCraft II [4], etc.

Problem Statement

- Agent searches scene $S \subset \mathbb{R}^d$.
- Scene contains set of targets $\{t_0, \ldots t_n\}$, $t_i \in S$.
- Agent perceives view $V \subset S$.
- Move actions transform view to new subspace.
- Final action indicates that a target is in view.
- Select actions that maximize the probability of finding all targets while minimizing cost in time.
- NP complete [5], intractable to solve optimally.

Environments

- Three simulated environments used for experiments.
- Search space discretized into 10×10 camera positions.
- Each camera position has a unique view $V \subset S$.
- Three targets in all scenes.
- Target probability correlated with scene appearance.
- Should be possible to do better than exhaustive search on average.
- Scenes procedurally generated:
  - Pseudorandom seed determines scene appearance and target positions.
  - Gives control over difficulty to solve.
  - Can vary training and test set sizes.

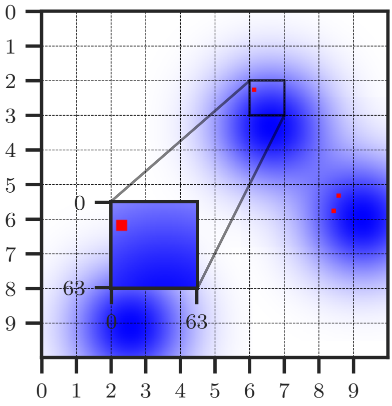
Observation, Action and Reward

At each time step t:

- The agent receives observation $o_t = \langle x_t, p_t \rangle$, where
  - $x_t \in \mathbb{R}^{3\times 64\times 64}$ is an RGB image of current view, and
  - $p_t \in \{0, \ldots, 9\} \times \{0, \ldots, 9\}$ is the position of the camera.
- Takes action $a_t \in \{\texttt{INDICATE}, \texttt{UP}, \texttt{DOWN}, \texttt{LEFT}, \texttt{RIGHT}\}$, where
  - `INDICATE` indicates that a target is in view, and
  - `UP`, `DOWN`, `LEFT`, `RIGHT` move the view in each cardinal direction.
- Receives reward $r_t = h - 0.01 + 0.005d + 0.005e$ where
  - $h = |T \cap V|$ if $a_t = \texttt{INDICATE}$.
  - $d = 1$ if a_t moves closer to nearest target.
  - $e = 1$ if a_t moves to new position.
  - Rewarded for finding targets, moving towards them and exploring environment.
  - Constant penalty encourages quick episode completion.

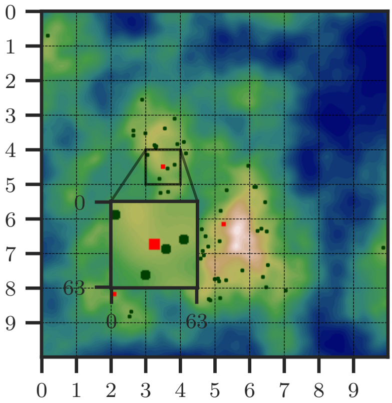
Environment I: Gaussian

- 2D scene.
- Three gaussian kernels with random center.
- Sum of kernels determine appearance of scene and probability of targets.
- Clear correlation between appearance and desired behavior.



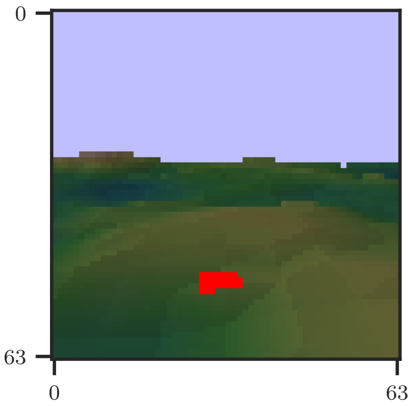
Environment II: Terrain

- Similar to previous environment.
- Terrain seen from above.
- Gradient noise used to generate height map.
- Color determined by height.
- Targets placed with uniform probability across coastlines.
- More realistic, higher variance.
- Analogous to search and rescue with UAV.



Environment III: Camera

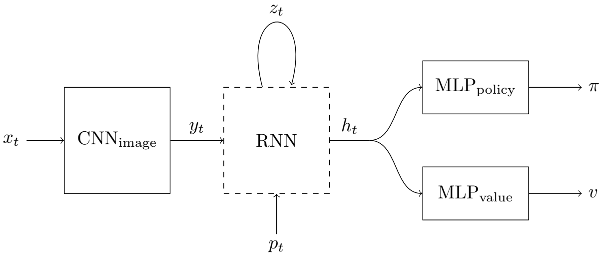
- 3D scene viewed from a perspective projection camera.
- Height map from terrain environment turned into mesh, same appearance and target probability as before.
- Camera location fixed at center of scene.
- Moving actions control pan and tilt (pitch and yaw).
- Visually complex, difficult to interpret.



Approach

Architecture

- Actor-critic method trained with PPO [6].
- Image x_t passed through CNN.
- Latent image representation y_t and position p_t passed through RNN with recurrent state z_t.
- Policy and value heads approximate π and v_π from h_t with MLPs.



Memory

- Recurrent step serves as memory.
- Necessary to remember features of explored environment.
- Two variants evaluated:
    1. Temporal memory (LSTM)
    2. Spatial memory.

1. LSTM:
   - Proven to work for POMDPs [7, 8, 9, 10].
   - May struggle with remembering over many time steps.
   - Important for exhaustive search and scene understanding.
2. Spatial memory (inspired by [11]):
   - Structured memory $M_t \in \mathbb{R}^{C \times 10 \times 10}$ as hidden state (one slot per camera position p_t / unique view V / image x_t).
   - Read vector $r_t = f(M_t)$, f is CNN.
   - Write vector $w_t = g([h_t, r_t])$, g is MLP.
   - Action probabilities $\pi([r_t, w_t])$ and value $v([r_t, w_t])$.
   - r_t contains information from the whole explored scene.
   - w_t written to index p_t of M_{t+1}.

Experiments

1. Search Performance
2. Scaling to Larger Search Spaces
3. Generalization from Limited Samples

- Train for 25M time steps.
- Results reported across 3 runs with different seeds.
- Separate training and test sets.
- Same hyperparameters in all runs.

Implementation

- OpenAI Gym environment interface.
- PyTorch for models and automatic differentiation.
- Intel Core i9-10900X CPU.
- NVIDIA GeForce RTX 2080 Ti GPU.

Experiment I: Search Performance

- Compare to random searcher, exhaustive searcher, human searcher with prior knowledge of scenes.
- Use held out samples as test set.
- Average number of steps on test set.
- SPL metric [12], with N as the number of test samples, S_i indicating success, p_i as the number of steps and l_i as the shortest path length:

$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)} \tag{1}$$

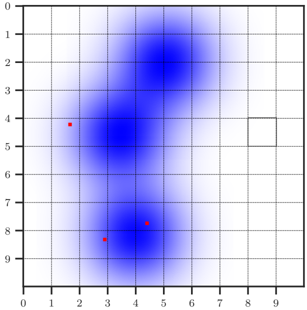
Baselines

- . . .

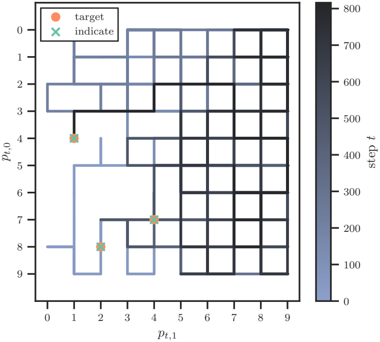
Gaussian Environment

Agent	SPL	Success	Length
random	0.06 ± 0.01	0.92 ± 0.06	369.07 ± 24.93
greedy	0.17 ± 0.00	1.00 ± 0.00	147.12 ± 2.38
exhaustive	0.21 ± 0.00	1.00 ± 0.00	83.37 ± 2.88
handcrafted	0.33 ± 0.00	1.00 ± 0.00	65.20 ± 1.41
human	0.23 ± 0.03	1.00 ± 0.00	80.97 ± 13.49
temporal	0.24 ± 0.03	0.99 ± 0.01	101.25 ± 13.32
spatial	0.29 ± 0.02	0.99 ± 0.01	72.16 ± 5.97

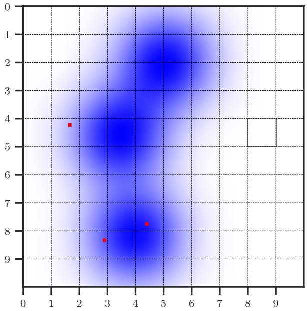
video 1



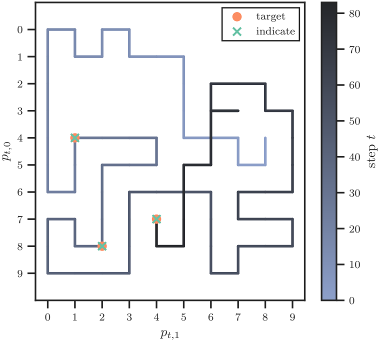
Environment sample



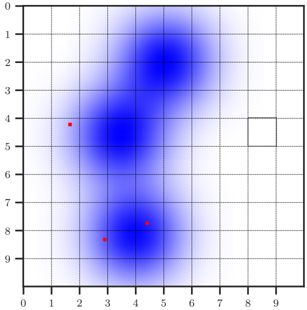
Random baseline



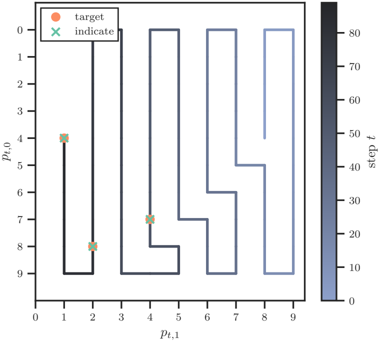
Environment sample



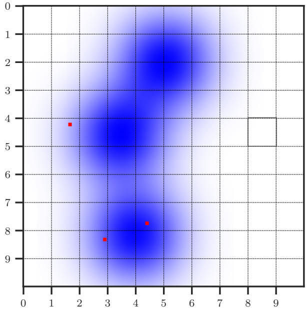
Greedy baseline



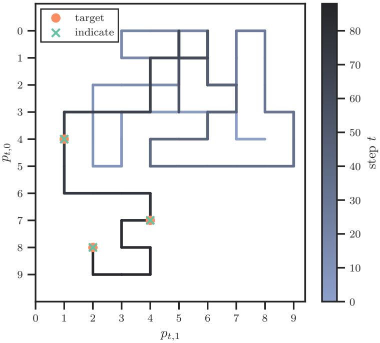
Environment sample



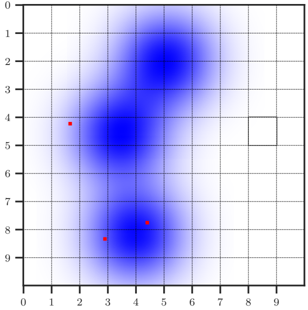
Exhaustive baseline



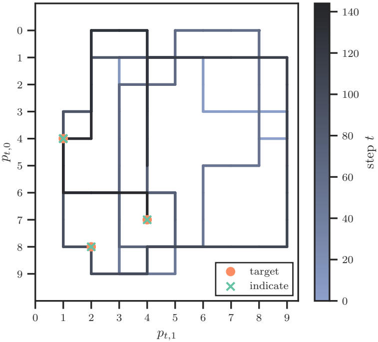
Environment sample



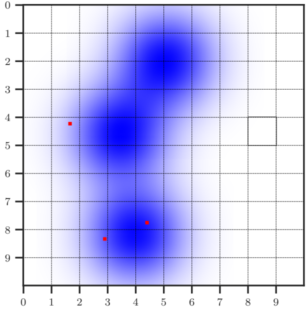
Handcrafted baseline



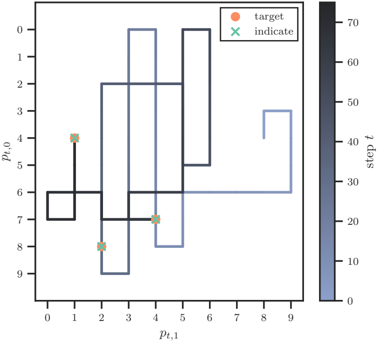
Environment sample



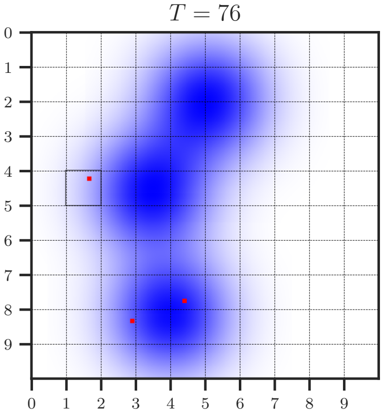
Temporal memory

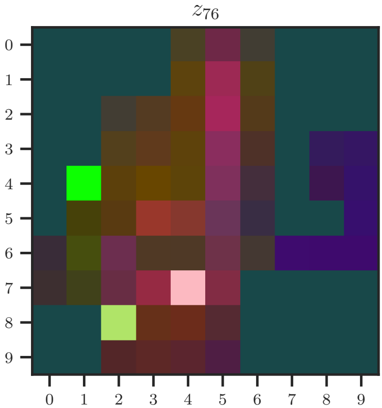


Environment sample



Spatial memory





PCA decomposition of spatial memory after episode.

Terrain Environment

Agent	SPL	Success	Length
random	0.06 ± 0.01	0.89 ± 0.04	366.05 ± 26.96
greedy	0.17 ± 0.01	1.00 ± 0.00	141.01 ± 2.31
exhaustive	0.22 ± 0.00	1.00 ± 0.00	84.11 ± 0.84
human	0.26 ± 0.02	1.00 ± 0.00	76.73 ± 5.33
temporal	0.25 ± 0.02	1.00 ± 0.01	103.76 ± 11.69
spatial	0.27 ± 0.01	1.00 ± 0.00	79.60 ± 6.88

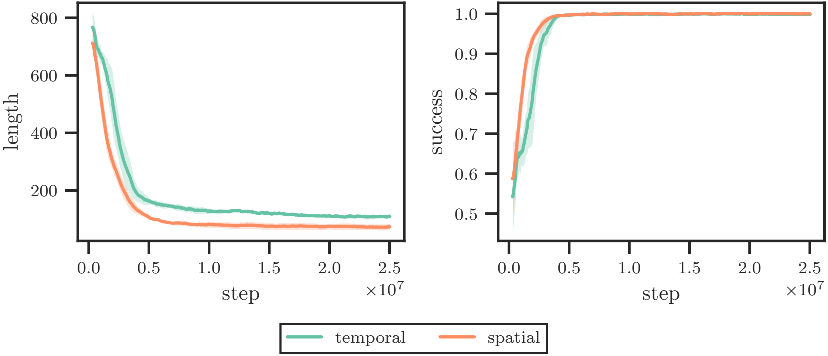
Camera Environment

Agent	SPL	Success	Length
random	0.04 ± 0.00	0.62 ± 0.03	545.09 ± 56.25
greedy	0.12 ± 0.01	0.97 ± 0.01	255.60 ± 10.44
exhaustive	0.37 ± 0.00	1.00 ± 0.00	67.03 ± 0.00
human	0.68 ± 0.08	1.00 ± 0.00	38.10 ± 5.72
temporal	0.70 ± 0.02	1.00 ± 0.00	42.36 ± 2.05
spatial	0.66 ± 0.03	1.00 ± 0.00	42.90 ± 1.73

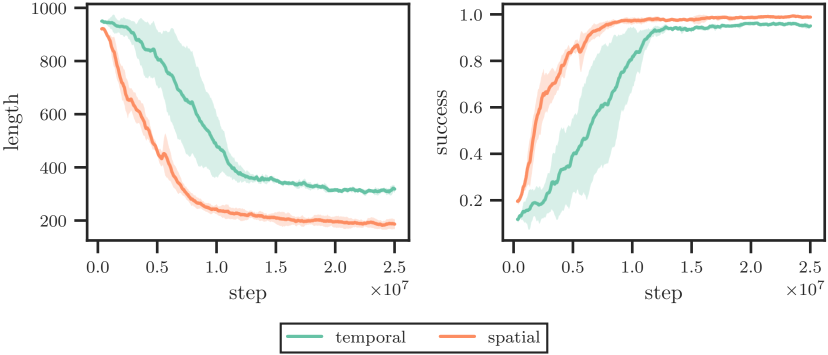
Experiment II: Scaling to Larger Search Spaces

- Larger search spaces take longer to train:
  - More states to explore and exploit.
  - Stronger demands on memory (remember searched positions, scene understanding).
- Investigate impact by comparing agents on 10×10, 15×15, and 20×20 versions of gaussian environment.

10×10

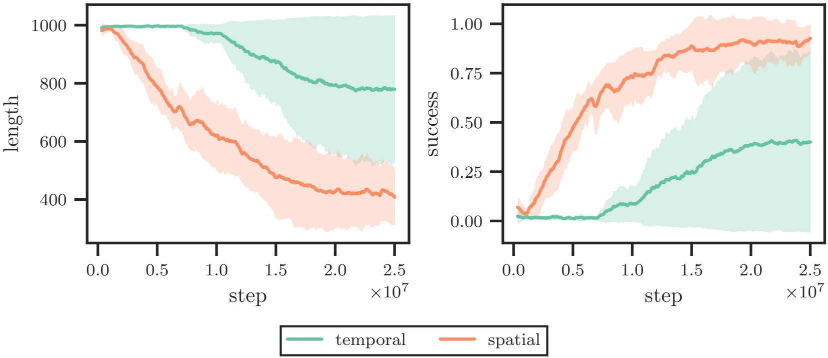

length
800
600
400
200
0.0
0.5
1.0
1.5
2.0
2.5
step
×10^7
success
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.5
1.0
1.5
2.0
2.5
step
×10^7
temporal
spatial

15×15



length
1000
800
600
400
200
0.0
0.5
1.0
1.5
2.0
2.5
step
$\times 10^7$
success
1.0
0.8
0.6
0.4
0.2
temporal
spatial

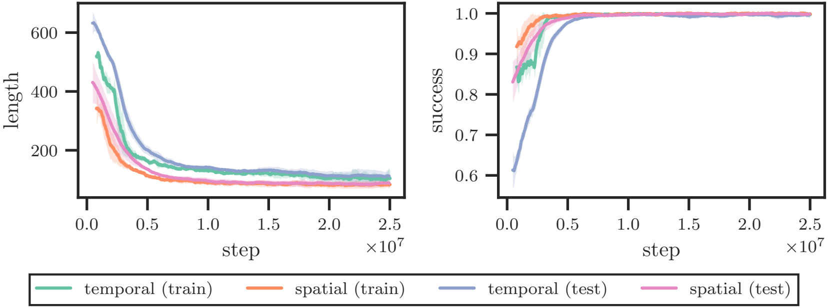
20×20


length
1000
800
600
400
0.0
0.5
1.0
1.5
2.0
2.5
step
×10^7
success
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
2.5
step
×10^7
temporal
spatial

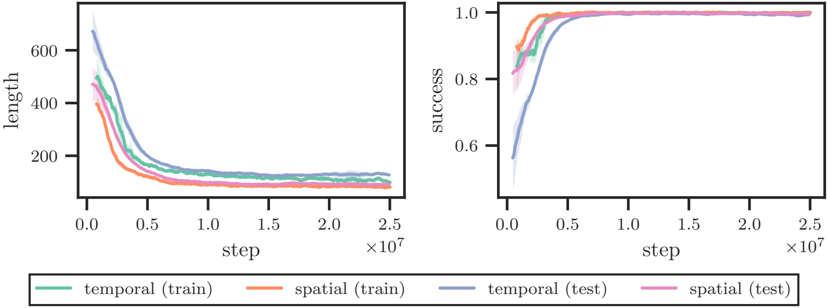
Experiment III: Generalization From Limited Samples

- Limit number of scene samples seen during training to 100, 1000, 10 000, ...
- Use terrain environment, high appearance variance and somewhat realistic.
- Fix seed pool used to generate scenes seen during training.
- Train agents until convergence (or for a fixed number of time steps).
- Test on held out scenes from full distribution.

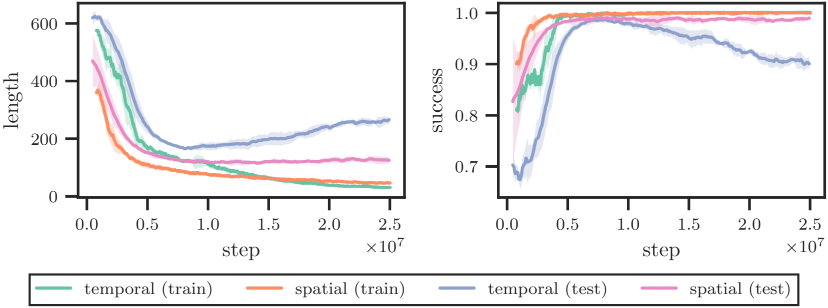
10000 samples



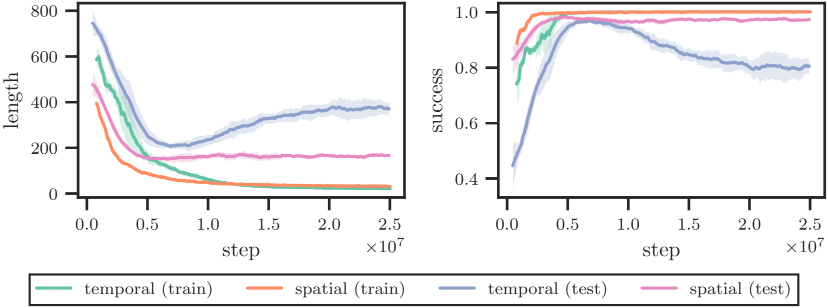
5000 samples



1000 samples



500 samples



length
800
600
400
200
0
0.0
0.5
1.0
1.5
2.0
2.5
step
×10^7
success
1.0
0.8
0.6
0.4
temporal (train)
spatial (train)
temporal (test)
spatial (test)

- Proposed a method for solving visual search with reinforcement learning
- Three environments for evaluating visual search agents
- Compared two neural network architectures with different strengths
- ...

References I

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," vol. 101, no. 1, pp. 99–134.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," vol. 518, no. 7540, pp. 529–533. Number: 7540 Publisher: Nature Publishing Group.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," vol. 529, no. 7587, pp. 484–489. Number: 7587 Publisher: Nature Publishing Group.

References II

[4] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," vol. 575, no. 7782, pp. 350–354. Number: 7782 Publisher: Nature Publishing Group.

[5] A. Andreopoulos and J. K. Tsotsos, "A theory of active object localization," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 903–910. 15 citations (Crossref) [2022-05-19] ISSN: 2380-7504.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms,"

[7] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable MDPs,"

[8] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning,"

References III

[9] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, D. Kumaran, and R. Hadsell, “Learning to navigate in complex environments,”

[10] S. Gupta, V. Tolani, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, “Cognitive mapping and planning for visual navigation,”

[11] E. Parisotto and R. Salakhutdinov, “Neural map: Structured memory for deep reinforcement learning,”

[12] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, “On evaluation of embodied navigation agents,”