
One-Shot Reinforcement Learning for Robot Navigation with Interactive Replay

Jake Bruce

QUT, Brisbane

jacob.bruce@hdr.qut.edu.au

Niko Sünderhauf

QUT, Brisbane

niko.suenderhauf@qut.edu.au

Piotr Mirowski

DeepMind, London

piotrmirowski@google.com

Raia Hadsell

DeepMind, London

raia@google.com

Michael Milford

QUT, Brisbane

michael.milford@qut.edu.au

Abstract

Recently, model-free reinforcement learning algorithms have been shown to solve challenging problems by learning from extensive interaction with the environment. A significant issue with transferring this success to the robotics domain is that interaction with the real world is costly, but training on limited experience is prone to overfitting. We present a method for learning to navigate, to a fixed goal and in a known environment, on a mobile robot. The robot leverages an interactive world model built from a single traversal of the environment, a pre-trained visual feature encoder, and stochastic environmental augmentation, to demonstrate successful zero-shot transfer under real-world environmental variations without fine-tuning.

1 Introduction

Learning from experience is an important capability with potential implications in all areas of robotics. However, interaction with the world can be an expensive undertaking due to constraints such as power usage and human supervision. As a result, a priority for any robotic learning system is minimization of the amount of environmental interaction required to learn a task. Model-free reinforcement learning systems have been shown to solve complex Markov decision processes (MDPs) in a variety of challenging domains, but usually at the cost of a large amount of agent experience (not to be confused with the number of training steps required by the optimization method). In this work we learn to reliably navigate towards a fixed goal in a known, real world environment, using *interactive replay* of a single traversal of the environment. Additional contributions of this paper include using pre-trained visual features, and augmenting the training set with stochastic observations, to demonstrate zero-shot transfer to real-world variations in the environment unseen during training.

To collect enough raw experience to train a model-free system on a real-world task is prohibitively expensive or undesirable in many robotics situations; and a common limitation of mobile robots is limited battery capacity with long recharge times, in which the physical platform is constrained but computational resources are unoccupied. To exploit these limitations and opportunities inherent in mobile robotics, we propose to learn offline using interactive replay from an internal world model [1] that we generate from a single traversal of the environment. In order to combat the tendency to overfit on such a small training environment we make use of a pre-trained and fixed visual feature encoder, and we apply stochastic augmentations to the environment; these modifications enable zero-shot transfer to variations in the environment unseen by the agent during training. Our system differs from previous work on learning to navigate from real-world imagery [2] in that our environments are an order of magnitude larger, and we transfer to a validation environment gathered under natural everyday variation with no fine-tuning required.

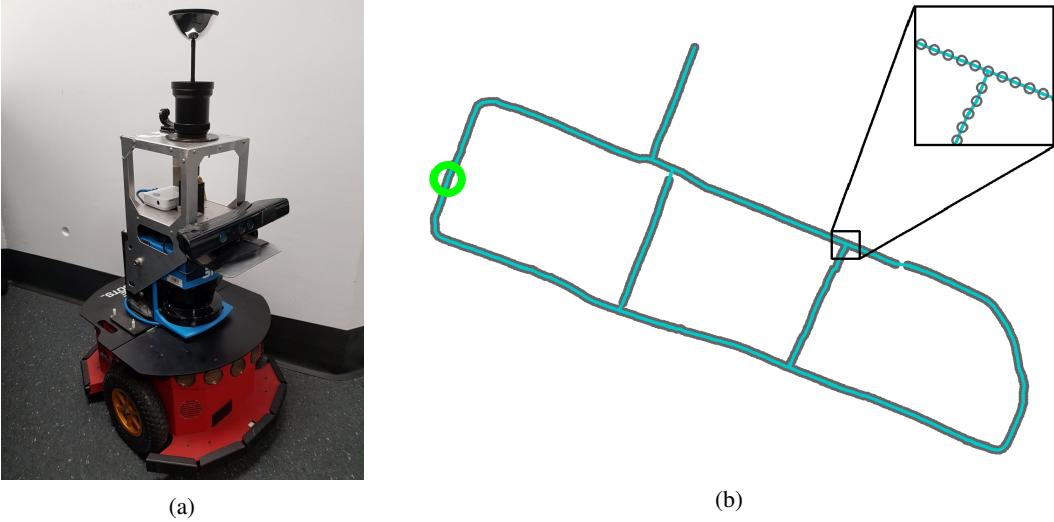


Figure 1: a) Pioneer 3DX mobile robot used in our experiments. b) Graph of the environment built from a single traversal; green circle indicates goal location, and inset shows local connection structure.

2 Related Work

In this section, we briefly review reinforcement learning in general, challenging problems in which it has achieved recent success, and the specific problem of navigation.

2.1 Reinforcement Learning

We formulate the navigation problem as an MDP described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, \mathcal{F})$: \mathcal{S} is the set of agent-environment states, \mathcal{A} is the set of possible actions the agent can perform, $\mathcal{T}(s, a) \rightarrow p(s'|s, a)$ maps a state-action pair to the next state, $\mathcal{O}(s) \rightarrow p(x|s)$ generates agent-visible observations given the true state, and the reward function $\mathcal{F}(s) \rightarrow r$ describes whether the agent has achieved the goal.

Note that the formulation of our MDP allows the observation function $\mathcal{O}(s)$ to be non-deterministic, and demonstrating the transfer performance implications of stochastic observations is a key contribution of this work. The transition function $\mathcal{T}(s, a)$ can also be stochastic, but we consider deterministic transitions for simplicity.

Reinforcement learning (RL) involves finding a policy $\pi(x) \rightarrow a$ to map observations to actions that maximize the expected sum of future rewards, often referred to as the *return*, denoted by R . The return can take many forms, but a common choice is the sum of γ -discounted future rewards $R = \sum_t \gamma^t r_t$. Broad flavors of RL include model-based control in which a predictive model is used to plan actions or learn a policy, episodic control in which highly rewarding trajectories can be replayed from known states, and model-free control in which a policy is learned directly from observations. We consider the model-free setting in this work.

In model-free RL, the policy is implemented by a function approximator, commonly a deep neural network. The system is usually trained either to map observations and actions to the expected future reward (as in Q-learning) or to directly approximate the probability distribution over actions that maximizes expected reward (as in policy search). When using a deep neural network as a function approximator, a policy is learned from experience with the environment, usually by gradient descent on trajectories of (x, a, R) -tuples. In value-based methods such as Q-learning, the policy consists of using a value estimate (the Q function) to choose actions with the highest return, whereas in policy search methods, the policy is parameterized directly and a value function is used to update the parameters of the policy in the direction of positive outcomes. We primarily consider a value-based method known as *bootstrapped Q-learning* [3].

2.2 Recent Successes

Many challenging problems have recently been addressed with model-free RL, with a particular abundance of research on the topic of playing video games directly from pixels. Video games make effective testbeds for evaluating learning algorithms due to full observability and manipulability, the ability to run at high speeds to gather large amounts of agent experience, and the potential for structural and perceptual similarities to real-world problems. Recent examples include learning to play a diverse suite of dozens of low-resolution Atari games with deep Q-networks [4], challenging board games such as Go [5], and first-person shooters such as Doom [6, 7].

Real world tasks have also been a target for RL algorithms, especially for difficult perception and control problems in robotics [8]. Recent contributions include continuous control for simulated robots [9], motor control for manipulation directly from pixels [10], and simulation-to-real transfer for manipulation [11, 12].

The successful application of deep Q-networks [4] involved the use of a passive replay buffer of past experience, which enabled more efficient use of interaction samples and stabilized training. We build on the concept in this work by introducing *interactive replay*, where a single trajectory through a real environment comprises a virtual environment through which an agent can learn to navigate.

2.3 Reinforcement Learning for Navigation

Recent work has also addressed the task of navigating 3D environments in simulation, e.g. maze navigation from perspective view imagery [13, 14]. In robotics, agents have been trained to navigate using laser sensors in simulation [15] and simulated depth imagery [16]; structure-based sensors tend to exhibit less difference between simulation and reality (a narrower “reality gap”) than does visual sensing. However, visual navigation is also an established direction, including work on target-driven image-based navigation in small grid worlds with simulation and fine-tuning in the real world [2].

In this work, we build on the success of navigation learning in virtual environments [14] and preliminary work in the real world [2] to develop techniques for learning to navigate large office environments on a mobile robot, with goal-generated reward as the only supervised training signal.

3 Approach

In this section we describe our approach consisting of the construction of virtual training and validation environments from single traversals of the real world; our use of a pre-trained and fixed visual encoder; stochastic observations to augment the training environment; and the bootstrapped Q-learning algorithm for RL.

3.1 Interactive Replay

The success of deep Q-networks [4] can be attributed partially to the use of a passive replay memory of past experience, from which training batches are sampled in order to make more efficient use of interaction samples and to stabilize training. We build on the concept of experience replay [17, 18], and more specifically that of a replay buffer, by introducing *interactive replay*, in which a rough world model is memorized from a single traversal of the environment. This allows an agent to interact with the model to generate a large number of diverse trajectories for learning to navigate, while minimizing the amount of real-world experience required by the robot.

In this section we describe how the training environment is gathered in one pass, and minimal human annotation is used to construct a pose graph for virtual interactive replay; a validation environment is then constructed from a second pass under different environmental conditions and approximately aligned with the training environment.

Training environment Recording the training environment consists of two phases. First, the robot executes a complete traversal of the environment while recording sensor data. Second, the recorded data is aligned into a topological map of the environment. In our experiments, alignment is conducted by minimal human interaction in which the user is shown the approximate poses of the agent in two-dimensional space (see Fig. 1b), connects loops that were closed during the traversal, and removes

duplicated regions of the space. In the general case, both phases can be accomplished autonomously using simultaneous localization and mapping (SLAM) techniques [19] without human intervention.

The space is then discretized into a pose graph [20] consisting of sensory snapshots taken every Δ_{pos} meters, which are connected by edges representing temporal adjacency and loop closures. In the office environment considered in this work, the agent can rotate in Δ_{rot} increments in either direction, and move forward by Δ_{pos} (if there is no node in that direction, the action has no effect). Nodes in the pose graph may have edges along the forward, backward, left, and right directions, but the pose graph concept generalizes to environments that are not grid-like. The resulting graph comprises the training environment in which the agent will learn to navigate.

Validation environment In order to evaluate the transfer performance of the agent under environmental variations that it has not seen during training, we gather a separate environment to use as validation during training. Collection of the validation environment is simpler than the training environment: the robot is used to conduct a second traversal of the environment, not necessarily in the same order as the first. Then the stored training environment is used in conjunction with an approximate localization method to associate nodes in the pose graph of the training environment with the closest matching pose in the validation recording; in our work we use laser-based localization as an approximate localization system and a 360° camera to align the orientation of the imagery, but the pose correspondences can be annotated by hand if a localization system is not available. The resulting aligned graph comprises a validation environment to estimate transfer performance under realistic environmental variation.

3.2 Rich Visual Encoding

A significant issue with training on a small set of real-world experience is the tendency to overfit to this constrained environment. One area where overfitting can manifest is the learning of visual features, since most current approaches in model-free RL learn a feature extractor from scratch [4, 13, 14], and a single traversal of a small environment provides limited variety of visual structure. Using a visual encoder network that has been trained on a large vision dataset can help compensate for the lack of variety in the training environment [21], so we use a ResNet-50 network trained on ImageNet [22, 23] as in [2] and extract the pre-final fully-connected layer as a 2048-dimensional rich visual encoding of each of the four 90° views from the camera. These are concatenated into an 8192-dimensional vector to form the agent’s observation $f(x_t)$. We also freeze the weights on this encoder to avoid overfitting to the limited structure present in the training environment; this also allows us to pre-compute the features in our recordings, yielding a significant computational advantage during training.

3.3 Stochastic Observations

Data augmentation techniques have been established to compensate for limited training data, in which random transformations are applied to the training examples in order to improve the diversity of the training samples [24]. We leverage this idea in our training environment by manipulating the observation function $\mathcal{O}(s) \rightarrow p(x|s)$ of our MDP to return observations drawn from a distribution of positions and orientations centered on the true pose of the state s . In this work, we use the nodes in the environment’s pose graph as the set of states \mathcal{S} , and generate observations by sampling from a distribution over both position and orientation.

For position, we sample input observations x_t from images recorded “in between” the discretized nodes of the pose graph according to a normal distribution centered on the true state s_t with a variance of 5cm. Orientations are sampled from a normal distribution centered on the true orientation of s_t , with a variance of 5°; the distributions approximate the accuracy of our laser-based localization system. In Section 4 we investigate the degree to which a stochastic observation function reduces overfitting in our navigation problem, as measured by relative performance on the validation set.

3.4 Bootstrapped Q-Learning

We consider as our model-free RL algorithm a double dueling n -step Q-learning framework [25, 26, 13] based on [3] with N_Q parallel Q-function heads. Each head is a deep neural network with shared visual encoder and recurrent layers, but with its own Q-function output layers. One head is chosen at

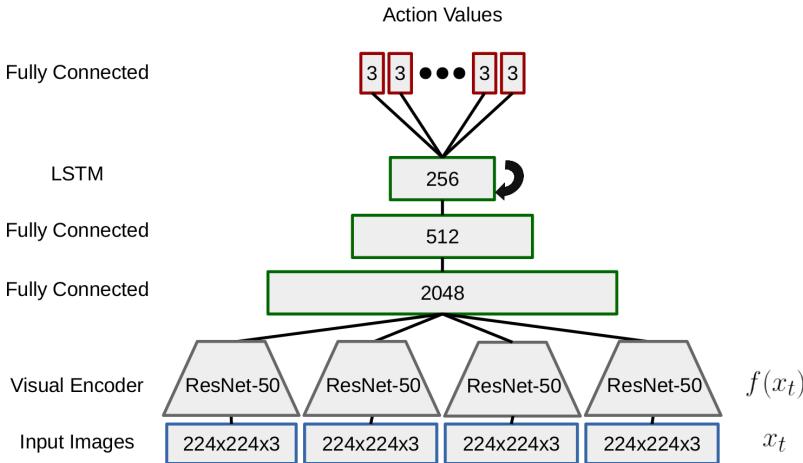


Figure 2: Schematic of our model architecture. Blue outlines indicate inputs; gray outlines indicate modules with fixed weights; green outlines indicate layers with trainable weights; and red outlines indicate the N_Q parallel Q-function output layers. The agent sees in four directions at a time, each consisting of a 90° crop from the robot’s 360° camera.

random as the behavior generator each episode, and each head is trained on a fraction p_Q of the data of all the heads (see [3] for details). We refer to the architecture as *Bootstrapped Q-learning*.

The bootstrapped Q-network is trained by minimizing the traditional Q-learning loss function: $\mathcal{L} = \sum_{\text{batch}} (\tilde{Q}(f(x_t)) - R_t)^2$, where R_t is the exponentially-discounted sum of future rewards. As in [13], we run N_w parallel workers exploring the environment simultaneously using a single shared network, which is updated by stochastic gradient descent on the batches of experience composed of all of the workers’ experience. Note that this parallelism is distinct from the N_Q separate Q-function heads, which comprise an additional level of parallel structured exploration on top of the parallel workers. In this work, we use $N_w = 64$, $N_Q = 10$, and $p_Q = 0.5$.

For an intuitive explanation of the advantage of the bootstrapped approach on our problem, consider performing a random walk. In our environment, as in most robot navigation problems, the optimal prior over actions heavily favors moving forward. By sampling from a diverse set of random priors in the form of randomly-initialized Q-functions, the agent can more reliably encounter trajectories that result from near-optimal priors, enabling faster convergence. As a bonus for the robotics context in particular, the ensemble nature of the technique enables the interpretation of the distinct Q-estimates as a probability distribution, enabling reasoning about the uncertainty in the model. This provides a significant practical advantage over most RL algorithms (although, see [27] for an exception). The architecture of the network is shown in Fig. 2 and corresponds to a standard convolutional recurrent network used in previous work [13, 28, 14]. We opted for the use of a recurrent network (specifically a Long Short-Term Memory [29]) as policy inputs, instead of a plain feed-forward network, because the recurrent network enables the agent to accumulate visual evidence over time.

4 Experiments

Our experiments consider a mobile robot navigation task in which the robot spawns at a random location in the environment and must reach a pre-designated goal location by executing actions from the set {turn_left, turn_right, move_forward}. The goal location is fixed for all experiments. Unlike previous work on goal-driven navigation in synthetic environments [14], the goal location is not visible on the images. The agent learns to recognize the environmental landmarks solely from the photographic inputs. The training environment was constructed from a single traversal of an office environment on a Pioneer 3DX mobile robot with a 360 -degree camera (see Fig. 1a, discretizing position and orientation to $\Delta_{\text{pos}} = 10\text{cm}$ and $\Delta_{\text{rot}} = 90^\circ$. The validation traversal was collected in

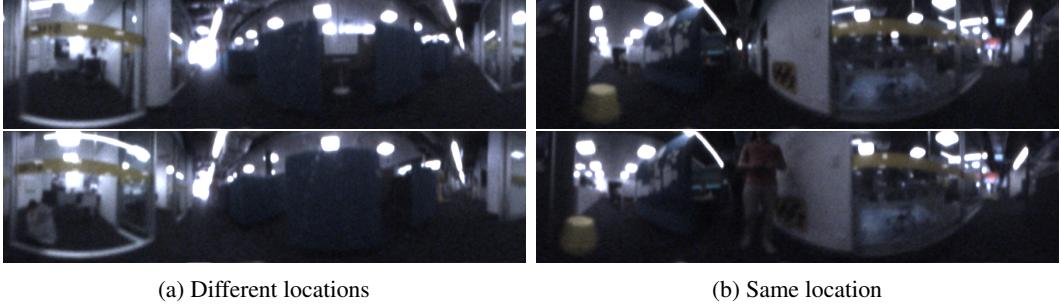


Figure 3: Example imagery from the recorded environment; a) two physically different locations that are visually similar, and b) two views of the same physical location with visual differences.

the same environment on a different day, and the route was traversed in a different order. As a result, the two datasets exhibit typical day-to-day differences in the appearance of an office environment, including moderate variations in lighting and furniture placement, and variations in dynamic obstacles such as humans (see Fig. 3).

The navigation problem consists of each worker spawning at a random node and orientation in the environment at the beginning of each episode, and choosing actions according to its policy until $T_{\max} = 3000$ timesteps have elapsed. A node in the graph is chosen at random as the goal (reward of 1.0), and $N_a = 10$ small sub-goals (reward of 0.1) are placed at random to encourage exploration, as in [14]; both types of goal locations were chosen arbitrarily without researcher input and are fixed for the entire duration of the experiment. Reaching the goal causes the agent to re-spawn in a new random location. Agents were trained for a total of 300 million frames of virtual experience, which is many orders of magnitude more experience than most robots ever obtain in the real world.

We evaluate absolute performance as measured by total reward achieved per episode by the worst-performing worker in the pool, denoted R_{\min} . The metric R_{\min} is important for judging whether the agent has learned to *reliably* navigate the environment: a useful robot should be able to reach its goal from any location. The pose graph recorded for these experiments consisted of 572 nodes, and the longest path in the graph was 227 timesteps. Given this length and T_{\max} of 3000, an optimal agent should be able to achieve worst case $R_{\min} \geq 13$. We include as a baseline the performance of a random walk with near-optimal action prior, using turn probability equal to the fraction of nodes in the graph representing intersections (locations where rotation may be necessary). Note that a uniform action prior achieves negligible reward (< 0.1), which further supports the use of the bootstrap sampling technique.

In addition to measuring reward, we compare the transfer-oriented components of the system according to R_{relative} , the ratio of R_{\min} on the validation environment to R_{\min} on the training environment: this metric measures the fraction of training performance achieved in validation, and is important for determining whether a particular component of the system specifically affects transfer; results are shown in Fig. 5. Qualitatively, successful agents learn to navigate directly to the goal; see Fig. 4 for example trajectories from a trained agent on the validation environment.

RL algorithm Our experiments compare three types of RL algorithms that learn from parallel experience: A2C [13], n -step Q-learning [13], and n -step bootstrapped Q-learning [3]. A2C (a synchronous implementation of A3C in which parallel actions are always chosen from the same model) is a policy search method that directly optimizes a probability distribution over actions, and n -step Q-learning is a value-based method like the bootstrapped approach but which maintains only one Q-function estimate. Results are shown in Figs. 5a and 5b; the bootstrapped Q-learning algorithm achieves significantly better performance. For the robotics context in particular, the ensemble nature of the technique provides a significant advantage in addition to its performance, because the ensemble can be interpreted as a distribution over Q-values to enable reasoning about model uncertainty.

Visual encoding Learning visual features from scratch on a small dataset can result in reduced representational diversity due to a lack of variation in visual structure in the environment. We compare performance when training the convolutional encoder of [13, 14] from scratch, against the fixed

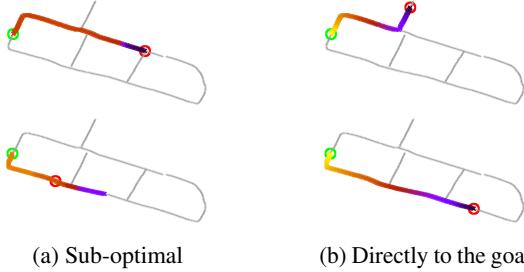


Figure 4: Example trajectories through the validation environment by a trained agent. Path color indicates progressive timesteps (purple to yellow), and red and green circles indicate start and goal locations respectively. a) Two sub-optimal trajectories in which the agent got stuck or backtracked but ultimately reached the goal, and b) two successful trajectories directly to the goal.

ResNet-50 architecture pre-trained on ImageNet; results are shown in Fig. 5f. Pre-trained visual features perform better on the validation environment, incur a lower penalty in transfer performance (see Fig. 5h), and are much more computationally efficient because a fixed encoder allows us to pre-encode all of the images in our environment during training.

Environmental stochasticity Dataset augmentation is a proven technique in machine learning, and we exploit this idea by augmenting our training environment with stochastic observations; results are shown in Figs. 5c and 5d. Stochastic observations significantly improve performance on both the training and validation environments. Agents trained on deterministic observations have poor minimum performance, and cannot be said to reliably solve the task; agents with stochastic observations however achieve nearly optimal minimum performance as computed by an optimal path planner. In addition, transfer performance as shown in Fig. 5g is significantly improved by stochasticity, indicating that the stochastic variation we inject into the observations improve learning in general, and also capture some degree of the natural variation encountered between traversals.

5 Discussion and Future Work

We have presented a reinforcement learning approach for learning to navigate on a mobile robot from a single traversal of the environment. The contributions of this work consist of an interactive replay buffer technique, augmenting the training environment with stochastic observations, and the use of pre-trained visual features. The first contribution enables the generation of extensive training experience needed by model-free systems from real world data; the latter two contributions significantly improve performance on both the training and validation environments, and proved crucial for obtaining reliable zero-shot transfer to variations in the environment unseen during training.

Future work will focus on transfer to situations with agent *and* robot in the loop. Performance on the validation set provides encouraging evidence for the feasibility of closed-loop transfer, although several issues remain to be addressed. In particular since our actions are coarse and discrete, once an action is selected by the agent, movement to the next node in the pose graph will need to be handled by a low-level motion controller. In our setup this will likely be accomplished by a laser-based approximate localization system; however, contributions in visual servoing [30], teach and repeat [31], and image alignment techniques for robot and animal homing [32, 33] provide evidence that a purely visual solution is also possible.

Intriguing evidence from psychology and neuroscience suggests that animals may replay known and novel sequences of experience during sleep, in service of consolidating memory and discovering new routes [34, 35]. In the long term outlook for learning in robotics, clever re-use of recorded experience is likely to be crucial for making the best of limited interaction with the environment, and forms of virtual interactive replay may play an important role in this process.

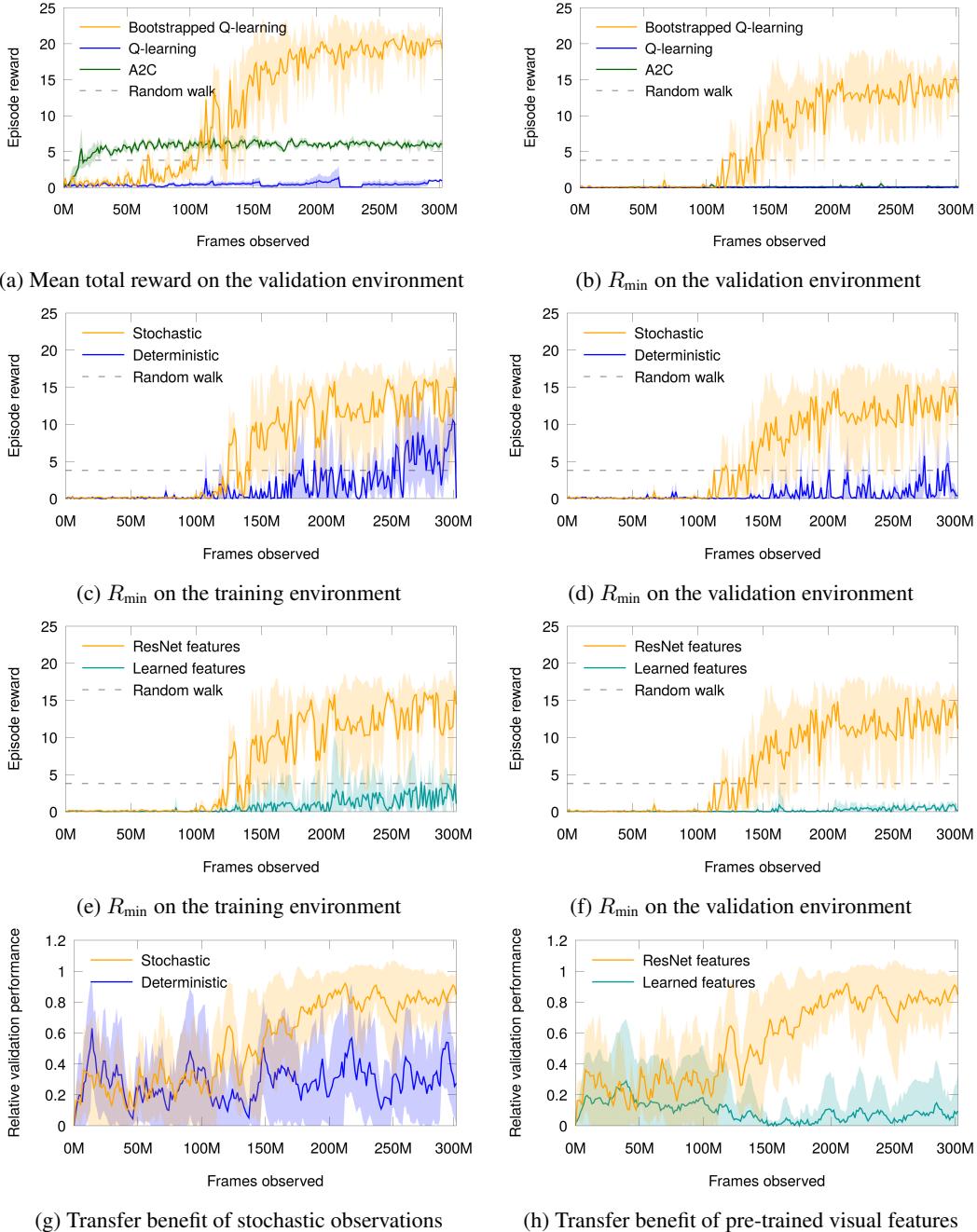


Figure 5: Experimental results on the navigation task. a-f) Minimum episode score R_{\min} , indicating worst case performance, g-h) ratio of validation to training performance. a-b) Bootstrapped Q-learning performs best; the other algorithms fail to reliably solve the task. c-h) Augmenting the environment with stochastic observations and using a pre-trained visual encoder dramatically improve training and transfer performance.

References

- [1] Richard Vaughan and Mauricio Zuluaga. Use your illusion: Sensorimotor self-simulation allows complex agents to plan with incomplete self-knowledge. In *International Conference on Simulation of Adaptive Behavior*, pages 298–309. Springer, 2006.
- [2] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pages 3357–3364. IEEE, 2017.
- [3] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, pages 4026–4034, 2016.
- [4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [5] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.
- [6] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. VizDoom: A Doom-based AI research platform for visual reinforcement learning. In *IEEE Conference on Computational Intelligence and Games*, pages 1–8. IEEE, 2016.
- [7] Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.
- [8] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
- [9] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.
- [10] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.
- [11] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv preprint arXiv:1511.03791*, 2015.
- [12] Andrei A Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. *arXiv preprint arXiv:1610.04286*, 2016.
- [13] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.
- [14] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.
- [15] Lei Tai, Giuseppe Paolo, and Ming Liu. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. *arXiv preprint arXiv:1703.00420*, 2017.
- [16] Jingwei Zhang, Jost Tobias Springenberg, Joschka Boedecker, and Wolfram Burgard. Deep reinforcement learning with successor features for navigation across similar environments. *arXiv preprint arXiv:1612.05533*, 2016.

- [17] James L McClelland, Bruce L McNaughton, and Randall C O’reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419, 1995.
- [18] Long-Ji Lin. Reinforcement learning for robots using neural networks. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.
- [19] MWM Gamini Dissanayake, Paul Newman, Steve Clark, Hugh F Durrant-Whyte, and Michael Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, 2001.
- [20] Sebastian Thrun and Michael Montemerlo. The graph SLAM algorithm with applications to large-scale mapping of urban structures. *The International Journal of Robotics Research*, 25(5-6):403–429, 2006.
- [21] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*, 2015.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [24] Zhe Gan, Ricardo Henao, David Carlson, and Lawrence Carin. Learning deep sigmoid belief networks with data augmentation. In *Artificial Intelligence and Statistics*, pages 268–276, 2015.
- [25] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.
- [26] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1995–2003, 2016.
- [27] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.
- [28] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.
- [29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [30] François Chaumette, Seth Hutchinson, and Peter Corke. Visual servoing. In *Springer Handbook of Robotics*, pages 841–866. Springer, 2016.
- [31] Michael Paton, Kirk MacTavish, Chris J Ostafew, and Timothy D Barfoot. It’s not easy seeing green: Lighting-resistant stereo Visual Teach & Repeat using color-constant images. In *IEEE International Conference on Robotics and Automation*, pages 1519–1526. IEEE, 2015.
- [32] Aleksandar Kodzhabashev and Michael Mangan. Route following without scanning. In *Conference on Biomimetic and Biohybrid Systems*, pages 199–210. Springer, 2015.
- [33] Barbara Webb and Antoine Wystrach. Neural mechanisms of insect navigation. *Current Opinion in Insect Science*, 15:27–39, 2016.
- [34] H Freyja Ólafsdóttir, Francis Carpenter, and Caswell Barry. Coordinated grid and place cell replay during rest. *Nature Neuroscience*, 19(6):792–794, 2016.
- [35] Brigitta Gundersen. Forming artificial memories during sleep. *Nature Neuroscience*, 18(4):483–483, 2015.