RAPID TASK-SOLVING IN NOVEL ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When thrust into an unfamiliar environment and charged with solving a series of tasks, an effective agent should (1) leverage prior knowledge to solve its current task while (2) efficiently exploring to gather knowledge for use in future tasks, and then (3) *plan* using that knowledge when faced with new tasks in that same environment. We introduce two domains for conducting research on this challenge, and find that state-of-the-art deep reinforcement learning (RL) agents fail to plan in novel environments. We develop a recursive implicit planning module that operates over episodic memories, and show that the resulting deep-RL agent is able to explore and plan in novel environments, outperforming the nearest baseline by factors of 2-3 across the two domains. We find evidence that our module (1) learned to execute a sensible information-propagating algorithm and (2) generalizes to situations beyond its training experience.

1 INTRODUCTION

A truly general-purpose AI should be able to enter a new environment and get to work immediately, leveraging general world knowledge to (1) solve its current task and (2) efficiently explore to gather knowledge for use during future tasks in the same environment. After a small amount of experience, the agent should be able to (3) solve new tasks on the first try by *planning* using knowledge gathered during earlier tasks in the same environment. Consider for example the challenge faced by a household robot solving its first task in its new home: let's say, cleaning the bathroom. Ideally the robot would use general knowledge about household layouts to find the bathroom and cleaning supplies, efficiently exploring when necessary to complete the task. As it carries out the task, it should gather information it anticipates will be useful in later tasks, looking for example to see where the clothes hampers are in the rooms it passes. When faced with its next task, say, doing the laundry, it should use its newfound knowledge of the home's hamper locations to plan a solution that solves this second task more quickly than it would have had it not done the first.

Humans make this kind of *rapid task-solving in novel environments* (RTS) look easy (Lake et al., 2017), but as yet it remains an aspiration for AI. Two recent developments suggest progress can be made. First, recent work has shown that deep reinforcement learning (RL) can *learn* general knowledge useful for solving new tasks quickly (Wang et al., 2017; Duan et al., 2016; Finn et al., 2017). These methods have so far been used only to solve one task in each environment, so that environment knowledge need not be stored and then re-purposed for planning. This is understandable – solving a lengthy sequence of complex tasks requires memory and planning capacities unlikely to emerge from the neural network architectures typically used in these systems (e.g., LSTMs). The second important development may address the memory limitation: recent work has shown that episodic memory systems for deep-RL agents can perform well in a variety of memory-intensive tasks (Oh et al., 2016; Wayne et al., 2018; Ritter et al., 2018b; Fortunato et al., 2019).

We develop a minimal diagnostic test for the ability to explore and plan in novel environments, and find that state-of-the-art episodic memory deep-RL agents are able to explore but not to plan. We hypothesize that this is due to insufficiently powerful mechanisms for planning using retrieved memories, and we develop a novel neural network module for this purpose. We find that the resulting agent explores and plans effectively in the diagnostic domain as well as in One-Shot StreetLearn, a scaled-up RTS domain. We analyze the learned algorithms to discover desirable interpretability and generalization properties.

In summary, our contributions are to:

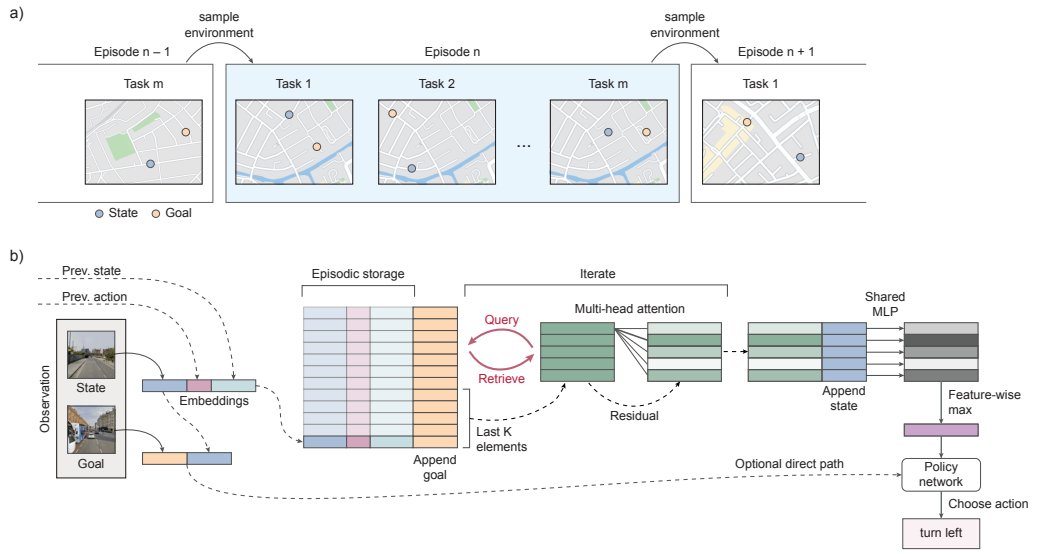

Figure 1: (a) *Rapid Task Solving in Novel Environments* (RTS) setup. A new environment is sampled in every episode. Each episode consists of a sequence of tasks which are defined by sampling a new goal state and a new initial state. The agent has a fixed number of steps per episode to complete as many tasks as possible. (b) The Episodic Planning Network (EPN). The EPN uses multiple iterations of a single self-attention function (shared across iterations) over memories retrieved from an episodic storage.

1. Develop two domains – the minimal and highly interpretable Memory&Planning Game and the richer, scaled-up One-Shot StreetLearn - for studying planning in novel environments. We make these available for public use.
2. Show that previous agents are severely limited in their ability to plan using their episodic memory stores.
3. Design a new architecture – the Episodic Planning Network (EPN) – that is capable of exploring and planning in both domains, widely outperforming prior agents. We find that EPNs generalize to larger problems than those seen in training.
4. Analyze our agent's behavior and internal activations to find evidence that EPNs learn to execute an iterative planning algorithm that propagates information about state-state connectivity outward from the target state.

2 Connections to Past Work

Deriving repurposable environment knowledge to solve arbitrary tasks has been a long-standing goal in the field of RL (Foster & Dayan, 2002; Dayan, 1993) toward which progress has recently been made (Sutton et al., 2011; Schaul et al., 2015; Borsa et al., 2018). While quick to accommodate new tasks, these approaches require large amounts of time and experience to learn about new environments. This can be attributed to their reliance on gradient descent for encoding knowledge and their inability to bring prior knowledge to new environments This is still a bit of an uncomfortable concept for me, since I imagine NN's as very easily bringing prior knowledge to new environments, via their learned weights (Botvinick et al., 2019). Our approach overcomes these problems by recording information about new environments in activations instead of weights and learning prior knowledge in the form of exploration and planning policies.

The possibility of learning to explore with recurrent neural networks was demonstrated by Wang et al. (2017) and Duan et al. (2016) in the context of bandit problems. Wang et al. (2017) further demonstrated that an LSTM could learn to display the behavioral hallmarks of model-based control – in other words, *learn to plan* – in a minimal task widely used for studying planning in humans Daw et al. (2011). The effectiveness of learned, implicit planning was demonstrated further in fully-observed spatial environments using grid-structured memory and convolutional update functions

(Tamar et al., 2016; Lee et al., 2018b; Guez et al., 2019). Gupta et al. (2017) extended this approach to partially observed environments with ground-truth ego-motion by training agents to emulate an expert that computes shortest paths. Our present work can be seen as extending learned planning by (1) solving partially-observed environments by learning to gather the information needed for planning, and (2) providing an architecture appropriate for a broader range of (e.g. non-spatial) planning problems.

Learning to plan can be seen as a specific case of learning to solve combinatorial optimization problems, a notion that has been taken up by recent work (for review see Bengio et al., 2018). Especially relevant to our work is Kool et al (2018), who show that transformers can learn to solve large combinatorial optimization problems, comparing favorably with more resource-intensive industrial solvers. This result suggests that in higher-depth novel-environment planning problems than we consider in our current experiments, the transformer-based architecture may continue to be effective.

Savinov et al. (2018) develop an alternative to end-to-end learned planning that learns a distance metric over observations for use with a classical planning algorithm. Instead of learning to explore, this system relies on expert trajectories and random walks. It's worth noting that hand-written planning algorithms do not provide the benefits of domain-adapted planning demonstrated by Kool et al. (2018), and that design modifications would be needed to extend this approach to tasks requiring *abstract planning* - e.g. jumpy planning and planning over belief states - whereas the end-to-end learning approach can be applied to such problems out-of-the-box.

Recent work has shown episodic memory to be effective in extending the capabilities of deep-RL agents to memory intensive tasks (Oh et al., 2016; Wayne et al., 2018; Ritter et al., 2018b; Fortunato et al., 2019). We chose episodic memory because of the following desirable properties. First, because the episodic store is non-parametric, it can grow arbitrarily with the complexity of the environment, and approximate k-nearest neighbors method make it possible to scale to massive episodic memories in practice, as in Pritzel et al. (2017). This means that memory fidelity need not decay with time. Second, episodic memory imposes no particular assumptions about the environment's stucture, making it a potentially appropriate choice for a variety of non-spatial applications such as chemical synthesis Segler et al. (2018) and web navigation Gur et al. (2018), as well as abstract planning problems of interest in AI research, such as Gregor et al.'s (2019) Voxel environment tasks.

Our approach can be seen as implementing episodic model-based control (EMBC), a concept recently developed in cognitive neuroscience (Vikbladh et al., 2017; Ritter, 2019). While episodic control (Gershman & Daw, 2017; Lengyel & Dayan, 2008; Blundell et al., 2016) produces value estimates using memories of individual experiences in a model-free manner, EMBC uses episodic memories to inform a model that predicts the outcomes of actions. Ritter et al. (2018a) showed that a deep-RL agent with episodic memory and a minimal learned planning module (an MLP) could learn to produce behavior consistent with EMBC (Vikbladh et al., 2017). Our current work can be seen as using iterated self-attention to scale EMBC to much larger implicit models than the MLPs of past work could support.

3 Problem Formulation

Our objective is to build agents that can maximize reward over a sequence of tasks in a novel environment. Our basic approach is to have agents *learn* to do this through exposure to distributions over multi-task environments. To define such a distribution, we first formalize the notion of an environment e as a 4-tuple $(S, A, P_a, \mathcal{R})$ consisting of states, actions, a state-action transition function, and a *distribution over* reward functions. We then define the notion of a task t in environment e as a Markov decision process (MDP) (S, A, P_a, r) that results from sampling a reward function r from $\mathcal{R}$.

We can now define a framework for learning to solve tasks in novel environments as a simple generalization of the popular meta-RL framework (Wang et al., 2017; Duan et al., 2016). In meta-RL, the agent is trained on MDPs (S, A, P_a, r) sampled from a task distribution $\mathcal{D}$. In the rapid task-solving in novel environments (RTS) paradigm, we instead sample problems by first sampling an environment e from an environment distribution E, then sequentially sampling tasks, i.e. MDPs, from that environment's reward function distribution (see Figure 1).

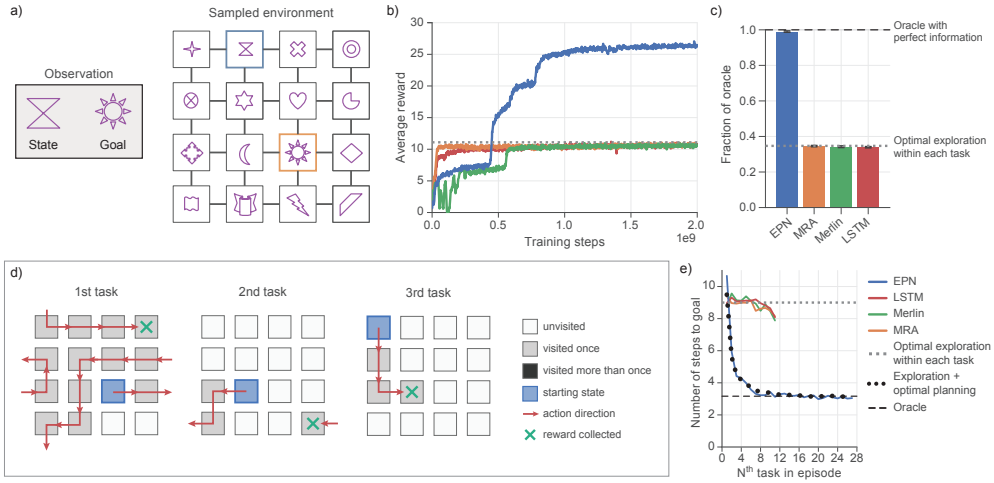

Figure 2: Memory&Planning Game. (a) Example 4×4 environment (not observable by the agent) and state-goal observation. (b) Training curves. Performance measured by the average reward per episode, which corresponds to the average number of tasks completed (showing the best runs from a large hyper-parameter sweep for each model). (c) Performance measured in the last third of the episodes (post-training), relative to an oracle with perfect information that takes the shortest path to the goal. (d) Example trajectory of a trained EPN agent in the first three tasks of an episode. In the first task, the agent explores optimally without repeating states. In the subsequent tasks, the agent takes the shortest path to the goal. (e) Number of steps taken by an agent before completing the nth task of an episode.

An agent can be trained for RTS by maximizing the following objective:

$$\mathbb{E}_{e \sim E}[\mathbb{E}_{r \sim \mathcal{R}_e}[J_{e,r}(\theta)]],$$

where $J_{e,r}$ is the expected reward in environment e with reward function r. When there is only one reward function per environment, the inner expectation disappears and we recover the usual meta-RL objective $\mathbb{E}_{e \sim E}[J_e(\theta)]$. RTS can be viewed as meta-RL with the added complication that the reward function changes within the inner learning loop.

While meta-RL formalizes the problem of learning to learn, RTS formalizes both the problems of learning to learn and learning to plan. To maximize reward while the reward function is constantly changing in non-trivial novel environments, agents must learn to (1) efficiently explore and effectively encode the information discovered during exploration (i.e., *learn*) and (2) use that encoded knowledge to select actions by predicting trajectories it has never experienced (i.e., *plan*[1]).

4 The Limitations of Prior Agents

To test whether past deep-RL agents can explore and plan in novel environments, we introduce a minimal problem that isolates the challenges of (1) exploring and remembering the dynamics of novel environments and (2) planning over those memories. The problem we propose is a simple variation of the well-known Memory Game[2], wherein players must remember the locations of cards in a grid. The variation we propose, which we call the Memory&Planning Game, extends the challenge to require planning as well as remembering (see Figure 2).

In the Memory&Planning Game, the agent occupies an environment consisting of a grid of symbols. The agent can see one symbol which corresponds to the agent's current location, and it sees a "goal" symbol, which it is tasked with navigating to. The agent can *not* see its relative location with respect to other symbols in the grid. The agent's actions are: move left, move right, move up, move down,

[1]Following Sutton & Barto (1998), we refer to the ability to choose actions based on predictions of not-yet-experienced events as "planning".

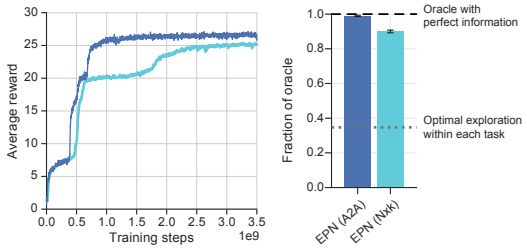

Figure 3: Comparison between architecture variants. The Nxk architecture variant, which scales linearly with the total number of memories, recovers 92% of the performance of the A2A variant, which scales quadratically.

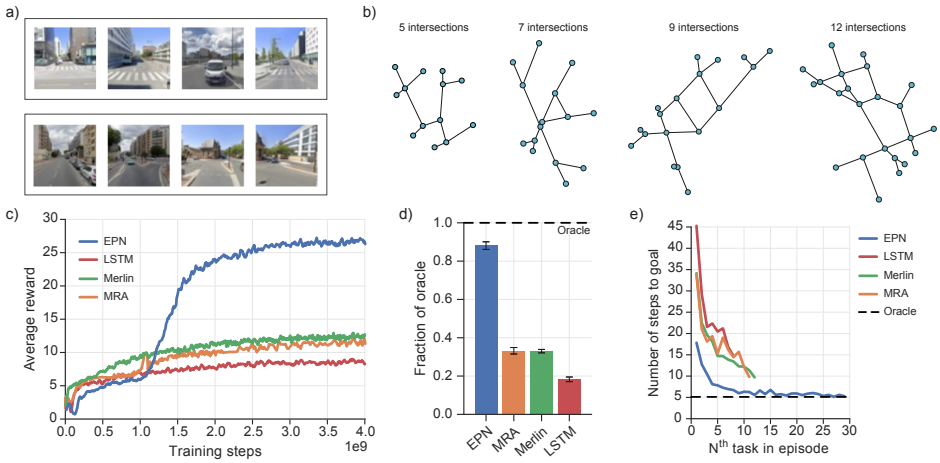

Figure 4: One-Shot StreetLearn. (a) Four example states from two randomly sampled neighborhoods. (b) Example connectivity graphs. (c) Evaluation performance measured on neighborhoods of a held-out city throughout the course of training (showing the best run from a large hyper-parameter sweep for each model). (d) Performance relative to an oracle with perfect information. (e) Number of steps taken by an agent before reaching the nth goal in an episode.

and choose. The agent receives reward when it takes the "choose" action while its current location matches the goal location. At the beginning of each episode, a new set of symbols is sampled, effectively inducing a new transition function. Each time the agent finds a goal – which corresponds to completing a task –, a new goal is sampled in and the transition function stays fixed.

A successful agent will (1) efficiently explore the grid to discover the current set of symbols and their connectivity and (2) plan shortest paths to goal symbol if it has seen (or can infer) all of the transitions it needs to connect its current location and the current goal location. This setup supplies a minimal case of the RTS problem: at the beginning of each episode the agent faces a new transition structure (S, A, P_a) and must solve the current task while finding information that will be useful for solving future tasks on that same transition structure. In subsequent tasks, the agent must use its stored knowledge of the current grid's connectivity to plan shortest paths.

In this symbolic domain, we find evidence that previous agents learn to explore but not to plan. Specifically, they match the within-trial planning optimum; that is, a strategy that explores optimally within each task, but forgets everything about the current environment the task ends (Figure 2). We hypothesize that this failure is the result of the limited expressiveness of past architectures' mechanisms for processing retrieved memories. The following architecture overcomes this limitation.

[2]In an interesting parallel, the Memory Game played a key role in the development of memory abilities in early episodic memory deep-RL agents (Wayne et al., 2018).

5 EPISODIC PLANNING NETWORKS

Past episodic memory agents used their memory stores by querying the memories, summing the retrieved slots, then projecting the result through a multi-layered perceptron (MLP) (Fortunato et al., 2019; Wayne et al., 2018). We hypothesize that these agents fail to plan because the weighted sum of retrieved slots is not sufficiently *representationally* expressive, and the MLP is not sufficiently *computationally* expressive to support planning in non-trivial environments. To test this hypothesis, we replace the weighted sum and MLP with an an iterative self-attention-based architecture designed to support implicit planning.

Implicit planning has recently been studied using recurrent convolutional functions that learn to plan by iteratively updating working memory representations of the environment (Lee et al., 2018b; Tamar et al., 2016; Guez et al., 2019). The key idea behind this approach was that the update function could learn to implement algorithms akin to value iteration, computing state-state reachability by propagating value through the grid-structured environment representation.

We would like to plan in a similar manner over the environment information represented in episodic memory, in order to support arbitrarily-structured environments and get the benefits of non-parametric storage. We accomplished this by replacing the convolutions of past work with self-attention, a powerful method for computing relationships among an arbitrary number of items which does not assume any particular structure among them (Vaswani et al., 2017). The following self-attention architecture iteratively computes relationships among slots in episodic memory, channeling the intuition that this architecture can learn to execute value-iteration like computations over representations of the environment's structure stored in episodic memory.

The architecture, which we call the Episodic Planning Network (EPN, see Figure 1b), begins with a set M of episodic memories which reflect the agent's experience in the episode so far. In the present experiments, the agent stores a memory in a new slot on each timestep. The memory represents that timestep's transition: it is the concatenation of embeddings of the current observation, the previous action, and the previous observation. The EPN appends an embedding of the current goal g to each slot in M, producing M^g. Then, a self-attention-based update function S is applied to produce a processed representation reflecting the agent's belief about the environment B_1^g. S is iterated some number of times u (sharing weights) to produce B_u^g. Finally, the current state c is appended to each slot in B_u^g to produce B_u^{gc}, and then each resulting slot is passed through a shared MLP. The resulting vectors are aggregated by a feature-wise max operation, then fed to the policy network to influence behavior.

Note that the self-attention network does not have access to the current state until the very end – it does all of the self-attention updates with access only to the target and episodic memories. This design reflects the intuition that the function S might learn to compute something like a value map, which represents the distance from all states in memory to the target. If the final self-attention state B_u^g does in fact come to represent such a reachability map, then the simple MLP over B_u^{gc} should be sufficient to compute the action from c that leads to the nearest/highest value state with respect g. We find in Section 6.2, Figure 5 evidence that B^g does in fact come to resemble an iteratively improved reachability map.

We experiment with two variants of the update function S. In the first, all of the episodic memories attend to all of the others to produce B^g with same dimensionality as M. This variant, which we refer to as all-to-all (A2A), scales quadratically with the number of memories (and the number of steps per episode). The second version, which scales more favorably, takes the k most recent memories as the initial belief state B_0^g. On each iteration of S, each slot in B_i^g attends to each slot in M^g, producing $|B^g|$ vectors (i.e. k vectors), which then self-attend to one another to produce the next belief state B_{i+1}^g. The idea behind this design is that the query from the k vectors to the full memory can learn to select and summarize information to be composed via self-attention. The self-attention among the k vectors may then learn to compute relationships, e.g. state-state reachability, using this condensed representation. The benefit is that this architecture scales only linearly with the size of the full memory N, because the quadratic time self-attention need only be applied over the k vectors. This approach is similar to the inducing points of Lee et al. (2018a). Because this scales with $N * k$ rather than N^2, we call this variant N-by-k (abbreviated, Nxk).

The specific self-attention function we used was crucial for full performance. In all references above to self-attention, the specific update function we used was:

$$B_{i+1} = f(B_i + \phi(B_i))$$
$$\phi(x) = \text{MHA}(\text{LayerNorm}(x))$$

where MHA is the multi-head dot-product attention described by Vaswani et al. (2017). In our experiments, f was a ReLU followed by a 2-layer MLP shared across rows. We trained all agents to optimize an actor-critic objective function using IMPALA, a framework for distributed RL training (Espeholt et al., 2018). For further details, please refer to the Supplemental Material.

Our results show that the EPN succeeds in both exploration and planning in the Memory&Planning Game. It matches the shortest-path optimum after a few tasks (Figure 2c), and its exploration and planning performance closely matches the performance of a hand-coded agent which combines a strong exploration strategy – avoiding visiting the same state twice – with an optimal planning policy (Figure 2d,e). Figure 3 compares the performance obtained with the two architecture variants, A2A and Nxk. With k set to one half of the original memory capacity ($k = 50$), the Nxk agent recovers 92% of the performance of the A2A. This result indicates that we can indeed overcome the quadratic complexity and opens up the possibility of applying EPNs to problems with longer timescales.

6 ONE-SHOT STREETLEARN

We now test whether our agent can extend its success in the Memory&Planning Game to a domain with high-dimensional pixel-based state observations, varied real-world state-connectivity, longer planning depths, and larger time scales. We introduce the One-Shot StreetLearn domain (see Figure 1a), wherein environments are sampled as neighborhoods from the StreetLearn dataset of Google StreetView images and their connectivity (Mirowski et al., 2019). Tasks are then sampled by selecting a position and orientation that the agent must navigate to from its current location.

In past work with StreetLearn, agents were faced with a single large map (e.g. a city), in which they could learn to navigate over billions of experiences (Mirowski et al., 2018). StreetLearn city maps are fully connected graphs where nodes are generated from point-wise samples every few meters or so in the given city. For each node and neighbour pair there is a 64×64 RGB pixel input unique to that point-wise sample and direction of gaze - this pixel representation is the observation state of the environment. We partition the StreetLearn data into a many different maps (e.g. neighborhoods of different cities), so that agents can be trained to rapidly solve navigation tasks in new maps. This approach is analogous to the partitioning of the ImageNet data into many one-shot classification tasks that spurred advances in one-shot learning (Vinyals et al., 2016). In this case, rather than learning to classify in one shot, agents should learn to *plan* in one shot, that is, to plan after one or fewer observations of each transition in the environment. Code for generating "One-Shot StreetLearn" tasks, and a subset of the data we used will be made publicly available.

In One-Shot StreetLearn, the agent's observations consist of an image representing the current state and an image representing the goal state (Figure 1). The agent receives no other information from the environment. The agent's actions are "turn left", which orients the agent toward the next available direction of motion from its current node; "turn right", which does the same in the other direction; and "move forward", which moves the agent along the direction it's facing to the next available node.

One-Shot StreetLearn has several important characteristics for developing novel-environment planning agents. First, we can generate many highly varied environments for training by sampling neighborhoods from potentially any place in the world. This allows us to generate virtually unlimited maps with real-world connectivity structure for agents to learn to take advantage of. Second, the agent's observations are rich, high-dimensional visual inputs that simulate a person's experience when navigating. These inputs can in principle be used by an agent to infer some of the structure of a new environment. Third, we can scale the planning depth arbitrarily by changing the size of the sampled neighborhoods – a useful feature for iteratively developing increasing capable planning systems. Fourth, the time-scale of planning can be varied independently of the planning depth by varying the number of nodes between intersections. This is a useful feature for developing temporally abstract (i.e. jumpy) planning abilities (see Section 7 for discussion).

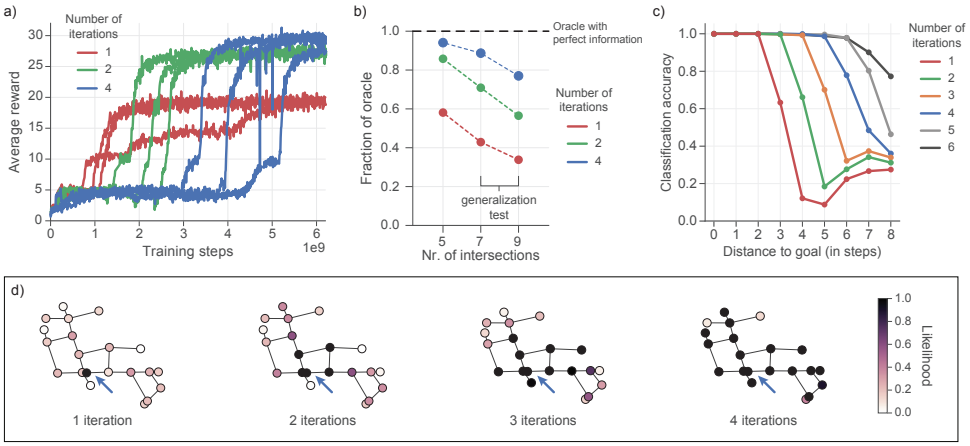

Figure 5: Iteration analysis and generalization. (a) Evaluation performance of an EPN agent with a planner using 1, 2 and 4 self-attention iterations (showing 3 runs for each condition). (b) Performance evaluation on neighborhoods larger (7 and 9 iterations) than the ones used during training (5 iterations). (c) Distance-to-goal accuracy of six decoders which had access to the output of the planner after 1–6 iterations. (d) The ability of EPN activations to predict state-distance expands out from the target (blue arrow) as the number of self-attention iterations increases. See section 6.2 for details.

6.1 Results

We trained EPN-based agents and previous deep-RL agents on One-Shot StreetLearn neighborhoods with 5 intersections. In order to reduce the exploration problem while keeping the planning difficulty constant, we removed all the locations between intersections that corresponded to degree-2 nodes. This resulted in neighborhoods with a median of 12 nodes (see example connectivity graphs in Figure 4b). Our agent significantly outperformed the baselines, successfully reaching 28.7 goals per episode (averaged over 100 consecutive episodes, Figure 4c). It is important to note that this performance was measured on neighborhoods of a held-out city, which the agent was visiting for the first time. Baseline agents with an episodic memory system (Merlin and MRA) did not exceed 14.5 goals per episode, performing better than ones without (LSTM) that were only able to reach 10.0.

The EPN performance approached optimal planning ($88\% \pm 2\%$), as measured by the average reward obtained in the last third of every episode relative to an oracle with perfect information that systematically takes the shortest path to the goal. Plotting the average number of steps taken by the agent before completing the nth task of an episode reveals that the agents needs fewer and fewer steps to complete a new task, matching the minimum number of steps required to collect each goal after 15–20 tasks (Figure 4d,e).

6.2 Iteration Analysis and Generalization

In this section, we test whether the planner has learned a general, iterative planning function – general, in that it is effective in problems larger than the ones it was trained on; and iterative in that performance increases as the number of iterations (i.e., "thinking time") increases.

To test this hypothesis, we start by looking at the effect of number of self-attention iterations used during training on the agent's performance at evaluation time.[3] Training on problems with a fixed planning depth (neighborhoods with 5 intersections), we observed a systematic performance boost with each additional iteration used in training (see Figure 5a), with a stronger effect for lower numbers of iterations. We then evaluated the trained agents, without further training, on larger maps while keeping fixed the number of agent steps per episode. We observed robust generalization – agents' performance degrades gracefully as the planning depth increases. For example, an agent trained with with 4 self-attention iterations on neighborhoods with 5 intersections achieved a per-

[3]The experiments described here were restricted to one-hot inputs for quicker turnaround.

formance of $77\% \pm 0.7\%$ relative to an oracle with perfect information when evaluated on neighborhoods with 9 intersections (Figure 5b). The effect of number of iterations is still present in this condition.

We then investigate how additional iterations contribute to improved performance. To do this we took a trained EPN agent, freezing its weights, and replaced its policy network with an MLP decoder of the same size. In a supervised classification set up, we trained the decoder (with a stop gradient to the planner) to infer the distance from a random state to a random goal, while providing the planner with memories containing all the transitions of a randomly sampled environment. We repeated this experiment with six decoders each one receiving inputs from the output of the planner after a fixed number of iterations, from 1 to 6.[4]

The training loss of the different decoders revealed a steady increase in the ability to infer distance to goal as we increase the number of self-attention iterations from 1 to 6 (see Suppl. Material). This gain is a direct consequence of improved decoding for longer distances, as made evident by the gradual rightward shift in the drop-off of the classification accuracy plotted against distance to goal (Figure 5c). This holds true even when we increase the number of iterations beyond the number of iterations used during the training of the planner. Altogether, the results suggest that the planner is able to "look" further ahead when given more iterations.

We can visualize this result, and its spatial implications, by selecting a single evaluation environment – in this case, a neighborhood with 12 intersections – with a fixed goal location, and measuring the likelihood resulting from all possible states in that environment. This manipulation reveals a spatial pattern in the ability to infer distance to goal, spreading radially from the goal location as we increase the number of self-attention iterations (see Figure 5e and Suppl. Material).

7 Discussion

We considered the problem of maximizing reward over sequences of tasks in novel environments, and provided two domains for researching this problem. We demonstrated that past deep-RL agents fail even in minimal cases of the problem class, then showed that our new architecture, EPNs, succeed at rapid task planning. We found evidence that EPNs learn an iterative information-propagating algorithm that generalizes well both to larger maps and more iterations than those used in training. The discovery of this effective and general planning module paves the way for important future work.

In the current experiments, the agent could succeed by planning over observed states, i.e. states of the environment MDP represented by e.g. StreetLearn images. However, there is nothing preventing EPNs from being used to plan over *belief states*, a potential critical ability for operating in dynamic partially-observed environments. Planning over belief states might be accomplished simply by storing belief states such as those developed by Gregor et al. (2019) and training on a problem distribution that requires planning over information stored in those states.

Another important direction is temporal abstraction, i.e., jumpy planning (Akilesh et al., 2019). This is an extremely important capability for acting in temporally extended environments like the real world. Unlike other prominent approaches to learning to plan (Schrittwieser et al., 2019; Racanière et al., 2017), our method makes no assumptions about the timescale of planning. Training and evaluating EPNs on problems that benefit from jumpy planning, such as One-Shot StreetLearn with all intermediate nodes, may be enough to obtain strong temporal abstraction performance.

The tasks in our current experiments are drawn from a very narrow task class; i.e., they are all solved by occupying a particular observable state. In contrast, humans are able to solve a seemingly *open-ended* variety of tasks in the real world, as highlighted for example by Lake and colleagues' (2017) Frostbite challenge. Future work may approach RTS problems with broader task distributions, e.g., by using generative grammars over tasks or by taking human input (Fu et al., 2019). The EPN architecture places minimal constraints on the class of tasks it can be trained on - it is suitable for any task class wherein the reward function can be represented by a vector. Future work may test the extent to which EPNs are effective in solving broader classes of tasks.

[4]Note that while the decoders were trained in a supervised setting using 1 to 6 iterations, the weights of the planner were trained once in the RL setting using 4 iterations.

REFERENCES

B Akilesh, Suriya Singh, Anirudh Goyal, Alexander Neitz, and Aaron Courville. Toward jumpy planning. In *International Conference on Machine Learning*, 2019.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *arXiv preprint arXiv:1811.06128*, 2018.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. *arXiv preprint arXiv:1812.07626*, 2018.

Mathew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 2019.

Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl˜2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Meire Fortunato, Melissa Tan, Ryan Faulkner, Steven Hansen, Adrià Puigdomènech Badia, Gavin Buttimore, Charles Deck, Joel Z Leibo, and Charles Blundell. Generalization of reinforcement learners with working and episodic memory. In *Advances in Neural Information Processing Systems*, pp. 12448–12457, 2019.

David Foster and Peter Dayan. Structure in the space of value functions. *Machine Learning*, 49 (2-3):325–346, 2002.

Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

Samuel J Gershman and Nathaniel D Daw. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, 68:101–128, 2017.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. In *Advances in Neural Information Processing Systems*, pp. 13475–13487, 2019.

Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, et al. An investigation of model-free planning. *arXiv preprint arXiv:1901.03559*, 2019.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2017.

Izzeddin Gur, Ulrich Rückert, Aleksandra Faust, and Dilek Hakkani-Tür. Learning to navigate the web. *CoRR*, abs/1812.09195, 2018. URL `http://arxiv.org/abs/1812.09195`.

Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. *arXiv preprint arXiv:1810.00825*, 2018a.

Lisa Lee, Emilio Parisotto, Devendra Singh Chaplot, Eric Xing, and Ruslan Salakhutdinov. Gated path planning networks. *arXiv preprint arXiv:1806.06408*, 2018b.

Máté Lengyel and Peter Dayan. Hippocampal contributions to control: the third way. In *Advances in neural information processing systems*, pp. 889–896, 2008.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, pp. 2419–2430, 2018.

Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.

Junhyuk Oh, Valliappa Chockalingam, Satinder Singh, and Honglak Lee. Control of memory, active perception, and action in minecraft. *arXiv preprint arXiv:1605.09128*, 2016.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2827–2836. JMLR. org, 2017.

Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomenech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. Imagination-augmented agents for deep reinforcement learning. In *Advances in neural information processing systems*, pp. 5690–5701, 2017.

Samuel Ritter. *Meta-Reinforcement Learning with Episodic Recall: An Integrative Theory of Reward-Driven Learning*. PhD thesis, Princeton University, 2019.

Samuel Ritter, Jane X Wang, Zeb Kurth-Nelson, and Matthew M Botvinick. Episodic control as meta-reinforcement learning. *bioRxiv*, pp. 360537, 2018a.

Samuel Ritter, Jane X Wang, Zeb Kurth-Nelson, Siddhant M Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. Been there, done that: Meta-learning with episodic recall. *arXiv preprint arXiv:1805.09692*, 2018b.

Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604, 2018.

Richard S Sutton and A Barto. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge, 1998.

Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 761–768. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems*, pp. 2154–2162, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Oliver Vikbladh, Daphna Shohamy, and Nathaniel Daw. Episodic contributions to model-based reinforcement learning. In *Annual conference on cognitive computational neuroscience, CCN*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

JX Wang, Z Kurth-Nelson, D Tirumala, H Soyer, JZ Leibo, R Munos, C Blundell, D Kumaran, and M Botivnick. Learning to reinforcement learn. arxiv 1611.05763, 2017.

Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z Leibo, Adam Santoro, et al. Unsupervised predictive memory in a goal-directed agent. *arXiv preprint arXiv:1803.10760*, 2018.

A APPENDIX

A.1 ONE-SHOT STREETLEARN

Dataset preparation – The StreetLearn dataset Mirowski et al. (2019) provides a connectivity graph between nodes, and panoramic images (panos) for each node. As movement in the environment is limited to turning left, turning right, or moving forward to the next node, we only need from the panoramic images the frames corresponding to what an observer would see in a given location, when oriented towards a next possible location. This defines a mapping from oriented edge in the full StreetLearn graph to a single frame. We compute it ahead of time and store it in a data structure enabling efficient random accesses.

Observation associated to an edge – As our agent operates on a reduced graph containing only degree-2 nodes, we define the observation to an oriented edge A–B in the reduced graph to be the image associated to the first edge in the shortest path from A to B on the full graph.

Movement – When the agent moves forward, its new location becomes the neighbor it was oriented towards, and is reoriented towards the neighbor most compatible with its current orientation.

Distributed learning setup –

Each actor is assigned to a random city, out of 12 cities/regions in Europe: Amsterdam, Brussels, Dublin, Lisbon, Madrid, Moscow, Rome, Vienna, Warsaw, Paris North-East, Paris North-West and Paris South-West. A special *evaluator* actor, which does not send send trajectories to the learner, is assigned to a withheld region: Paris South-East.

Sampling neighborhoods – In the beginning of a new episode, we start by sampling a random location in the region assigned to the actor – i.e. a random node of the region's connectivity graph. This node will become the center-node of the new neighborhood and is added as the first node of the sampled graph. We then proceed to traverse the connectivity graph, breadth first, adding visited nodes to the sampled graph, until it contains the number of intersections required. An intersection is a node with degree greater than 2; the number of intersections is a parameter of environment that

we can set for each experiment and it determines the difficulty (depth) of the planning problem. In our experiments we decided to remove all degree-2 nodes of the sampled graph. This manipulation allowed us to simplify the exploration problem substantially without reducing the planning difficulty.

A.2 RL SETUP

We used an actor-critic setup for all RL experiments reported in this paper, following the distributed training and V-Trace algorithm implementations described by Espeholt et al. (2018). The distributed agent consisted of 1000 actors that produced trajectories of experience on CPU, and a single learner running on a Tensor Processing Unit (TPU), which learned a policy π and a baseline V using mini-batches of actors' experiences provided via a queue. The length of the actors' trajectories is set to the unroll length of the learner. Training was done using the RMSprop optimization algorithm. Please see the table below for values of fixed hyperparameters and intervals used for hyperparameter tuning.

Hyperparameter	**Values**
Agent	
Mini-batch size	[32, 128]
Unroll length	[10, 40]
Entropy cost	$[1\mathrm{e}{-3}, 1\mathrm{e}{-2}]$
Discount γ	$[0.9, 0.95]$
RMSprop	
Learning rate	$[1\mathrm{e}{-5}, 4\mathrm{e}{-4}]$
Epsilon ϵ	$1\mathrm{e}{-4}$
Momentum	0
Decay	0.99

Table 1: Hyperparameter values and tuning intervals used in RL experiments.

A.3 ARCHITECTURE DETAILS

Vision – The visual system of our agents, which produced embeddings for state and goal from the raw pixel inputs, was identical to the one described in Espeholt et al. (2018), except ours was smaller. It comprised 3 residual-convolutional layers, each one with a single residual block, instead of two. The number of output channels in each layer was 16, 32 and 32. We used a smaller final linear layer with 64 units.

Planner – For the multi-head attention, queries, keys and values were produced with an embedding size of 64, using 1 to 4 attention heads. In our experiments, we did not observe a significant benefit from using more than a single attention head. The feedforward block of each attention step was a 2-layer MLP with 64 units per layer (shared row-wise). Both the self-attention block and the feedforward block were shared across iterations. After appending the state to the output of the final attention iteration, we used another MLP (shared row-wise) consisting of 2-layers with 64 units. The output of this MLP applied to each row was then aggregated using a max pooling, feature-wise, operation.

Policy network – The input to the policy network was the result of concatenating the output of the planner and the state-goal embedding pair (which can be seen as a skip connection). We used a 2-layer MLP with 64 units per layer followed by a ReLU and two separate linear layers to produced the policy logits (π) and the baseline (V).

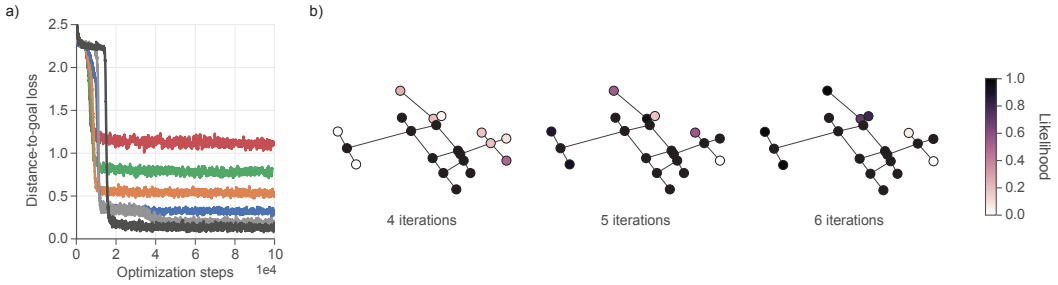

Figure 6: (a) Distance-to-goal training loss of six decoders with access to the output of the planner (with frozen weights) after 1–6 iterations. Each additional iteration improves distance decoding. (b) The ability of EPNs to infer distance to goal expands with the number of self-attention iterations, even when that number goes beyond the number of iterations used during training of the EPN. See Section 6.2 for more details.