

INTROSPECTIVE ACTION ADVISING FOR INTERPRETABLE TRANSFER LEARNING

Joseph Campbell*, Yue Guo, Fiona Xie, Simon Stepputtis, Katia Sycara
Robotics Institute, Carnegie Mellon University

ABSTRACT

Transfer learning can be applied in deep reinforcement learning to accelerate the training of a policy in a target task by transferring knowledge from a policy learned in a related source task. This is commonly achieved by copying pretrained weights from the source policy to the target policy prior to training, under the constraint that they use the same model architecture. However, not only does this require a robust representation learned over a wide distribution of states – often failing to transfer between specialist models trained over single tasks – but it is largely uninterpretable and provides little indication of what knowledge is transferred. In this work, we propose an alternative approach to transfer learning between tasks based on action advising, in which a teacher trained in a source task actively guides a student’s exploration in a target task. Through introspection, the teacher is capable of identifying when advice is beneficial to the student and should be given, and when it is not. Our approach allows knowledge transfer between policies agnostic of the underlying representations, and we empirically show that this leads to improved convergence rates in Gridworld and Atari environments while providing insight into what knowledge is transferred.

1 INTRODUCTION

Reinforcement learning (RL) has been used to successfully solve tasks that were once considered to be computationally intractable, in domains such as games (Silver et al., 2018; Kaiser et al., 2019; Lample & Chaplot, 2017; Mnih et al., 2013), robotics (Gu et al., 2017; Nguyen & La, 2019; Ibarz et al., 2021; Niroui et al., 2019), and health care (Yu et al., 2021; Gottesman et al., 2019; Coronato et al., 2020). However, state-of-the-art RL algorithms utilizing deep neural network models as function approximators are notoriously sample inefficient (Yarats et al., 2021; Yu, 2018; Lillicrap et al., 2015), requiring potentially many millions of environment samples in complex environments. This limits their use in applications where the environment is costly to sample from, particularly when the samples come from the real world as in robotics (Dulac-Arnold et al., 2019; Zhu et al., 2020b; Campbell & Yamane, 2020). Transfer learning (Zhao et al., 2020; Tan et al., 2018; Zhuang et al., 2020; Da Silva et al., 2020) is a technique commonly employed to help alleviate this issue, by transferring knowledge from a previous model that has been trained in a related task (source model) to a new, untrained model in the target task (target model), with the goal of accelerating learning.

We consider two general approaches to transfer learning with deep neural networks: 1) the source model is used to generate an intermediate representation which is directly provided as input to the target model, common in language (Shor et al., 2022; Dalvi et al., 2019; Stepputtis et al., 2020) and vision (Evcı et al., 2022; Lee et al., 2018) domains utilizing features from large pre-trained models; and 2) the source model is *fine-tuned* (Yosinski et al., 2014) wherein a *subset*¹ of the source model’s layers are trained in the target task while the rest of the layers are frozen. In both of these cases, however, knowledge transfer is performed prior to training the target model: the mechanism is simply to use the pre-trained weights of an existing source model, under the constraint that the architecture remains fixed between the source and target models. This has significant limitations, since it is not clear, a) what information learned by the source model is useful in the target task, and b) where this information is encoded in the source model, i.e., which layers should be frozen in the case of fine-tuning.

The result is a largely empirical approach to transfer learning where practitioners often attempt to transfer different subsets of layers until they achieve satisfactory performance in the target task (Yosinski et al., 2014), with little insight into what was transferred and how it was used. Correspondingly, the most successful fine-tuned models are those with robust feature representations learned over a wide input distribution with vast amounts of data (Killian et al., 2017; Shafahi et al., 2019). Despite these issues, fine-tuning remains an exceedingly common approach to transfer learning

* Correspondence to jcampbell@cmu.edu

¹This is not a strict subset as all layers may be fine-tuned.

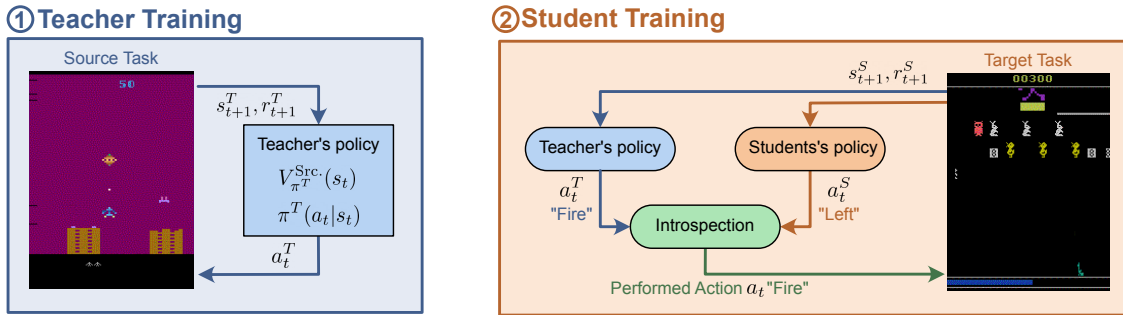


Figure 1: Overview of Introspective Action Advising: Given a trained policy π^T in a source task, we estimate if actions learned in this task are transferable and beneficial for the student policy π^S during training in a target task. Here, we show the transfer of action “Fire” between a source task (Air Raid), to our student in a target task (Carnival).

due to its simplicity – no specialized model architectures or algorithms are required and it may be applied to any existing policy. Thus there is a gap – while large models amenable to fine-tuning receive much attention, many models deployed in the real world are single-task specialist models trained to fulfill stringent accuracy requirements over small amounts of data. Consequently, it is often difficult to generalize task-specific features using fine-tuning-based methods.

In this work, we propose a new approach to transfer learning for deep reinforcement learning which occurs dynamically *during* the training process of the target model. Inspired by the teacher-student paradigm in action advising (Torrey & Taylor, 2013; Fachantidis et al., 2017; Zhu et al., 2020a; Guo et al., 2023; Omidshafiei et al., 2019), a teacher which has access to the source model selectively transfers advice in the form of state-action pairs to a student as they are training the target model, shown in Fig. 1. Intuitively, this serves as a form of guided exploration and represents the recommended action that a student should take in a given state. Unlike prior action advising methods which issue advice based on simple temporal heuristics (Torrey & Taylor, 2013), our proposed method introduces an intelligent teacher which identifies *when* its advice would be useful, even in cases where the task differs between the source and target. We refer to this capability as *introspection*, which is achieved by contrasting the expected reward for an advised action in the source task with the observed reward in the target task. If this difference is small enough – the outcome for an action in both tasks is expected to be similar – then the teacher’s advice is seen as transferable and advice is issued.

Our approach has several advantages over fine-tuning: 1) knowledge is *selectively transferred*, allowing us to distinguish task-specific knowledge which is not generalizable from knowledge that is generalizable; 2) the transferred knowledge is *interpretable* as an observer can examine the state-action pairs to see precisely what information has been exchanged; and 3) our method is agnostic of the underlying model architectures and representations since knowledge is transferred in terms of state-action pairs which are properties of the environment, not the policy. We empirically show that this leads to improved convergence rates over fine-tuning between tasks in Gridworld and Atari environments while allowing qualitative insights into where transfer is useful.

2 RELATED WORK

Action Advising: Action advising has a rich history in reinforcement learning, as practitioners have long been aware of the sample inefficiency associated with the credit assignment problem in model-free methods such as those derived from temporal difference learning (Watkins, 1989; Whitehead, 1991). Early works explored the idea of injecting advice from human observers, through the use of experience replay (Lin, 1992) and logic-based instructions (Maclin & Shavlik, 1994). The notion of the teacher-student paradigm was formalized shortly after (Clouse & Utgoff, 1992), in which a human teacher may intervene to directly provide an action that a student will follow in a particular state. This field has continued to receive attention over the years, including early attempts at leveraging action advising for knowledge transfer between simple RoboCup tasks (Torrey et al., 2005). Subsequent work (Torrey et al., 2006; 2010) leveraged inductive logic programming for similar purposes in order to accommodate imperfect advice, however, these methods were heavily constrained by the rigid incorporation of domain knowledge.

More recent action advising methods have focused on tackling the problem of *when* to give advice in order to maximize efficiency, resulting in the introduction of budget-based teaching methods where advice-giving is initiated both by the teacher (Torrey & Taylor, 2013), the student (Amir et al., 2016), and peers (Da Silva et al., 2017). Other approaches have framed this problem as a reinforcement learning one in which the teacher is an agent learning an optimal policy for issuing advice while considering task returns (Zimmer et al., 2014; Fachantidis et al., 2017). However, the heavy

computational burden of such an approach has led to simpler methods involving the explicit storage of previously received advice for future use (Ilhan et al., 2019; Zhu et al., 2020a). Decision trees have also been proposed as a more general advice re-use mechanism (Guo et al., 2023), with the added benefit of increasing interpretability into the reasoning for why advice was issued.

Transfer Learning: Transfer learning and multi-task learning (Caruana, 1998) have long been studied in the scope of deep supervised learning (Tan et al., 2018), with pre-training-based methods (Hinton & Salakhutdinov, 2006; Mesnil et al., 2012; Bengio, 2012) proving effective in a large variety of domains. However, transferring knowledge between tasks in reinforcement learning agents (Zhu et al., 2020c; Xie et al., 2022) has often proved more difficult due to the variability in tasks (Rusu et al., 2015) and Markov Decision Process-related assumptions (Parisotto et al., 2015). Despite this, progress has been made in introducing methods capable of alleviating forgetting and generalizing knowledge (Rusu et al., 2016). These methods typically come at a cost in unbounded parameter growth, specialized model architectures, and training algorithms, however, which limits their application and contributes to the continued prevalence of pre-training and fine-tuning. More practical methods based on policy distillation (Czarnecki et al., 2019; Schmitt et al., 2018) and recurrent models (Ni et al., 2022; Caccia et al., 2022) have been shown to exhibit positive knowledge transfer, however, a detailed analysis examining transfer performance between agents trained on single tasks which vary from each other is lacking.

An alternative approach has been proposed in meta-learning (Finn et al., 2019) which seeks to learn models capable of effectively generalizing to new tasks and has been shown to be effective in reinforcement learning settings (Flennerhag et al., 2021). Nonetheless, the same limitations from prior methods still apply: such approaches require the practitioner to explicitly employ meta-learning algorithms during training and don't readily enable knowledge transfer from prior policies that have been trained to solve specific tasks. Furthermore, knowledge is typically transferred in an opaque manner (Ramakrishnan & Shah, 2016) such that an external observer has little insight into the process. Despite attempts to address interpretability (Kim et al., 2019; Zabounidis et al., 2023), this drawback is increasingly problematic as deep learning models are employed for tasks in which decisions have significant impact in the real-world and the rationale for which may need to be scrutinized.

3 BACKGROUND

We examine the reinforcement learning setting in which an agent interacts with an environment at each time step and receives a reward signal, acting in such a way so as to maximize the reward over time.

Notation: Consider a discounted infinite-horizon Markov Decision Process (MDP) in which an agent observes environment state s_t at a discrete timestep t , performs action a_t , and receives the next state s_{t+1} and reward r_{t+1} from the environment. The MDP is formalized as a tuple consisting of $(\mathcal{S}, \mathcal{A}, R, T, \gamma)$, where \mathcal{S} is a set of states in the environment, \mathcal{A} is a set of agent actions, $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is a reward function such that $R(s, a) = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is a state transition probability $T(s', s, a) = P(s_{t+1} = s' | s_t = s, a_t = a)$, and $\gamma \in [0, 1]$ is a discount factor. Let $\pi(a|s) : \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ denote a stochastic policy which is the probability that an agent chooses action a given state s . Using standard definitions (Sutton & Barto, 2018) for the state-value and action-value functions,

$$V_\pi(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t \right],$$

$$Q_\pi(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t, a_t \right],$$

where $\mathbb{E}_\pi[\cdot]$ is the expected value given that an agent follows policy π . An agent's goal is to learn π through training such that the expected discounted return V_π is maximized, given an initial state distribution.

Action advising is based on the teacher-student framework in which a teacher agent attempts to help a student agent learn an optimal policy more quickly by issuing advice during training. We define a general form of action advising in which the teacher has access to a policy π^T which has been pre-trained in a *source* task, and gives advice in the form of state-action pairs to a student that is training policy π^S in a *target* task. As above, let s_t denote the student's observed environment state at time t and a_t the action that the student takes, such that:

$$a_t = \begin{cases} a_t^T \sim \pi^T(a_t | s_t) & \text{if } H(s_t) = 1, \\ a_t^S \sim \pi^S(a_t | s_t) & \text{if } H(s_t) = 0, \end{cases}$$

where $H : \mathcal{S} \mapsto \{0, 1\}$ is an indicator function which takes the value of 1 when the teacher gives advice, and 0 otherwise. We assume the student always follows advice when given, and thus takes either the teacher's action a_t^T or its

Algorithm 1 INTROSPECTIVE ACTION ADVISING (Actor-Critic)

Input: $\pi^T, \pi^S, V_{\pi^T}^{\text{Src}}, V_{\pi^T}^{\text{new}}$

- 1: **for** iteration = 1, 2, ... **do**
- 2: Roll out student’s policy for T timesteps
- 3: $X \leftarrow \emptyset$
- 4: **for** $t = 1, 2, \dots, T$ **do**
- 5: $h_t \leftarrow \text{INTROSPECT}(s_t, V_{\pi^T}^{\text{Src}}, V_{\pi^T}^{\text{new}}, t)$
- 6: **if** $h_t = 1$ **then**
- 7: $a_t \sim \pi^T(a_t|s_t)$
- 8: **else**
- 9: $a_t \sim \pi^S(a_t|s_t)$
- 10: **end if**
- 11: $X \leftarrow X \cup (s_t, a_t, s_{t+1}, r_{t+1}, h_t)$
- 12: **end for**
- 13: Teacher critic update over X
- 14: $\rho^T, \rho^S \leftarrow \text{CORRECT}(X, \pi^S, \pi^T)$
- 15: $\theta_{V^T} \leftarrow \theta_{V^T} + \alpha \rho^T \nabla_{\theta_{V^T}} \mathcal{L}_t^V$
- 16: Student actor-critic update over X
- 17: $\theta_{V^S} \leftarrow \theta_{V^S} + \alpha \rho^S \nabla_{\theta_{V^S}} \mathcal{L}_t^V$
- 18: $\theta_{\pi^S} \leftarrow \theta_{\pi^S} + \alpha \rho^S \nabla_{\theta_{\pi^S}} \mathcal{L}_t^\pi$
- 19: **end for**

Algorithm 2 INTROSPECT

Input: $s_t, V_{\pi^T}^{\text{Src}}, V_{\pi^T}^{\text{new}}, t, \epsilon, \lambda$

- 1: $h_t \leftarrow 0$
- 2: $p \sim \text{Bern}(\lambda^{\max(0, t-\delta)})$
- 3: **if** $t > \delta$ and $p = 1$ **then**
- 4: **if** $|V_{\pi^T}^{\text{new}}(s_t, a_t) - V_{\pi^T}^{\text{Src}}(s_t, a_t)| \leq \epsilon$ **then**
- 5: $h_t \leftarrow 1$
- 6: **end if**
- 7: **end if**

Return: h_t

Algorithm 3 CORRECT

Input: X, π^S, π^T

- 1: $\rho^T \leftarrow \emptyset, \rho^S \leftarrow \emptyset$
- 2: **for** each a_i, s_i, h_i in X **do**
- 3: **if** $h_i = 1$ **then**
- 4: $\rho^T \leftarrow \rho^T \cup 1, \rho^S \leftarrow \rho^S \cup \frac{\pi^S(a_i|s_i)}{\pi^T(a_i|s_i)}$
- 5: **else**
- 6: $\rho^T \leftarrow \rho^T \cup \frac{\pi^T(a_i|s_i)}{\pi^S(a_i|s_i)}, \rho^S \leftarrow \rho^S \cup 1$
- 7: **end if**
- 8: **end for**

Return: ρ^T, ρ^S

own action a_t^S if $H(s_t) = 1$ or $H(s_t) = 0$, respectively. In this work we make the assumption that the state and action spaces for the source and target tasks are the same, although the distribution of states and actions in each task differ.

Problem Formulation: Prior works in action advising assume that the teacher and student are learning policies for the same task and domain – that is, an optimal policy in the source task is also optimal in the target task. Our goal is to relax this constraint, and develop an introspection module $H(\cdot)$ capable of identifying a subset of states for which the teacher agent issues beneficial advice. We define beneficial advice as an action that, if taken, leads to an increased reward from the given state when compared to an action generated by the student. Formally, let \mathcal{S} represent the set of possible states in the target task. We seek a function $H(\cdot)$ such that if $H(s) = 1$, then $Q_{\pi^S}(s, a^T) > Q_{\pi^S}(s, a^S)$ for all $s \in \mathcal{S}$.

4 INTROSPECTIVE ACTION ADVISING

We introduce an *introspective* teacher in the action advising setting which determines when its advice is transferable to a target task, and thus beneficial to the student. This is necessary when the advice is derived from a policy that has been learned in a different source task, as otherwise the teacher may actively guide the student to repeatedly explore regions of the state space with low rewards. We define advice as being transferable when the associated action leads to expected discounted returns that are sufficiently similar across tasks; that is, advice should be issued if it leads to an outcome in the target task that is “close enough” to the outcome in the source task. Formally, a_t is transferable for a given state s_t if

$$|Q_{\pi^T}^{\text{Tar}}(s_t, a_t) - Q_{\pi^T}^{\text{Src}}(s_t, a_t)| \leq \epsilon \quad (1)$$

is satisfied, where $Q_{\pi^T}^{\text{Tar}}$ and $Q_{\pi^T}^{\text{Src}}$ are the action-value functions of the teacher in the source and target tasks respectively, and ϵ is a similarity threshold. Rather than considering whether a specific *action* is transferable, however, we consider whether a *policy* is transferable in a given state. That is, we modify Eq. 1 to operate over state-value functions such that

$$|V_{\pi^T}^{\text{Tar}}(s_t) - V_{\pi^T}^{\text{Src}}(s_t)| \leq \epsilon, \quad (2)$$

which yields a transferable action $a_t \sim \pi^T(a_t|s_t)$. Recall that in our problem formulation in Sec. 3 we seek to identify beneficial states for which $Q_{\pi^S}(s, a^T) > Q_{\pi^S}(s, a^S)$. Let the student’s policy be randomly initialized at $t = 0$, then an action generated by the student’s policy, a_t^S , is largely random for $t < \tau$ where τ is a temporal cut-off. We assume then that actions sampled from a teacher’s policy yield a higher expected reward than actions sampled from a student’s policy, $Q_{\pi^S}(s_t, a_t^T) > Q_{\pi^S}(s_t, a_t^S)$, for $t < \tau$ if the teacher is better than random. Therefore, we define a_t^T to be both beneficial and transferable if Eq. 2 holds for $a_t^T \sim \pi^T(a_t|s_t)$ and $t < \tau$. Intuitively, τ represents a point in time after which we no longer assume the teacher’s actions will lead to better rewards than the student’s.

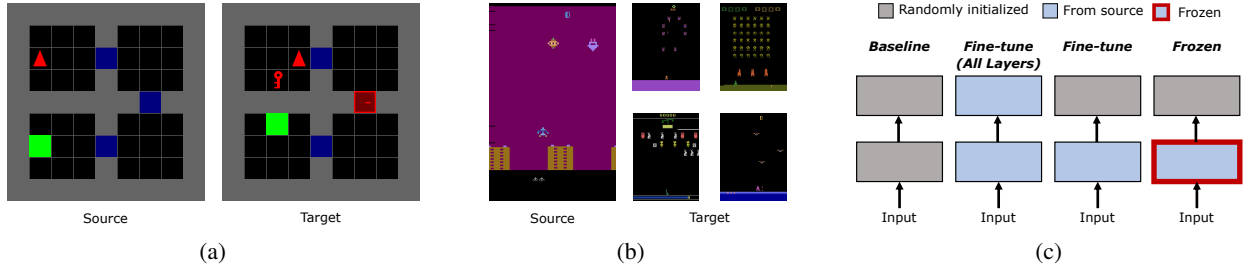


Figure 2: (a) The source and target Gridworld tasks. Agent start (red triangle), goal (green square), and key positions are randomized. The agent must learn to navigate to the goal. The source task contains no locked doors while the target task contains exactly one. (b) The source task (Air Raid) and a subset of target tasks in Atari (Phoenix, Space Invaders, Carnival, and Demon Attack shown). The nature of the task is dependent on the game. (c) The model architectures for the four baseline models. In Baseline, all layers are randomly initialized and trained in the target task. In Fine-tune (All Layers), all layers have weights copied from the source policy but are fine-tuned in the target task. In Fine-tune, only the input layers have weights copied from the source policy while the remaining are randomly initialized. Frozen is the same as Fine-tune, except the input layer is frozen and not updated during training in the target task.

Correspondingly, our approach involves accurately estimating $V_{\pi^T}^{\text{Tar}}$ in the target task. However, we assume advice is only beneficial in the early stages of training when $t < \tau$ and it is unlikely that our estimate of $V_{\pi^T}^{\text{Tar}}$ has sufficiently converged to the optimal value function in the target task, $V_{\pi^T}^*$, after a limited number of environment samples. That is,

$$|V_{\pi^T}^*(s_t) - V_{\pi^T}^{\text{Tar}}(s_t)| > \epsilon, \quad (3)$$

rendering it difficult to determine whether advice is not transferable due to a mismatch in state-values or due to the student not having a sufficiently accurate estimate of the state-value (and thus ultimately being transferable *if* the estimate were more accurate).

Introspection: We address this issue by having the teacher itself directly estimate the state-value function in the target task, with the assumption that refining the teacher’s existing estimate of the state-values in the source task will lead to quicker convergence to $V_{\pi^T}^*$ than estimating it from scratch. The process is as follows: we fine-tune a copy of the teacher’s state-value function learned in the source task, $V_{\pi^T}^{\text{new}}$, with observed returns from the student’s exploration in the target task and use this as our estimate of $V_{\pi^T}^*(s_t)$. This leads to the following modification of Eq. 2,

$$|V_{\pi^T}^{\text{new}}(s_t) - V_{\pi^T}^{\text{Src}}(s_t)| \leq \epsilon. \quad (4)$$

We incorporate a fuzzy approximation of the temporal cut-off τ through the introduction of a decay hyperparameter, $\lambda \in [0, 1)$. The teacher issues advice according to an annealed probability over time, λ^t , with the goal of issuing advice at a sufficiently low probability for $t > \tau$. The intuition behind this is that while a given action may result in similar outcomes in both the source and target tasks, it may no longer be the *optimal* outcome, and the student should be increasingly allowed to perform its own exploration over time. Additionally, there is a trade-off inherent to this process between how many returns the teacher should observe and incorporate into $V_{\pi^T}^{\text{new}}$ and the need to issue advice early in the training to obtain maximum effect (Guo et al., 2023). This is controlled through a *burn-in* hyperparameter $\delta \in \mathbb{R}_{\geq 0}$, which dictates how many timesteps the teacher should observe before issuing advice.

Off-Policy Correction: A consequence of introspective action advising is that the samples collected for training updates are no longer generated purely by rolling out the student’s policy π^S . Actions produced by the student’s policy are interspersed with advised actions from the teacher’s policy, meaning that the resulting trajectory samples will violate assumptions for on-policy algorithms and as such, need an off-policy correction applied to their losses. Let $X = x_0, \dots, x_n$ represent a trajectory of n samples obtained through the student’s interaction with the environment, where $x_t = (s_t, a_t, s_{t+1}, r_{t+1})$ is the observation for a single timestep t . Given a loss \mathcal{L} computed over X , we apply an off-policy correction ρ^S for the student such that we optimize $\rho^S \mathcal{L}$ with respect to the student’s model parameters. The off-policy correction is a vector of n scalars where each value is an importance sampling ratio $\frac{\pi^S(a_t|s_t)}{\pi^T(a_t|s_t)}$ if a particular action is sampled from the teacher’s policy and 1 otherwise. Similarly, we need to apply an off-policy correction ρ^T when optimizing the parameters of the teacher’s state-value function estimate $V_{\pi^T}^{\text{new}}$. The correction ρ^T contains an importance sampling ratio $\frac{\pi^T(a_t|s_t)}{\pi^S(a_t|s_t)}$ if an action is sampled from the student’s policy and 1 otherwise.

Algorithm: We describe our approach, *Introspective Action Advising* (IAA), in an actor-critic form in Alg. 1. Lines 3-12 roll out the student’s policy to collect samples, where actions come from either the teacher’s policy (Line 7) or

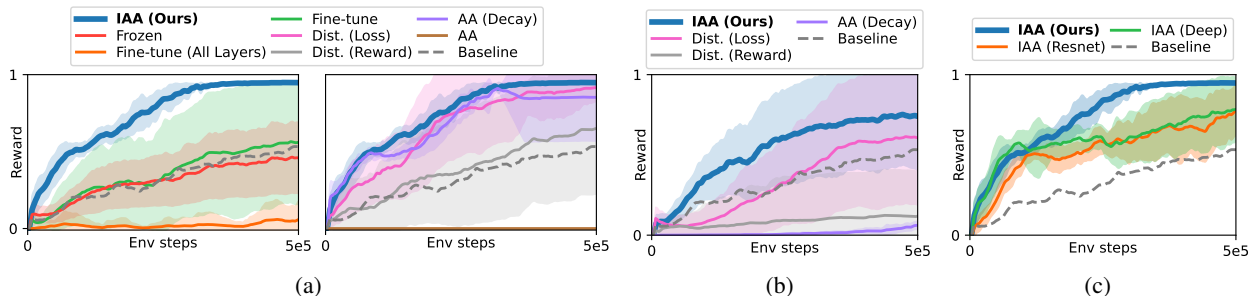


Figure 3: (a) The training curves between our proposed method and fine-tuning methods (left) and transfer methods methods (right) given a Gridworld environment where the source and target tasks are similar enough such that the teacher issues helpful advice. (b) The training curves for a Gridworld environment where the source and target tasks differ such that the teacher largely gives unhelpful advice. (c) The training curves for our proposed method in which the student and teacher models use different architectures.

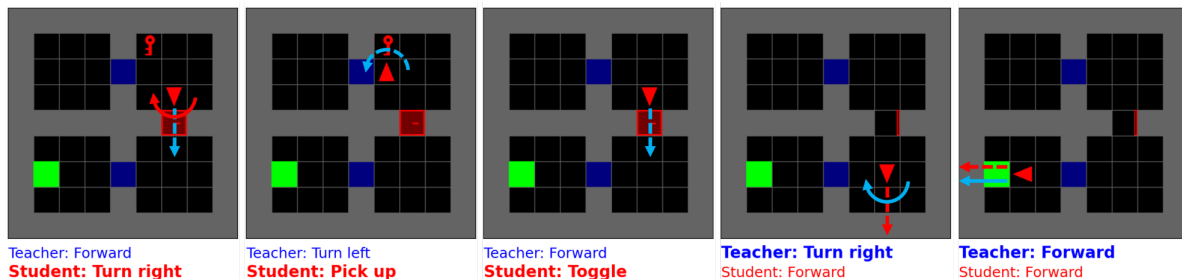


Figure 4: A sequence from a single training episode showing the student’s sampled action (red) and the teacher’s advised action (blue) in Gridworld. Bold text indicates the executed action as determined by IAA given $\epsilon = 0.1$.

student’s policy (Line 9). Our introspection function (Alg. 2) replaces $H(\cdot)$ and determines whether advice should be issued to the student. An off-policy correction is computed in Alg. 3 as described above, and then applied to the teacher’s critic loss \mathcal{L}_t^V in Line 15, and the student’s actor (\mathcal{L}_t^{π}) and critic (\mathcal{L}_t^V) losses in Lines 17 and 18, respectively. The losses are optimized over with respect to the model parameters of the teacher’s critic ($V_{\pi^T}^{\text{new}}$), θ_{V^T} , as well as the student’s actor and critic parameters: θ_{π^S} and θ_{V^S} given a learning rate α . This is a generic formulation and IAA can be readily applied to many on-policy algorithms, as we do for Proximal Policy Optimization (PPO) (Schulman et al., 2017) in our experiments, as well as off-policy algorithms with the exclusion of the proposed off-policy correction.

5 EXPERIMENTS

We quantitatively and qualitatively analyze our proposed algorithm over two different reinforcement learning domains: Gridworld and Atari. In each domain we examine learning performance when transferring between tasks of varying similarity with the goal of understanding: 1) does our proposed approach yield positive transfer performance between tasks; 2) how does our approach compare to methods based on fine-tuning and action advising; and 3) does our approach allow us to understand what knowledge has been transferred between tasks.

5.1 EXPERIMENTAL SETUP

In each domain we define a fixed source task for which to train a source policy and then evaluate transfer learning performance to a policy trained in one or more target tasks. We compare our approach to the following six methods:

- **Fine-tuning variants:** Three fine-tuning variants in Fig. 2c.
- **Action Advising:** Action advising at every step with no budget constraint.
- **Action Advising (Decay):** Action advising which has been augmented with our proposed advice decay rate λ . Rather than issuing advice at every timestep, it is issued according to a decaying probability over time.

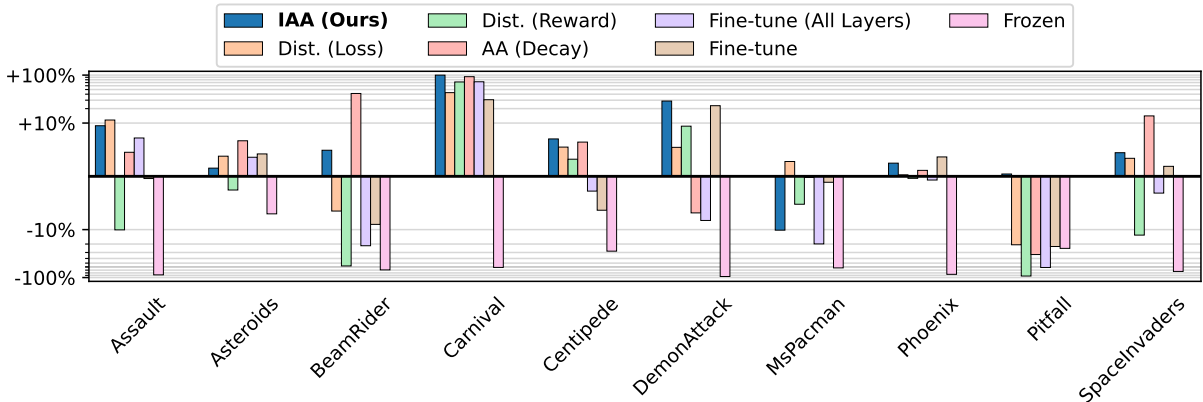


Figure 5: The relative performance of our proposed approach and the transfer learning baseline models defined in Fig. 2c to the Baseline policy learned in the target task without any transfer. Each value represents the percentage of improvement in the mean reward after 10M environment steps over four independent training runs with randomly initialized weights. The target task is shown along the x -axis with Air Raid used as the source task in all scenarios.

Metric	IAA (Ours)	Dist. (Loss)	Dist. (Reward)	AA (Decay)	Fine-tune (All Layers)	Fine-tune	Frozen
Positive Transfer Ratio	9/10	8/10	3/10	7/10	3/10	5/10	0/10
Minimum Transfer	-10%	-21%	-93%	-48%	-90%	-33%	-95%
Median Transfer	4.6%	3.6%	-3.9%	5.4%	-3.0%	0.7%	-66%
Maximum Transfer	99%	43%	72%	93%	73%	31%	-7%

Table 1: Summary of transfer performance shown in Fig. 5. Positive Transfer Ratio is the number of tasks for which performance improved, Minimum Transfer is the smallest relative performance achieved over all tasks, and Maximum Transfer is the greatest relative performance achieved. Bold represents the best value for each metric.

- **Policy Distillation (Loss)**: Policy distillation through the introduction of a cross-entropy auxiliary loss between the teacher’s and student’s policies (Schmitt et al., 2018).
- **Policy Distillation (Reward)**: Policy distillation through a reward shaping term which captures the difference in the teacher’s critic between the current and previous timesteps (Czarnecki et al., 2019).
- **Baseline**: A policy trained from scratch in the target task.

In addition to fine-tuning methods used in prior works (Rusu et al., 2016), we chose to compare to two widely-used policy distillation methods which transfer knowledge from a teacher policy through auxiliary losses and reward shaping terms. The source and target policies are trained using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with identical hyperparameters and model architectures as described in Appendix C. Our approach is agnostic of the specific reinforcement learning algorithm used to train the underlying policies, and we have chosen PPO solely due to its simplicity and consistency. We evaluate each method according to the *relative performance* with respect to the Baseline policy trained over the target task in the absence of any transfer. A larger return with respect to the Baseline is referred to as positive transfer, while a lower return is negative transfer. All methods are evaluated with the same set of hyperparameters so as to make consistent comparisons. We evaluate IAA with up to three different values for the threshold ϵ and show the results for the best performing one, with full results in Appendix A.

Gridworld: The Gridworld environment consists of four inter-connected rooms and an agent which must learn to navigate to a goal position as shown in Fig. 2a. In the source task, there are no doors and the agent is free to navigate to any room in order to reach the goal. In the target task, there is exactly one locked door separating the agent from the goal; a key must first be picked up and used on the door in order to unlock it and provide access to the goal. The agent, goal, and key positions are randomized in every episode and policies are trained for 10 random seeds with 500K environment steps each. A sparse reward is given when the agent reaches the goal, however, dense rewards are also explored in Appendix A.

Atari: The Atari environment consists of a subset of games selected from the Arcade Learning Environment (Machado et al., 2018). This is a challenging domain owing to the differences in visuals, strategy, and actions in each game,

and has previously been shown to yield negative transfer for fine-tuning-based transfer methods (Rusu et al., 2016). We have chosen a set of nine games sharing similar visuals and strategies to evaluate for transfer – Air Raid, Assault, Asteroids, Beam Rider, Carnival, Centipede, Demon Attack, Phoenix, and Space Invaders – as well as two games which substantially differ – Ms. Pacman and Pitfall. Air Raid was arbitrarily chosen as the source task and the remaining ten games serve as the target tasks. All policies are trained with 4 random seeds and 10M environment steps each; while this is not enough to converge, it is enough to analyze transfer performance with limited computational resources.

5.2 RESULTS - GRIDWORLD

Transfer Performance: The training curves of each method are shown in Fig. 3. In Fig. 3a we can see that IAA significantly out-performs all of the fine-tuning based variants. In particular, we can observe that the fine-tuning methods fail to yield positive transfer and under-perform the Baseline after $2e5$ environment steps, while IAA results in positive transfer with a 66% increase in reward. This is an interesting observation and seems to indicate that the target policy is unable to separate features useful for navigation in an environment without doors to one with locked doors. Additionally, this is one of the only cases we encountered where fine-tuning with a frozen input layer outperforms any of the other fine-tuning variants, indicating that randomly initialized layers after the input are more beneficial – likely due to over-fitting to the source task. Our decayed version of action advising, AA (Decay), fares better in this regard, producing positive transfer with a 48% reward increase. Contrast this with the non-decayed action advising variant, AA, which always gives advice at every timestep and fails to produce any meaningful policy. It is clear that while *some* of the advice is useful in the target task, the target policy must still be allowed to explore on its own. We hypothesize that this is also why Distillation (Reward) yields no performance improvement; the shaping term as proposed in Czarnecki et al. (2019) does not decay over time and the student never becomes fully independent. Distillation (Loss) does decay over time, however, and subsequently we see performance gains in line with IAA and AA (Decay).

Figure 3b shows performance when the source task differs more significantly – the agent starts in the bottom rooms and must navigate to the top rooms – and consequently the teacher issues more unhelpful advice. This is where we see the advantage of IAA, as it is able to successfully filter out unhelpful advice and still achieve modest performance improvements, at the expense of increased variance due to the sparse reward. The policy distillation methods and AA (Decay) all perform substantially worse than before due to transferring all knowledge, helpful or not.

Additionally, we evaluate how well IAA performs when the model architecture differs between the student and the teacher. In Fig. 3c, we examine performance when the teacher uses a similar but deeper architecture as the student with more than 4x the number of parameters, IAA (Deep), and when the architecture completely differs, IAA (Resnet). In both cases IAA still manages to achieve positive transfer, albeit to a less degree than when the architecture is the same.

Interpretability: Figure 4 shows a sequence of states from a training episode for IAA. In the first three states, the teacher gives seemingly unhelpful advice as it advises the student to move in the direction of the locked door before the student has picked up the key needed to unlock it. This is expected behavior, as the source task in which the teacher was trained contains no doors and keys and so the teacher is unaware of the key. However, this advice is filtered out by IAA and the student instead performs its own action and collects the key. Once the agent enters the bottom rooms, the teacher’s advice becomes helpful again and is issued to the student in the fourth and fifth states. The fourth state in particular demonstrates helpful advice, where the student would have moved in the wrong direction with its own action, but instead moves towards the goal by following the teacher’s advised action.

5.3 RESULTS - ATARI

Transfer Performance: Figure 5 shows the relative performance values for each method over each target task after 10M environment steps. Across all target tasks we can observe several trends: IAA consistently yields positive transfer between tasks, achieving the most number of tasks with improved performance; IAA achieves the best *negative* transfer performance for a single task, producing the least worst result when a task fails to transfer; and IAA also results in the best positive transfer performance for a single task. These results are summarized in Table 1. This means that while IAA does not produce the best *median* transfer performance – that falls to AA (Decay) – it does provide the highest positive transfer rate while minimizing performance losses in the event transfer fails. We opted not to show the results for AA as they are not meaningful; this reduces to simply running the source policy in the target task which consistently yields performance worse than the Frozen baseline.

Figure 6 shows training performance for two target tasks. The training returns for Carnival are shown in Fig. 6a where we can see that IAA produces a significantly higher return than the Baseline while dramatically improving the convergence rate. This task is interesting in that it is one of the few examples where fine-tuning over all layers performs quite well, out-performing fine-tuning over just the input layers. This seems to indicate that the state-action distribution

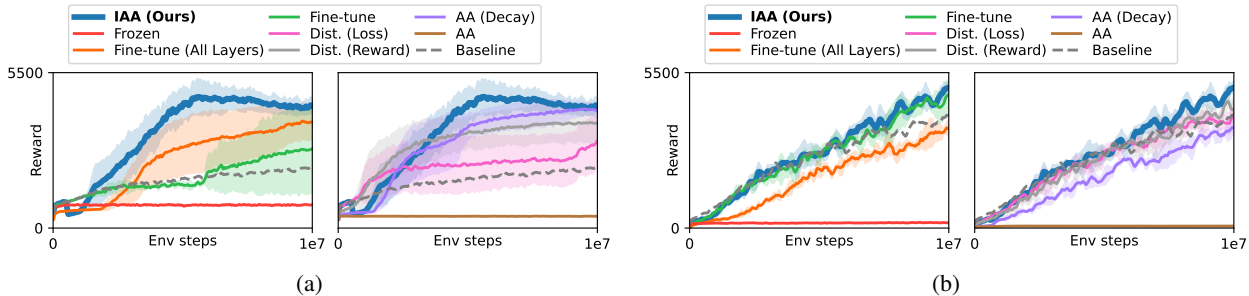


Figure 6: The training curves between our proposed method and transfer learning methods (left) and advising methods (right) for (a) Carnival and (b) Demon Attack. Each line is the average reward over four independent training runs for 10M environment steps, with the shaded region indicating standard deviation.

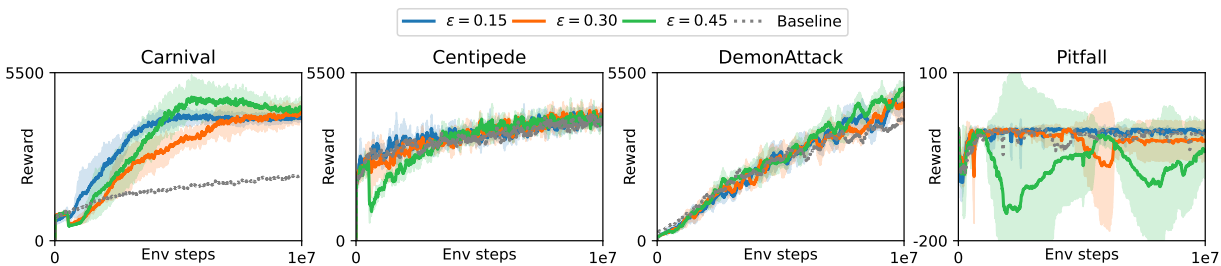


Figure 7: Training curves for multiple introspection thresholds of IAA for different Atari tasks.

is quite similar between source and target tasks, which would also contribute to the high performance of both IAA and AA (Decay). Another interesting trend is present which sometimes emerges for IAA, which is the drop in performance once the burn-in period is over and advice begins to be issued. This drop occurs after 500K steps – corresponding with $\delta = 500K$ – and has two potential causes: 1) the burn-in period wasn’t long enough for $V_{\pi_T}^{new}$ to sufficiently converge to $V_{\pi_T}^*$, and 2) the behavior distribution shift induced by issuance of advice means that even if it *had* converged for the prior behavior policy distribution, it may not be anymore. This invites the possibility of future work which investigates an adaptive introspection threshold conditioned on distribution shifts in behavior policy.

Figure 6b shows training performance from another game, Demon Attack, exhibiting more modest improvements. Rather than the significant increase to convergence rate as seen in Carnival, this task produces slower, more consistent improvements at the same level as fine-tuning. From these two tasks we can observe a more general trend: fine-tuning with a frozen input layer nearly always results in negative task transfer and poor policies after 10M steps. This is unsurprising given that the features encoded from the observation are likely over-fit to the source task, but adds further evidence that issuing advice in the form of state-action pairs which is agnostic to the underlying visual features (like in IAA) may be a desirable strategy.

Figure 7 shows the sensitivity of IAA to different values of the introspection threshold. While in general the tested thresholds yield similar results – suggesting that comprehensive hyperparameter tuning may not be necessary – an overly high threshold can cause large performance drops for certain tasks. In Centipede for instance, $\epsilon = 0.45$ produces a sharp drop in performance right after the burn-in period has elapsed and advice is issued. Similarly, in Pitfall, a target task which differs significantly from the source task, we can observe increasing variance in the returns as more advice is issued. Intuitively, we observe that IAA should be used with a more conservative introspection threshold for target tasks which differ significantly from the source task, which consequently leads to a decrease in expected performance improvement ($\epsilon = 0$ reduces to baseline performance). Figure 8 examines this in further detail, showing that there is indeed a negative correlation between performance and distribution shift between source and target observation distributions, with $R^2 = 0.34$.

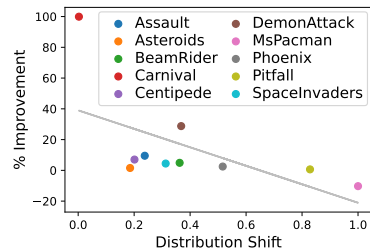


Figure 8: The relative performance of IAA compared to the distribution shift of each Atari task.

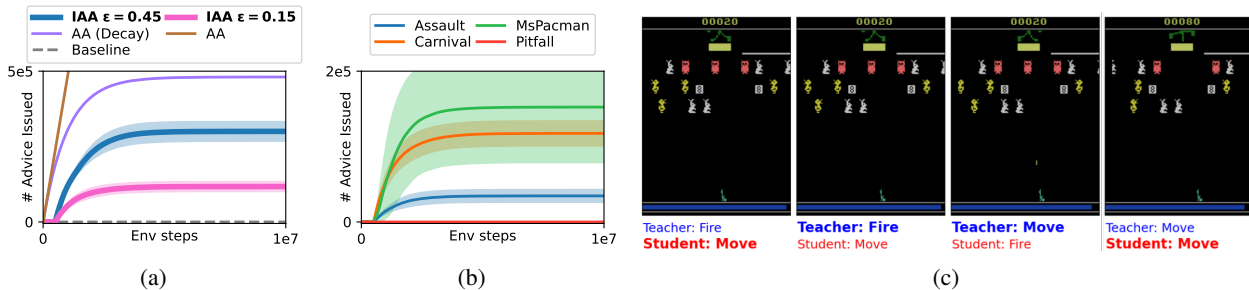


Figure 9: (a) The amount of advice issued by each advising algorithm for the game Carnival. (b) The amount of advice issued across Atari games by IAA given $\epsilon = 0.15$. (c) A sequence of states from a single training episode showing the student’s sampled action (red) and the teacher’s advised action (blue) in Carnival. Bold text indicates the executed action as determined by IAA given $\epsilon = 0.45$.

The amount of advice issued by the advising algorithms for a single game, Carnival, is shown in Fig. 9. As in the Gridworld environment, we see again that IAA is successful in identifying only useful knowledge and achieves higher returns by giving fewer pieces of advice. Figure 9b shows the amount of advice issued for IAA across target tasks when the introspection threshold is fixed. Unsurprisingly, this indicates that some target tasks have an expected reward distribution more similar to the source task than others, resulting in more advice being issued by IAA. However, this figure also indicates that $V_{\pi_T}^*$ may be more difficult to learn in some tasks, resulting in a large variance in issued advice as seen in Ms. Pacman. This variance is undesired and as seen in Fig. 5 led to the only negative task transfer for IAA, likely due to the introspective teacher failing to discriminate beneficial advice from non-beneficial advice.

Interpretability: Figure 9c shows a sequence of states from a training episode for IAA in Carnival. We can observe that in the first two frames the teacher advises the student to fire, an action which is filtered out in the first state but issued and executed in the second. Afterwards, the teacher advises the student to perform movement actions in order to reposition the agent so further fire actions can be taken. In the fourth state we can observe that the advice is indeed helpful, as the advised action results in a high-value target being destroyed in the furthest row and a 60 point increase to the score. This example is demonstrative of the insight which IAA provides into which knowledge is transferred between the teacher and the student; an external observer need only examine the states for which advice is issued. While prior methods such as policy distillation also transfer knowledge, it is much more difficult to analyze an auxiliary loss term for a particular state or even a reward shaping term – particularly as rewards require one to sample actions and perform a comparative analysis to other state-action rewards.

6 CONCLUSION

In this work, we have proposed Introspective Action Advising, an approach to transfer learning between tasks in reinforcement learning which is agnostic to model architectures and learning algorithms. We show that our method is capable of distinguishing which knowledge is beneficial to a target task, and as a result yields consistently positive task transfer while minimizing penalties to performance when there is little task overlap. As such, we think this affords a promising alternative to the standard approaches of fine-tuning and policy distillation, particularly for practitioners attempting to leverage knowledge contained in models that may be over-fit to a single task. At the same time, this allows one to examine the issued state-action advice and understand which knowledge was transferred, potentially providing an opportunity for more fine-grained knowledge transfer. We recognize that our proposed approach has limitations, particularly that an introspection threshold hyperparameter must be specified which appears to be source-task dependent and controls what advice is seen as transferable. While we believe an adaptive version may be possible, we leave this to future work. Moreover, our identification of transferable advice relies on estimating the expected value of a state sampled in the target task with respect to a policy trained in the source task (Eq. 4). We have shown that this works well empirically in two reinforcement learning domains in this work, however, we conjecture that if the source and target task were to differ more significantly then value estimates would become unstable due to the out-of-distribution nature of the samples. We expect that the resulting variance would result in more false negatives as opposed to false positives – failing to issue advice rather than failing to suppress advice as seen in the transfer from Air Raid to Pitfall – and plan to investigate this in the future.

ACKNOWLEDGEMENTS

This work has been funded in part by DARPA under grant HR001120C0036, the Air Force Office of Scientific Research (AFOSR) under grants FA9550-18-1-0251 and FA9550-18-1-0097, and the Army Research Laboratory (ARL) under grant W911NF-19-2-0146 and W911NF-2320007.

REFERENCES

- Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara Grosz. Interactive teaching strategies for agent training. In *Proceedings of IJCAI 2016*, 2016.
- Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36. JMLR Workshop and Conference Proceedings, 2012.
- Massimo Caccia, Jonas Mueller, Taesup Kim, Laurent Charlin, and Rasool Fakoor. Task-agnostic continual reinforcement learning: In praise of a simple baseline. *arXiv preprint arXiv:2205.14495*, 2022.
- Joseph Campbell and Katsu Yamane. Learning whole-body human-robot haptic interaction in social contexts. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10177–10183. IEEE, 2020.
- Rich Caruana. *Multitask learning*. Springer, 1998.
- Jeffery A Clouse and Paul E Utgoff. A teaching method for reinforcement learning. In *Machine learning proceedings 1992*, pp. 92–101. Elsevier, 1992.
- Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial intelligence in medicine*, 109:101964, 2020.
- Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1331–1340. PMLR, 2019.
- Felipe Leno Da Silva, Ruben Glatt, and Anna Helena Reali Costa. Simultaneously learning and advising in multiagent reinforcement learning. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, pp. 1100–1108, 2017.
- Felipe Leno Da Silva, Garrett Warnell, Anna Helena Reali Costa, and Peter Stone. Agents teaching agents: a survey on inter-agent transfer learning. *Autonomous Agents and Multi-Agent Systems*, 34:1–17, 2020.
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6309–6317, 2019.
- Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.
- Utku Evcı, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pp. 6009–6033. PMLR, 2022.
- Anestis Fachantidis, Matthew E Taylor, and Ioannis Vlahavas. Learning to teach reinforcement learning agents. *Machine Learning and Knowledge Extraction*, 1(1):21–42, 2017.
- Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930. PMLR, 2019.
- Sebastian Flennerhag, Yannick Schroecker, Tom Zahavy, Hado van Hasselt, David Silver, and Satinder Singh. Bootstrapped meta-learning. *arXiv preprint arXiv:2109.04504*, 2021.
- Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.
- Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

- Yue Guo, Joseph Campbell, Simon Stepputtis, Ruiyu Li, Dana Hughes, Fei Fang, and Katia Sycara. Explainable action advising for multi-agent reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.
- Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5): 698–721, 2021.
- Ercüment İlhan, Jeremy Gow, and Diego Perez-Liebana. Teaching on a budget in multi-agent deep reinforcement learning. In *2019 IEEE Conference on Games (CoG)*, pp. 1–8. IEEE, 2019.
- Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.
- Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. *Advances in neural information processing systems*, 30, 2017.
- Dowan Kim, Woohyun Lim, Minye Hong, and Hyeoncheol Kim. The structure of deep neural network for interpretable transfer learning. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–4. IEEE, 2019.
- Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.
- Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8:293–321, 1992.
- Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.
- Richard Maclin and Jude W Shavlik. *Incorporating advice into agents that learn from reinforcements*. University of Wisconsin-Madison. Computer Sciences Department, 1994.
- Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 97–110. JMLR Workshop and Conference Proceedings, 2012.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- Hai Nguyen and Hung La. Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pp. 590–595. IEEE, 2019.
- Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free rl can be a strong baseline for many pomdps. In *International Conference on Machine Learning*, pp. 16691–16723. PMLR, 2022.
- Farzad Niroui, Kaicheng Zhang, Zenda Kashino, and Goldie Nejat. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *IEEE Robotics and Automation Letters*, 4(2): 610–617, 2019.
- Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P How. Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6128–6136, 2019.

- Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*, 2015.
- Ramya Ramakrishnan and Julie Shah. Towards interpretable explanations for transfer learning in sequential tasks. 2016.
- Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.
- Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- Simon Schmitt, Jonathan J Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M Czarnecki, Joel Z Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, et al. Kickstarting deep reinforcement learning. *arXiv preprint arXiv:1803.03835*, 2018.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*, 2019.
- Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3169–3173. IEEE, 2022.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33: 13139–13150, 2020.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pp. 270–279. Springer, 2018.
- Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pp. 1053–1060, 2013.
- Lisa Torrey, Trevor Walker, Jude Shavlik, and Richard Maclin. Using advice to transfer knowledge acquired in one reinforcement learning task to another. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*, pp. 412–424. Springer, 2005.
- Lisa Torrey, Jude Shavlik, Trevor Walker, and Richard Maclin. Skill acquisition via transfer learning and advice taking. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pp. 425–436. Springer, 2006.
- Lisa Torrey, Jude Shavlik, Trevor Walker, and Richard Maclin. Transfer learning via advice taking. *Advances in Machine Learning I: Dedicated to the Memory of Professor Ryszard S. Michalski*, pp. 147–170, 2010.
- Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.
- Steven D Whitehead. A complexity analysis of cooperative mechanisms in reinforcement learning. In *AAAI*, pp. 607–613, 1991.
- Zhihui Xie, Zichuan Lin, Junyou Li, Shuai Li, and Deheng Ye. Pretraining in deep reinforcement learning: A survey. *arXiv preprint arXiv:2211.03959*, 2022.

- Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10674–10681, 2021.
- Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.
- Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
- Yang Yu. Towards sample efficient reinforcement learning. In *IJCAI*, pp. 5739–5743, 2018.
- Renos Zabounidis, Joseph Campbell, Simon Stepputtis, Dana Hughes, and Katia Sycara. Concept learning for interpretable multi-agent reinforcement learning. In *Proceedings of The 6th Conference on Robot Learning*, Proceedings of Machine Learning Research. PMLR, 2023.
- Wenshuai Zhao, Jorge Peña Queraltá, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744. IEEE, 2020.
- Changxi Zhu, Yi Cai, Ho-fung Leung, and Shuyue Hu. Learning by reusing previous advice in teacher-student paradigm. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1674–1682, 2020a.
- Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570*, 2020b.
- Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020c.
- Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.
- Matthieu Zimmer, Paolo Viappiani, and Paul Weng. Teacher-student framework: a reinforcement learning approach. In *AAMAS Workshop Autonomous Robots and Multirobot Systems*, 2014.

A ADDITIONAL RESULTS - GRIDWORLD

A.1 FULL RESULTS

We tested two variants of the Gridworld environment: one with a sparse reward in which a reward is only received when the agent reaches the goal, and one with a dense reward in which the agent receives a reward penalty when it leaves the optimal path to the goal (or key if the agent is operating in the target task and it hasn't been picked up yet). Training curves for the sparse environment are shown in Fig. 3a while those in the dense reward are shown below in Fig. 10.

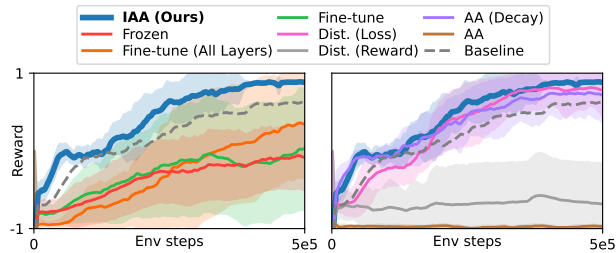


Figure 10: Training curves for IAA under varying hyperparameters in the sparse reward Gridworld environment. All other hyperparameters remain fixed except for the ones indicated.

The average returns over all seeds at different points in time are shown in Table 2.

	Return after X env steps	IAA (Ours)	Dist. (Loss)	Dist. (Reward)	AA (Decay)	Fine-tune (All Layers)	Fine-tune	Frozen	Baseline
Sparse	125K	0.57	0.40	0.23	0.47	0.01	0.27	0.24	0.21
	250K	0.84	0.76	0.37	0.78	0.02	0.34	0.32	0.31
	500K	0.95	0.92	0.65	0.85	0.06	0.56	0.46	0.53
Dense	125K	-0.06	-0.21	-0.75	0.01	-0.71	-0.52	-0.59	-0.04
	250K	0.65	0.46	-0.68	0.46	-0.28	-0.18	-0.27	0.23
	500K	0.88	0.78	-0.68	0.73	0.35	0.04	-0.09	0.63

Table 2: The average return over all seeds after the specified number of environment steps for sparse and dense Gridworld environments. The largest return in each row is shown in bold.

A.2 ADVICE ANALYSIS

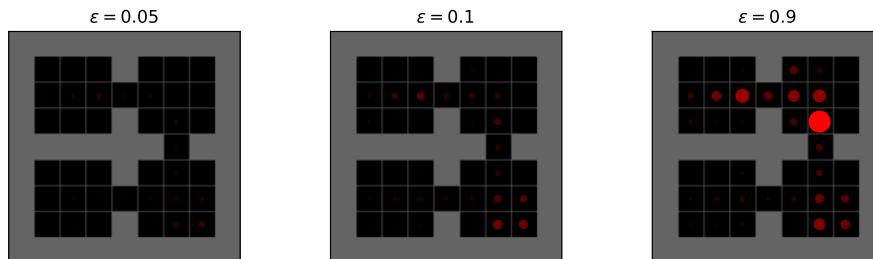


Figure 11: A heatmap visualization showing the states for which advice is issued. The size and opacity of the circles are proportional to the amount of advice issued in that state. The start and goal positions are not visualized as they are randomized each episode while the advice visualized here has been collected over many episodes.

States in which advice has been issued during training are shown in Fig. 11. The size and opacity of each circle indicates the relative frequency for which advice has been issued in that state. The distribution of advice can be seen to shift as the introspection threshold ϵ varies. With a high threshold, 0.9, advice is frequently given in the top area of the

environment where it may sometimes be unhelpful, e.g., see Fig. 4. With lower thresholds, advice is only given in the lower portions near the goal location where there is little difference between the source and target tasks. Note that there is also a trade-off between how quickly the teacher’s critic is fine-tuned in the source environment, i.e., the learning rate, and the introspection threshold, which determines when and where advice is issued.

A.3 HYPERPARAMETER SENSITIVITY ANALYSIS

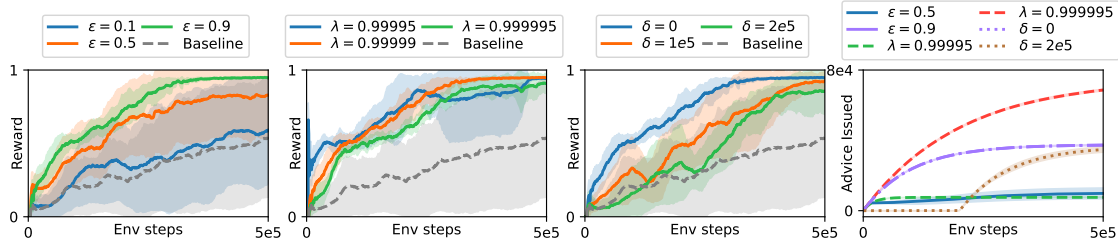


Figure 12: Training curves for IAA under varying hyperparameters in the sparse reward Gridworld environment. All other hyperparameters remain fixed except for the ones indicated.

Figure 12 shows how the performance of IAA varies with respect to the hyperparameters. Introspection threshold yields the largest variance, as expected, as $\epsilon = 0$ reduces to baseline performance. The decay rate has relatively little effect in the Gridworld although the impact can be seen at the beginning as larger thresholds start yielding lower performance numbers due to increased student exploration. The effect of burn-in can be seen in the third figure where the reward of IAA will see a rapid increase from baseline once advice begins to be issued. These hyperparameters all effect the amount of advice issued, which can be seen in the last figure.

A.4 EFFECTS OF REWARD LANDSCAPE

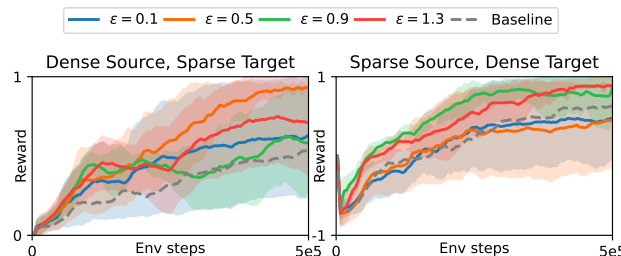


Figure 13: Training curves for IAA when the teacher and student utilize different reward types, i.e., sparse vs dense.

An assumption of IAA is that the rewards are of a similar scale and structure between the source task and the target task, as this is required for determining whether advice is transferable. This is often not a particularly strong assumption in practice, as reward scaling and clipping are often employed, e.g., all of our Atari experiments utilize reward clipping where $r = \text{sign}(r)$. Interestingly, Fig. 13 also shows that IAA is still capable of transferring advice when the source and target tasks utilize different reward structures, e.g., a dense reward and a sparse reward, albeit not as well as when the same reward structure is utilized in both tasks. However, we caution that this result is highly dependent on the type of reward, the learning rate applied to the teacher’s fine-tuned critic, and how quickly the student converges to an optimal policy itself.

B ADDITIONAL RESULTS - ATARI

	Return after X env steps	IAA (Ours)	Dist. (Loss)	Dist. (Reward)	AA (Decay)	Fine-tune (All Layers)	Fine-tune	Frozen	Baseline
Assault	2.5M	1139	1735	1716	863	1888	1558	332	1632
	5M	2158	2557	2022	1949	2336	2267	333	2461
	10M	3118	3177	2559	2976	3053	2836	345	2848
Astrds.	2.5M	1300	1267	1260	1252	1306	1238	1270	1274
	5M	1296	1307	1278	1309	1253	1356	1253	1340
	10M	1333	1363	1279	1401	1360	1368	1220	1313
Bm.Rdr.	2.5M	607	600	602	866	570	594	477	578
	5M	994	878	597	1431	756	855	495	914
	10M	1754	1564	715	2365	1310	1522	515	1673
Carnival	2.5M	2227	2191	2750	2277	1121	1449	806	1478
	5M	4312	2196	3370	3567	2862	1531	778	1731
	10M	4306	3082	3704	4149	3718	2814	831	2154
Centi.	2.5M	3431	3127	3014	3251	3361	3102	2730	3009
	5M	3371	3505	3345	3821	3643	3258	2604	3186
	10M	4137	4078	3990	4114	3758	3619	2777	3865
Dm.Attk.	2.5M	1627	1585	1381	891	707	1614	176	1650
	5M	2590	2653	2583	1805	2093	2510	183	2585
	10M	4964	4065	4218	3590	3536	4742	190	3855
MsPac.	2.5M	960	923	997	871	1094	1177	628	1192
	5M	1271	1487	1369	1369	1370	1459	672	1448
	10M	1642	1881	1734	1826	1466	1809	679	1830
Phoenix	2.5M	2774	3187	3295	2905	3053	3616	777	3161
	5M	4326	4157	4198	4254	3960	4422	760	4309
	10M	4890	4785	4754	4825	4739	4946	688	4772
Pitfall	2.5M	-22	-185	-198	-228	-118	-17	-20	-3
	5M	-6	-66	-263	-47	-36	-4	-14	-11
	10M	-4	-29	-115	-44	-78	-31	-34	-4
Sp.Inv.	2.5M	435	514	507	436	483	512	155	520
	5M	550	568	505	582	540	565	172	587
	10M	703	696	586	768	652	686	172	673

Table 3: The average return over all seeds after the specified number of environment steps for all Atari games. The largest return in each row is shown in bold.

C TRAINING DETAILS

Teacher Training: Our training process for both domains first consists of training a teacher policy using standard PPO in the source environment using the hyperparameters for the appropriate domain (shown below). A total of four independent training runs were performed, with the best performing seed selected as the teacher policy. This teacher policy was then used for all subsequent transfer learning experiments in the corresponding domain.

Student Training: Each target task in the domain was trained using standard PPO without transfer in order to collect Baseline results. In Gridworld, the ten training runs with random seeds were run for 500K environment steps each while in Atari four training runs were run for 10M environment steps each. Each algorithm was then run for an equivalent number of independent random seeds utilizing the corresponding transfer learning algorithm and the previously trained teacher policy. For example, in Atari a total of 280 training runs were performed over all target tasks, 4 runs \times 7 total

algorithms \times 10 target tasks, while in Gridworld there were 70 runs for the target task. All student and teacher policies used the same model architecture (shown in Fig. 14) for a given domain. If a method was a fine-tuning variant described in Fig. 2c, then prior to training weights would be copied from the teacher policy (and additionally frozen if the Frozen model was employed). In the case of IAA on Atari, four seeds were run per Introspection Threshold for a total of 12 runs. The best performing threshold (as measured by average return after 10M steps) was selected as the IAA method of choice for the given target task, and the relative performance results were computed from the corresponding four student policies.

D HYPERPARAMETERS - GRIDWORLD

Hyperparameter	Value
Introspection Threshold (ϵ)	0.9
Burn-in (δ)	0
Introspection Decay (λ)	0.99999

Table 4: IAA Hyperparameters for Gridworld. These are used for all experiments unless otherwise stated.

Hyperparameter	Value
Horizon	128
Learning Rate	0.0005
Clipping Parameter	0.2
Batch Size	256
Minibatch Size	128
Num. SGD epochs	4
VF Clip Parameter	10.0
VF Coeff.	0.5
KL Coeff.	0.5
Entropy Coeff.	0.01
GAE Parameter	0.8
Discount	0.99

Table 5: PPO Hyperparameters for Gridworld.

E HYPERPARAMETERS - ATARI

Hyperparameter	Value
Introspection Threshold (ϵ)	{0.15, 0.30, 0.45}
Burn-in (δ)	500000
Introspection Decay (λ)	0.999999

Table 6: IAA Hyperparameters for Atari. These are used for all experiments unless otherwise stated.

Hyperparameter	Value
Horizon	128
Learning Rate	0.0005
Clipping Parameter	0.2
Batch Size	3200
Minibatch Size	256
Num. SGD epochs	4
VF Clip Parameter	10.0
VF Coeff.	1.0
KL Coeff.	0.5
Entropy Coeff.	0.01
GAE Parameter	0.95
Gradient Clip	0.1
Discount	0.99

Table 7: PPO Hyperparameters for Atari. Note that the learning rate was linearly annealed over time. Standard observation pre-processing for Atari was utilized: 4-frame stack, 84x84 grayscale image, etc.

F MODEL ARCHITECTURE

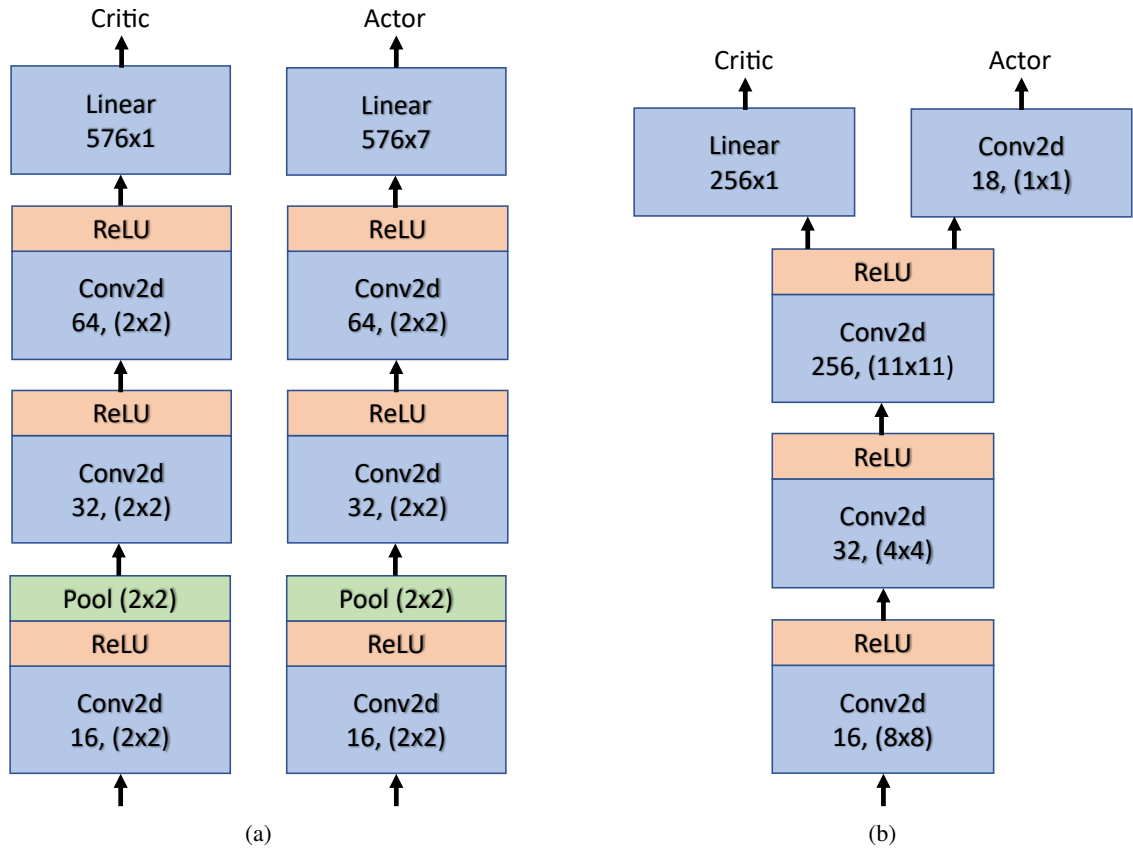


Figure 14: Model architectures used in experiments for (a) Gridworld and (b) Atari. The number out of out channels and kernel is given for each Conv2d layers. Linear layers represent input nodes by output nodes.