Mutual Information Based Knowledge Transfer Under State-Action Dimension Mismatch

Michael Wan  Tanmay Gangwani  Jian Peng

UAI 2020
Motivation

• Reinforcement Learning is **sample-inefficient**: often millions of samples required to learn a good policy.

• Environment interaction may be expensive
  • True of real-world robotics. Risk of causing damage to hardware/surroundings.
Transfer Learning in RL

• Assume access to a pre-trained teacher policy
  • Trained in source MDP with state space $S_{src}$, action space $A_{src}$
  • Teacher policy network $\pi_{\theta'}$, value network $V_{\psi'}$

• Train student policy in target MDP with state space $S_{targ}$, action space $A_{targ}$
  • Student policy network $\pi_{\theta}$, value network $V_{\psi}$

• Ideally, teacher should help accelerate the student learning
Transfer Learning – MDP Mismatch

• Prior work has considered setting where $S_{src} = S_{targ}$, $A_{src} = A_{targ}$
• What if $S_{src} \neq S_{targ}$, $A_{src} \neq A_{targ}$?

• How to handle difference in state space?
• How to handle difference in action space?

Teacher (6-legged Centipede)
State dimension : 139
Action dimension : 16

Transfer Learning

Student (4-legged Ant)
State dimension : 111
Action dimension : 8
Transfer Learning – MDP Mismatch

• How to handle difference in state space?
  • An embedding space learned through a network
  • This network acts a conduit between $S_{targ}$ and $S_{src}$

• How to handle difference in action space?

Teacher (6-legged Centipede)
State dimension : 139
Action dimension : 16

Transfer Learning

Student (4-legged Ant)
State dimension : 111
Action dimension : 8
Embedding Space

• Introduce learned embedding space
  • Parameterized by encoder function $\phi(\cdot)$
  • Defined as $S_{emb} := \{\phi(s) \mid s \in S_{targ}\}$

• Dimension of embedding space must match state-space dimension of source MDP: $|S_{emb}| = |S_{src}|$

• Can now utilize the teacher networks for knowledge transfer!
Embedding Space Desiderata

• **Desired property 1:** Embeddings must be task aligned
  • Embedding parameters should be updated to maximize the cumulative discounted rewards in the target MDP

• **Desired property 2:** Embeddings must have high correlation with input states in target MDP
  • We propose to maximize a lower bound to the mutual information (MI) between state $s_{\text{targ}}$ and embedding $\phi(s_{\text{targ}})$
Embeddings with Mutual Information Maximization

• Lower bound to MI between state $s$ and embedding $e$:

$$\mathcal{H}(s) + \mathbb{E}_{s, e} [\log q_\omega(s | e)]$$

• $q_\omega(s | e)$: variational distribution
  • Neural network that outputs mean of a multivariate Gaussian
  • Learned diagonal covariance matrix

• Loss: $L_{MI}(\phi, \omega) = -\mathbb{E}_{s \sim \rho_{\pi_\theta}} [\log q_\omega(s | \phi(s))]$
  • $\rho_{\pi_\theta}$: state-visitation distribution
  • Entropy $\mathcal{H}(s)$ is constant w.r.t. encoder parameters $\phi$ and variational parameters $\omega$, so we can omit it

Transfer Learning – MDP Mismatch

• How to handle difference in state space?
  • An embedding space learned through a network
  • This network acts a conduit between $S_{\text{targ}}$ and $S_{\text{src}}$

• How to handle difference in action space?

Teacher (6-legged Centipede)
State dimension : 139
Action dimension : 16

Transfer Learning

Student (4-legged Ant)
State dimension : 111
Action dimension : 8
Knowledge Transfer

• **Idea:** **Augment representations of student with teacher representations**

• Feed current state $s_{\text{targ}} \in S_{\text{targ}}$ into student networks, embedding $\phi(s_{\text{targ}})$ into teacher networks

• Weighted linear combination at each layer $j$:

$$h_{\pi_\theta}^j = \sigma(p_\theta z_\theta^j + (1 - p_\theta)z_{\theta'}^j)$$

$$h_{V_\psi}^j = \sigma(p_\psi z_\psi^j + (1 - p_\psi)z_{\psi'}^j)$$

**Notation**

- $\sigma$: activation function
- $z$: pre-activations
- $p$: mixing weights

---

Mixing Weights

- Lower values indicate heavier dependence on teacher representations

- Encourage student to be independent of the teacher by end of training
  
  \[ L_{coupling} = -\frac{1}{N_\pi} \sum_{j=1}^{N_\pi} \log(p_\theta^j) - \frac{1}{N_V} \sum_{j=1}^{N_V} \log(p_V^j) \]

Mutual Information based Knowledge Transfer

### Complete Algorithm (MIKT)
- Can incorporate into any base RL algorithm; we choose PPO
- Update $\phi$ with $\nabla_\phi [L_{MI}(\phi, \omega) + L_{PPO}(\theta, \psi, \theta', \psi', \phi)]$
- Update $\theta, \psi$ with $\nabla_{\theta, \psi} L_{PPO}(\theta, \psi, \theta', \psi', \phi)$
- Update $\omega$ with $\nabla_\omega L_{MI}(\phi, \omega)$
- Update $\{p\}$ with $\nabla_p [L_{coupling} + L_{PPO}]$
Experimental Setup

- MuJoCo locomotion tasks (Ant, Centipede$^1$)
- Centipede tasks differ in number of legs and disability (Cp variants have some legs disabled)

<table>
<thead>
<tr>
<th>Environment</th>
<th>State Dimension</th>
<th>Action Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentipedeFour</td>
<td>97</td>
<td>10</td>
</tr>
<tr>
<td>CentipedeSix</td>
<td>139</td>
<td>16</td>
</tr>
<tr>
<td>CentipedeEight</td>
<td>181</td>
<td>22</td>
</tr>
<tr>
<td>CpCentipedeSix</td>
<td>139</td>
<td>12</td>
</tr>
<tr>
<td>CpCentipedeEight</td>
<td>181</td>
<td>18</td>
</tr>
<tr>
<td>Ant</td>
<td>111</td>
<td>8</td>
</tr>
</tbody>
</table>

1. NerveNet: Learning Structured Policy with Graph Neural Networks, Wang et al.
Results

- Methods
  - MIKT *(ours)*
  - VPG: PPO on target task (no transfer learning)
  - MLPP: Re-use middle layers of pre-trained network, randomly initialize input and output layers
Encoder Ablation

• Are gradients from both $\{L_{PPO}, L_{MI}\}$ to the encoder beneficial?
• MIKT w/o MI: encoder does not receive gradients from $L_{MI}$
• MIKT w/o RL gradients: encoder does not receive gradients from $L_{PPO}$
Task Similarity Ablation

- How does task similarity impact MIKT?
- We experiment with transfer from Centipede-{Four, Six} to CentipedeEight (source and target tasks are similar) and transfer from Hopper (source and target tasks are different)

Student learns to trust its own representations rather than dissimilar Hopper teacher’s knowledge
Summary

• MIKT enables transfer learning between MDPs with different state and action spaces

• Learned embedding space
  • Mutual Information maximization: teacher representations depend on current state in the student MDP
  • Task aligned: encoder trained to maximize cumulative discounted reward

• Knowledge Transfer
  • Augment student representations with teacher representations