Policy Distillation

Muhammad Uzair
Overview

• Introduction
• Deep Q-Learning
• Single-Game Policy Distillation
• Multi-Task Policy Distillation
• Results
  • Single-Game Policy Distillation
  • Policy Distillation with Model Compression
  • Multi-Game Policy Distillation
  • Online Policy Distillation
Introduction

• Transfer of policy from Q-networks to an untrained network

• Advantages:
  • Compressed network size without degradation in performance
  • Combining multiple expert policies into a single multi-task policy
  • Real-time, online learning process

• Why?
  • DQN needs large networks and extensive training
Deep Q-Learning

• Reward based
• Requires long training times
• The Q function gives the maximum expected return:

\[ Q^*(s, a) = \max_\pi \mathbb{E}[R_t|s_t = s, a_t = a, \pi] \]

• The following loss is minimize to train a convolutional neural network:

\[ L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim U(D)} \left[ \left( r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i) \right)^2 \right] \]
Single-Game Policy Distillation

• Distillation is a method of transferring knowledge from teacher $T$ to student $S$

• Softmax function

• To transfer more knowledge, a relaxed Softmax function is used

• For a selected temperature $t$, the new teacher outputs are given by $\text{Softmax}(q^T/t)$. $q^T$ is the vector of Q-values of the $T$. 
Single-Game Policy Distillation
Single-Game Policy Distillation

• Problems:
  • Predicting Q-values of all actions is a difficult regression task
  • Scale of Q-values can be quite unstable
  • Computationally challenging
  • Training of S to predict only the single best action is also problematic
Single-Game Policy Distillation

- Negative Log Likelihood Loss (NLL)
  - Uses only the highest valued action from the teacher

\[
L_{\text{NLL}}(D^T, \theta_S) = - \sum_{i=1}^{|D|} \log P(a_i = a_{i,\text{best}} | x_i, \theta_S)
\]
Single-Game Policy Distillation

• Mean-Squared-Error Loss (MSE)
  • It preserves the full set of action-values in the resulting student model

$$L_{MSE}(D^T, \theta_S) = \sum_{i=1}^{|D|} ||q^T_i - q^S_i||_2^2.$$
Single-Game Policy Distillation

• Kullback-Leibler Divergence (KL)

\[
L_{KL}(\mathcal{D}^T, \theta_S) = \sum_{i=1}^{D} \text{softmax}(\frac{q_i^T}{\tau}) \ln \frac{\text{softmax}(\frac{q_i^T}{\tau})}{\text{softmax}(q_i^S)}
\]
Multi-Task Policy Distillation
Multi-Task Policy Distillation

• Compare the performance of multi-task DQN with multi-task distilled agents

• Multi-game DQN learning is extremely challenging for Atari games
  • Different policies, different reward scaling, instability of learning value functions

• Policy distillation may combine multiple policies into single network
Results

• Training and Evaluation
  • Separate DQN agent for each game
  • Same network used to train student as DQN
  • Scaled down the number of units in compression experiments
  • A large network for multi-task distillation
  • Professional human expert play was used to generate starting states

• Ten popular Atari games
Results

- Single-Game Policy Distillation Results

<table>
<thead>
<tr>
<th></th>
<th>DQN</th>
<th>Dist-MSE</th>
<th>%DQN</th>
<th>Dist-NLL</th>
<th>%DQN</th>
<th>Dist-KL</th>
<th>%DQN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakout</td>
<td>score</td>
<td>score</td>
<td>%DQN</td>
<td>score</td>
<td>%DQN</td>
<td>score</td>
<td>%DQN</td>
</tr>
<tr>
<td>303.9</td>
<td>102.9</td>
<td>33.9</td>
<td></td>
<td>235.9</td>
<td>77.6</td>
<td>287.8</td>
<td>94.7</td>
</tr>
<tr>
<td>Freeway</td>
<td>25.8</td>
<td>25.7</td>
<td>99.4</td>
<td>26.2</td>
<td>101.4</td>
<td>26.7</td>
<td>103.5</td>
</tr>
<tr>
<td>Pong</td>
<td>16.2</td>
<td>15.3</td>
<td>94.4</td>
<td>15.4</td>
<td>94.9</td>
<td>16.3</td>
<td>100.9</td>
</tr>
<tr>
<td>Q*bert</td>
<td>4589.8</td>
<td>5607.3</td>
<td>122.2</td>
<td>6773.5</td>
<td>147.6</td>
<td>7112.8</td>
<td>155.0</td>
</tr>
</tbody>
</table>
Results

• Policy Distillation with Model Compression
Results

• Multi-Game Policy Distillation Results
## Results

- Multi-Game Policy Distillation Results

<table>
<thead>
<tr>
<th>Game</th>
<th>DQN</th>
<th>Multi-Dist-KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beamrider</td>
<td>8672.4</td>
<td>4789.0</td>
</tr>
<tr>
<td>Breakout</td>
<td>303.9</td>
<td>216.0</td>
</tr>
<tr>
<td>Enduro</td>
<td>475.6</td>
<td>613.0</td>
</tr>
<tr>
<td>Freeway</td>
<td>25.8</td>
<td>26.6</td>
</tr>
<tr>
<td>Ms.Pacman</td>
<td>763.5</td>
<td>681.8</td>
</tr>
<tr>
<td>Pong</td>
<td>16.2</td>
<td>16.1</td>
</tr>
<tr>
<td>Q*bert</td>
<td>4589.8</td>
<td>6098.3</td>
</tr>
<tr>
<td>Seaquest</td>
<td>2793.3</td>
<td>4320.7</td>
</tr>
<tr>
<td>Space Invaders</td>
<td>1449.7</td>
<td>461.1</td>
</tr>
<tr>
<td>Riverraid</td>
<td>4065.3</td>
<td>4326.8</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

• Online Policy Distillation Results