Using Relative Novelty to Identify Useful Temporal Abstractions in Reinforcement Learning

Özgür Şimşek, Andrew G. Barto

Comp -767 Reinforcement Learning

Claudio Sole

The problem

- Planning, acting and learning at multiple levels of temporal abstraction can dramatically improve capabilities of autonomous agents
- To fully realize possible benefits, useful temporal abstractions should be automatically detected by the agent
- Purpose of the study:
  - Define the properties a subgoal should have
  - Based on these properties, classify states as subgoal or not

Definitions

- **Access states**:

  states that allow the agent to transition to a part of the state space that is otherwise unavailable or difficult to reach from its current region

- **Novelty** of a discrete state *s* **=** $\boldsymbol{\frac{1}{\sqrt{n_s}}}$, with n_s number of time a state *s* has been visited

- **Novelty of a set S** of states $= \frac{1}{\sqrt{\bar{n}_S}}$ with $\bar{n}_S$ average number of times states in S has been visited

Relative Novelty Score

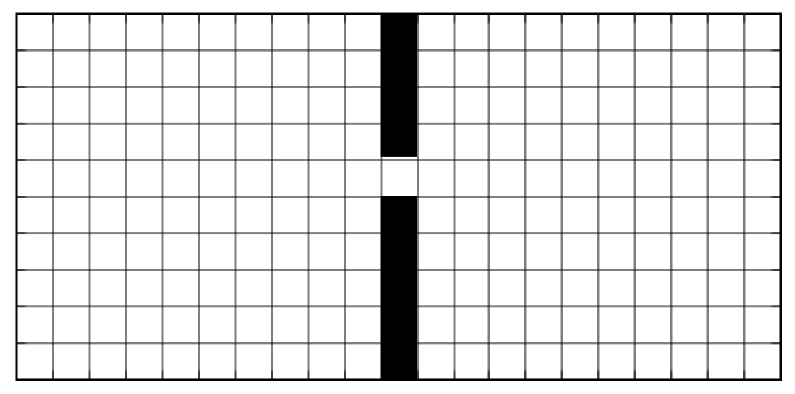
Given a transition sequence, with define the relative novelty of a state s as

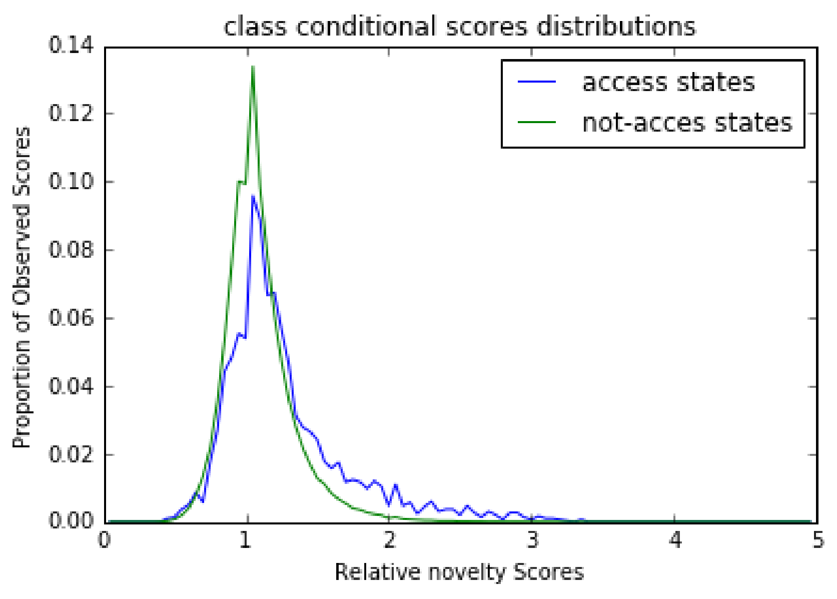
$$RN(s) = \frac{Set\ Novelty\ (states\ following\ s)}{Set\ Novelty\ (states\ precediing\ s)}$$

- Idea: the distribution of relative novelty scores of *access states* will be different than that of *non-access states*

Relative Novelty scores distributions

- We perform 1000 times a 1000-steps random walk, ignoring the goal state, on a two-rooms environment





Classifying subgoals(1)

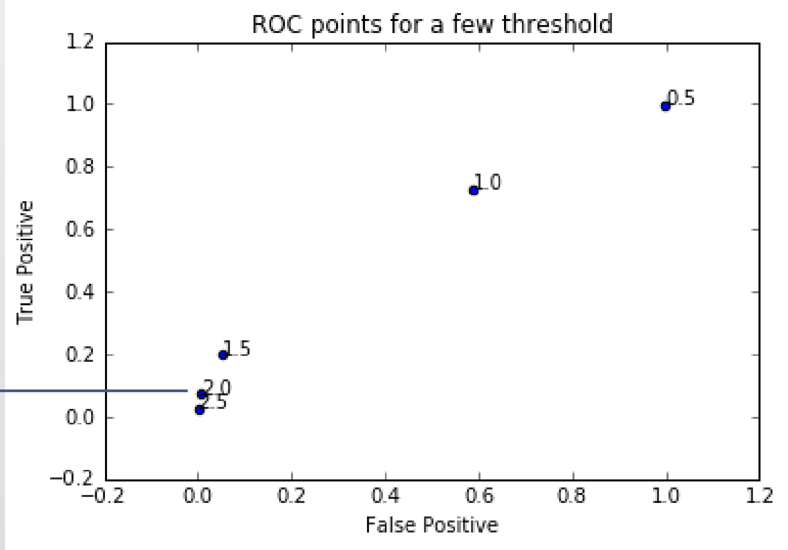
- For each state, we record *all* hi RN scores and, based on these, be classify it as target or not.
- Bayesian approach: label *s* as target if

$$\frac{P\{(s_1,\ldots,s_n)|T\}}{P\{(s_1,\ldots,s_n)|N\}} > \frac{\lambda_{fa}}{\lambda_{miss}} \frac{P\{N\}}{P\{T\}} \qquad (1)$$

- Where
  - λ_{fa} is the cost for a false positive, λ_{miss} is the cost for a false negative and *P{i} are the class priors*

Classifying subgoals(2)

- Transform the scores *s* to a binary feature x where
  - *x=1* if score >= t_{RN}
  - *x=0 otherwise*
- To establish t_{RN}, we perform a ROC analysis, keeping in mind that avoiding False Positives is much more important then trying to maximize the number of True Positives (otherwise we could end up with too many bad options, thus hurting the agent's exploration ability)
- We choose t_{RN}=2, which provide more TP then the 2.5 threshold but keeping FP rate close to 0% (see precedent distribution plot to get an intuition about the 2.0 value)



Classifying subgoals(3)

- Defining
  - *p = P{x=1|T}*
  - *q = P{x=1|N}*
  - *n number of scores for a given state*
  - n_1 *number of 1 in state scores*
- We can rewrite (1) as

$$\frac{p^{n_1}(1-p)^{n-n_1}}{q^{n_1}(1-q)^{n-n_1}} > \frac{\lambda_{fa}}{\lambda_{miss}}\frac{P\{N\}}{P\{T\}}$$

- And so, classify s as target if

$$\frac{n_1}{n} > \frac{ln\frac{1-q}{1-p}}{ln\frac{p(1-q)}{q(1-p)}} + \frac{1}{n}\frac{\ln\left(\frac{\lambda_{fa}}{\lambda_{miss}}\frac{p(N)}{p(T)}\right)}{ln\frac{p(1-q)}{q(1-p)}}$$

Remaining parameters

- *p* and *q* can be easily estimated from the class conditional distribution as
  - p= proportion of 1 in all target binary scores
  - q= proportion of 1 in all non-target binary scores

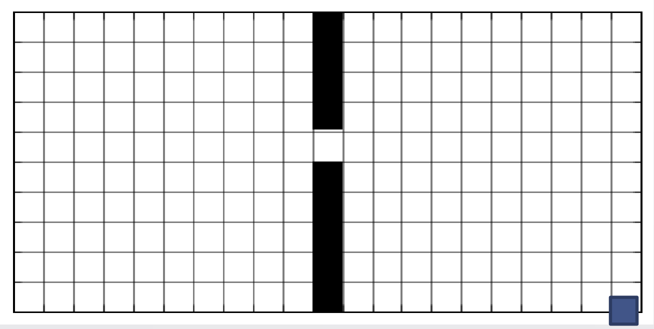
- For the ratios $\frac{\lambda_{fa}}{\lambda_{miss}}$ and $\frac{P\{N\}}{P\{T\}}$ we have two follow two guidelines:
  - The prior probability of target should be much smaller the that of non-target
  - Following the reasoning we made for t_{RN}, a false positive should have a much higher cost then a miss.

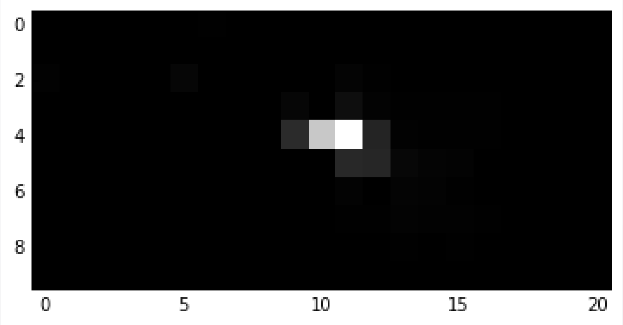
- Following these guidelines, the authors choose

$$\frac{\lambda_{fa}}{\lambda_{miss}} = 100, \qquad \frac{P\{N\}}{P\{T\}} = 100$$

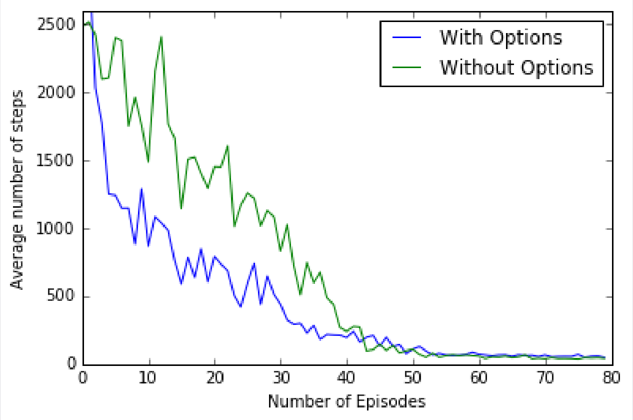
Results(1)

- Subgoals detected in 30 runs (for each episode, random start chosen in west region):



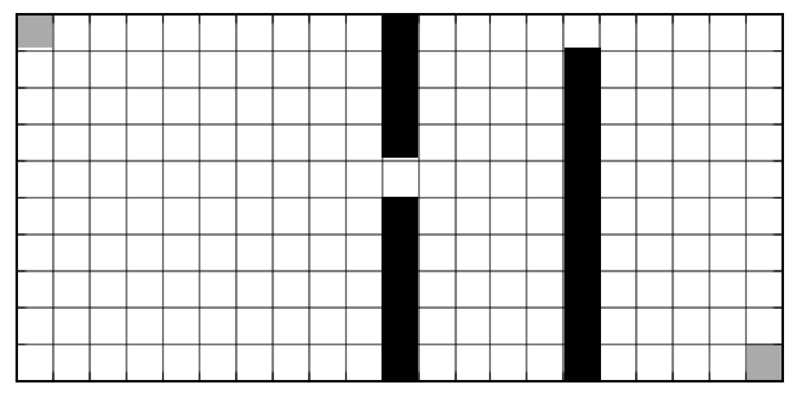


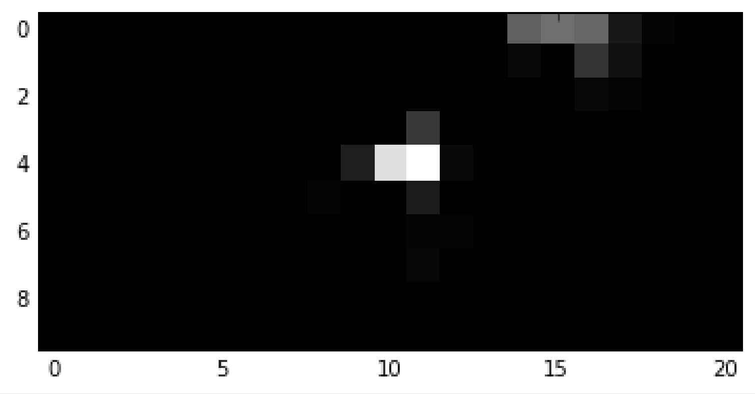
- Computational improvement:



Results(2)

- For the previous example, we trained and tested the classifier on the same environment. Anyway, the off-line procedure is supposed to work well also for similar environment (an alternative would to train the agent in a small part of the environment, if it is representative of the whole task)
- Frequency of subgoals detected in three-rooms environment during 30 runs





References

- Şimşek, Özgür, and Andrew G. Barto.
  "Using relative novelty to identify useful temporal abstractions in reinforcement learning."
  Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.