Exploration in RL

Prediction-based Intrinsic Exploration[1]

Marius Lindauer

[1]based on Blog by Lilian Weng

Prediction-based Exploration [Schmidhuber. 1991]

- Idea: If the agent is able to predict what will happen in the future, it is already well informed
- In contrast, if the agent is not able to predict the future, it is surprised.

$$f : (s_t, a_t) \mapsto s_{t+1}$$
$$e(s_t, a_t) = ||f(s_t, a_t) - s_{t+1}||_2^2$$

  - the higher the error e, the less familiar the agent is with that state / more surprised

Intelligent Adaptive Curiosity [Oudeyer et al. 2007]

- Memory of all observed state transitions $M = (s_t, a_t, s_{t+1})$
- Split the state space S similarly as in decision node:
  - Split only if enough states were observed
  - Variance of states in each leaf should be minimal
  - For each leaf, learn a forward dynamic predictor f
- Reward regions where we can make fast progress via decreasing error

$$r_t^i = \frac{1}{k} \sum_{i=0}^{k-1} (e_{t-i-\tau} - e_{t-i})$$

  - moving window with offset τ and moving window size k

Automated Machine Learning Hannover

Decay [Stadie et al. 2015]

- Normalize error to [0,1] by the maximal error observed so far

$$\bar{e}_t = \frac{e_t}{\max_{i \leq t} e_i}$$

- decay intrinsic reward over time

$$r_t^i = \frac{e_t(s_t, a_t)}{t \cdot C}$$

  - C being a constant

State Encoding

State should be encoded to reduce the state space

- hash-function [Tang et al. 2016]
- Autoencoder [Stadie et al. 2015]
- Self-supervised inverse dynamic models [Pathak et al. 2017]

$$g : (\phi(s_t), \phi(s_{t+1})) \mapsto a_t$$

  - The encoding will focus on the parts of state features that influenced the agent behavior
  - Use ϕ to learn a forward model $f : (\phi(s_t), a_t) \mapsto s_{t+1}$

$$r_t^i = ||f(\phi(s_{t+1})) - \phi(s_{t+1})||_2^2$$

Pure Curiosity-drive Learning [Burda et al. 2018]

- Pure exploration and ignoring the extrinsic reward signal

$$r_t = r_t^i = ||f(s_t, a_t) - \phi(s_{t+1})||_2^2$$

- Study on different state encodings: compact, sufficient and stable
  1. Raw image pixels; no encoding
  2. Neural network with fixed random weights
  3. VAE
  4. Inverse dynamics features (IDF)
- Insights:
  - No clear winner overall
  - Random network quite competitive
  - IDF can generalize better (e.g., learn IDF on one Super Mario Bros level and then test it on another)
  - On the noisy TV env, IDF performed best, followed by random network, but overall very hard to learn anything reasonable (wrt extrinsic reward)

Variational Information maximizing Exploration

[Houthooft et al. 2016]

- Idea: maximize information gain about the agent's belief of env dynamics.
- Information gain often measured by reduction in entropy

$$\begin{aligned} & \sum_t H(\Theta \mid e_t, a_t) - H(\Theta | s_{t+1}, e_t, a_t) \\ = \; & \mathbb{E}_{s_{t+1} \sim P(\cdot | e_t, a_t))} \left[D_{KL}(p(\theta \mid e_t, a_t, s_{t+1}) || p(\theta | e_t)) \right] \end{aligned}$$

  - $\theta \in \Theta$ parameterized forward dynamic model
  - e_t episode t
  - H entropy
  - D_{KL} Kullback-Leibler divergence

$\rightsquigarrow$ Use Bayesian Neural Network (BNN) for dynamics model which maintains a distribution over its weights θ

Automated Machine Learning Hannover

AutoML.org