Soft Actor Critic

Oliver, Leon Büttinghaus, Thilo Röthemeyer

18. April 2021

Contents

Part 1

Kontinuierlicher Aktionsraum

- kontinuierliche Aktionsräume benötigen
  $\Rightarrow$ Approximation für Q-Funktion
  $\Rightarrow$ Approximation für Strategie
- Schritt von Tabellen zu DNNs
- Optimierung mittels gradient descent

Funktionen und deren Netzwerke

- State Value Funktion:

  $V_\psi(s_t)$ $\rightarrow$ Skalar als Ausgabe

- Q-Funktion:

  $Q_\theta(s_t, a_t)$ $\rightarrow$ Skalar als Ausgabe

- Strategie:

  $\pi_\phi(s_t|a_t)$ $\rightarrow$ Mittelwert und Kovarianz als Ausgabe $\Rightarrow$ Gauss

Mit Parametervektoren ψ, θ und ϕ

State Value Funktion

- eigenes Netzwetk nicht notwendig, aber
  - stabilisiert Training
  - macht simultanes Training aller Netzwerke möglich

Optimierung State Value Funktion

- Minimierung des Fehlers
- Fehler: Residuenquadratsumme aus state-value und Erwartungswert

$$J_V(\psi) = \mathbb{E}_{s_t \sim D}\left[\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi}[Q_\theta(s_t, a_t) - log\pi_\phi(s_t|a_t)])^2\right]$$

$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(s_t)(V_\psi(s_t) - Q_\theta(s_t, a_t) + log\pi_\phi(s_t|a_t))$$

Optimierung Q-Funktion

- Minimierung des Fehlers
- Fehler: soft Bellman Restwert

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}\left[\frac{1}{2}(Q_\theta(s_t,a_t) - \hat{Q}_\theta(s_t,a_t))^2\right]$$

mit $\hat{Q}(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim p}[V_{\overline{\psi}}(s_{t+1})]$

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(a_t,s_t)(Q_\theta(s_t,a_t) - r(s_t,a_t) - \gamma V_{\overline{\psi}}(s_{t+1}))$$

Optimierung der Strategie

- Minimierung des Fehlers
- Fehler: KL-Divergenz

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D}\left[D_{KL}\left(\pi_\phi(\cdot|s_t)||\frac{exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)}\right)\right]$$

- der reparameterization trick wird angewandt

$$a_t = f_\phi(\epsilon_t; s_t)$$
$$\Rightarrow J_\pi(\phi) = \mathbb{E}_{s_t \sim D, \epsilon_t \sim N}\left[log\pi_\phi(f_\phi(\epsilon_t; s_t)|s_t) - Q_\theta(s_t, f_\phi(\epsilon_t; s_t))\right]$$

Algorithmus (1/2)

Algorithm 1: Soft Actor-Critic

Initialize parameter vectors $\psi, \overline{\psi}, \theta, \phi$
for *each iteration* **do**
  for *each environment step* **do**
    $a_t \sim \pi_\phi(a_t|s_t)$
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    $D \leftarrow D \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
  end
  for *each gradient step* **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\overline{\psi} \leftarrow \tau\psi + (1-\tau)\overline{\psi}$
  end
end

Algorithmus (2/2)

Ziel der Experimente

- Stabilität und Sample Komplexität im Vergleich zu anderen Algorithmen
  - Kontinuierliche Aufgaben
  - Verschiedene Schwierigkeitgrade
- OpenAI gym und rllab

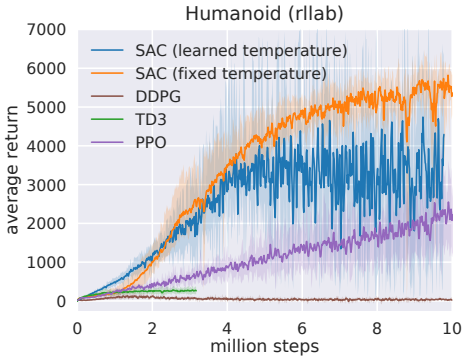
Vergleich zu anderen Algorithmen

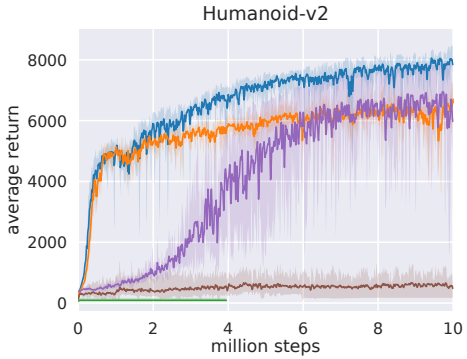
- SAC
  - Durchschnittswert (mean action)
  - feste und variable Temperatur(Anpassung im neuen Paper)
- PPO, DDPG
  - kein Explorationsrauschen
- TD3
- SQL mit zwei Q Funktionen
  - Evaluation mit Explorationsrauschen

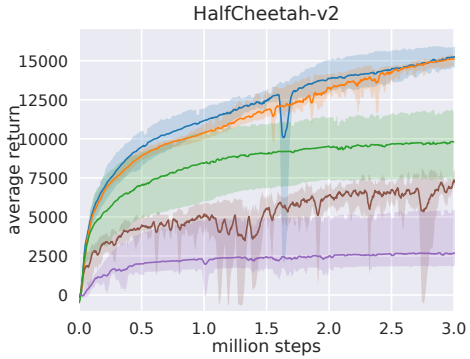
Vergleich zu anderen Algorithmen

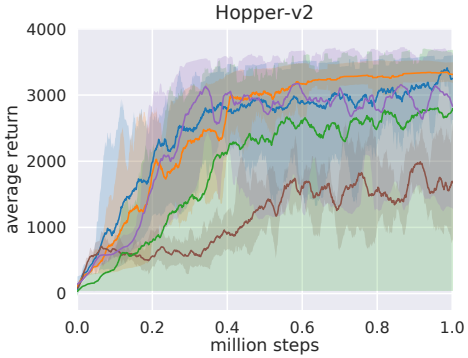
- 5 Instanzen mit einer Evaluation alle 1000 Schritte
- Schattierter Verlauf zeigt min und max der fünf Durchläufe

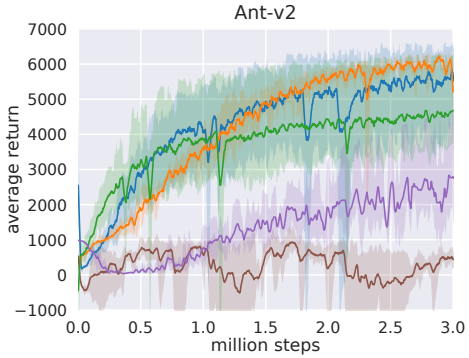
Ergebnisse

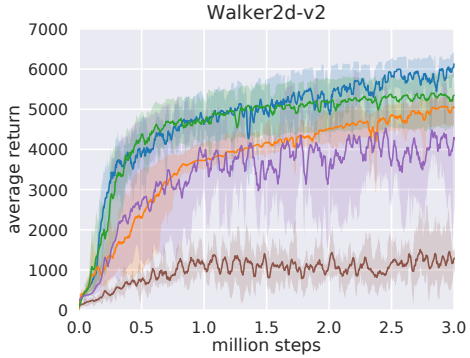

Humanoid (rllab)
SAC (learned temperature)
SAC (fixed temperature)
DDPG
TD3
PPO
average return
million steps


Humanoid-v2
average return
million steps


HalfCheetah-v2
average return
million steps


Hopper-v2
average return
million steps


Ant-v2
average return
million steps


Walker2d-v2
average return
million steps

Zusammenfassung

- soft actor critic vorgestellt
  - Off policy Algorithmus
  - Entropiemaximierung verbessert Stabilität
  - Besser als state-of-the-art Algorithmen
  - Gradientenbasiertes Temperatur Tuning


Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine.
Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.
CoRR, abs/1801.01290, 2018.