

Soft Actor Critic

Oliver, Leon Büttinghaus, Thilo Röthemeyer

18. April 2021

Contents

- 1 Part 1
- 2 Soft Actor-Critic im kontinuierlichen Raum
 - SAC Grundprinzip
 - SAC Update Regeln
 - SAC Algorithmus
- 3 Ergebnisse
 - Vergleich mit anderen Algorithmen
 - Zusammenfassung
- 4 Literaturverzeichnis

Part 1

Kontinuierlicher Aktionsraum

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

Funktionen und deren Netzwerke

- State Value Funktion:

$V_\psi(s_t)$ → Skalar als Ausgabe

- Q-Funktion:

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

- Strategie:

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

Mit Parametervektoren ψ , θ und ϕ

State Value Funktion

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

Optimierung State Value Funktion

Q-Funktion

Optimierung Q-Funktion

Optimierung der Strategie

Algorithmus (1/2)

Algorithmus (2/2)

Ziel der Experimente

- Stabilität und Sample Komplexität im Vergleich zu anderen Algorithmen
 - kontinuierliche Aufgaben
 - verschiedene Schwierigkeitsgrade
- OpenAI gym und rllab

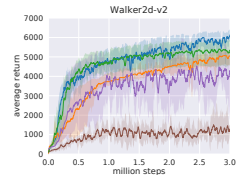
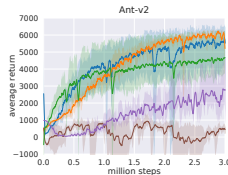
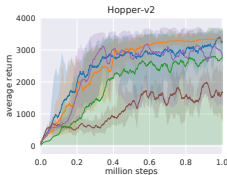
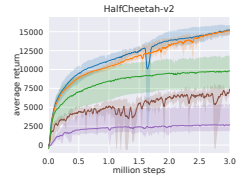
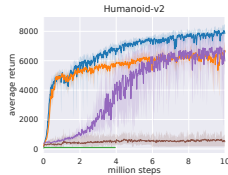
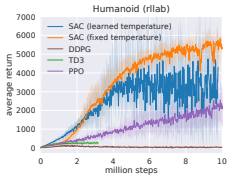
Vergleich zu anderen Algorithmen

- 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

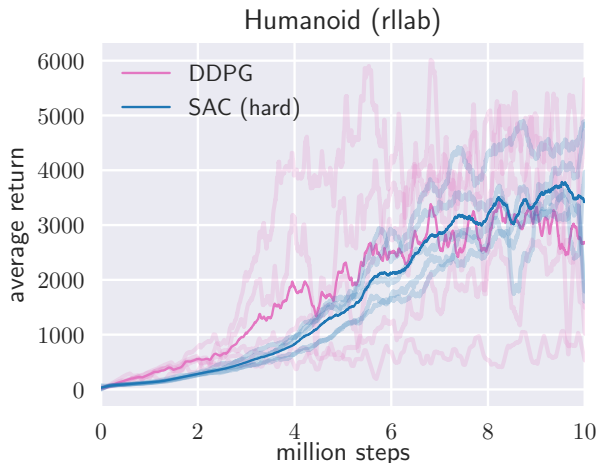
Vergleich zu anderen Algorithmen

- 5 Instanzen mit einer Evaluation alle 1000 Schritte
- Total average return shown in the following
- Schattierter Verlauf zeigt min und max der fünf Durchläufe

Ergebnisse



Ergebnisse



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.