

Soft Actor Critic

Oliver, Leon Büttinghaus, Thilo Röthemeyer

15. April 2021

Contents

1 SAC Grundprinzip

2 SAC Update Regeln

3 SAC Algorithmus

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 ⇒ 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)

Part 2

Conclusion

- Performance im Vergleich zu TD3 und A2C
- Vor- und Nachteile von SAC

Performance

- SAC soll stabiler laufen als TD3 Algorithmen
- Testumgebung gym

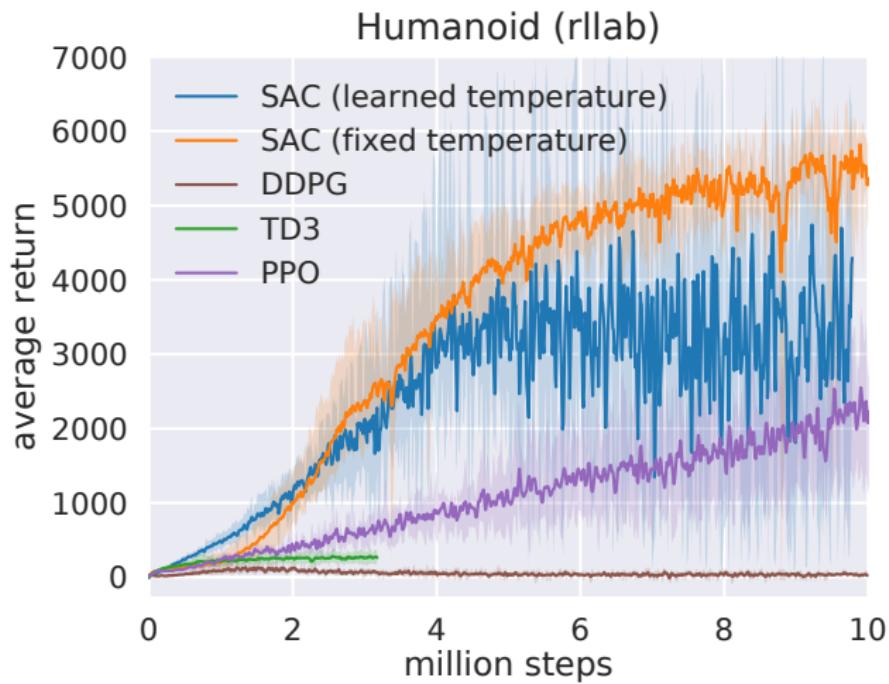
Ziel der Experimente

- Stabilität und Sample Komplexität im Vergleich zu anderen Algorithmen
 - kontinuierliche Aufgaben
- OpenAI gym und rllab
 - schwierig für off-policy Algorithmen
 - Parametertuning bei komplexeren Aufgaben schwierig

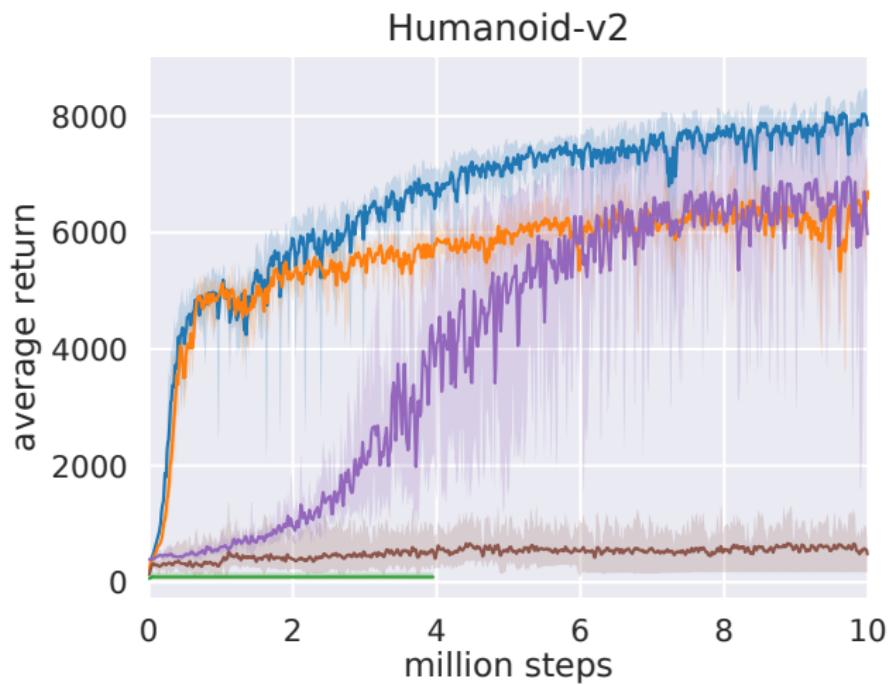
Vergleich zu anderen Algorithmen

- SAC
 - mean action
- DDPG
- PPO
- SQL mit zwei Q Funktionen
 - evaluated with exploration noise
- TD3

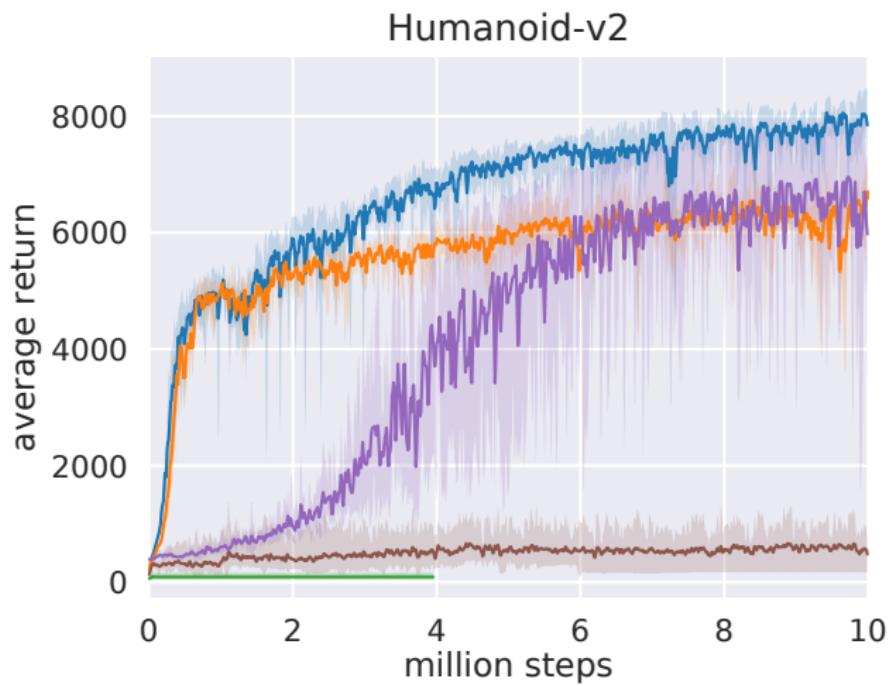
Ergebnisse



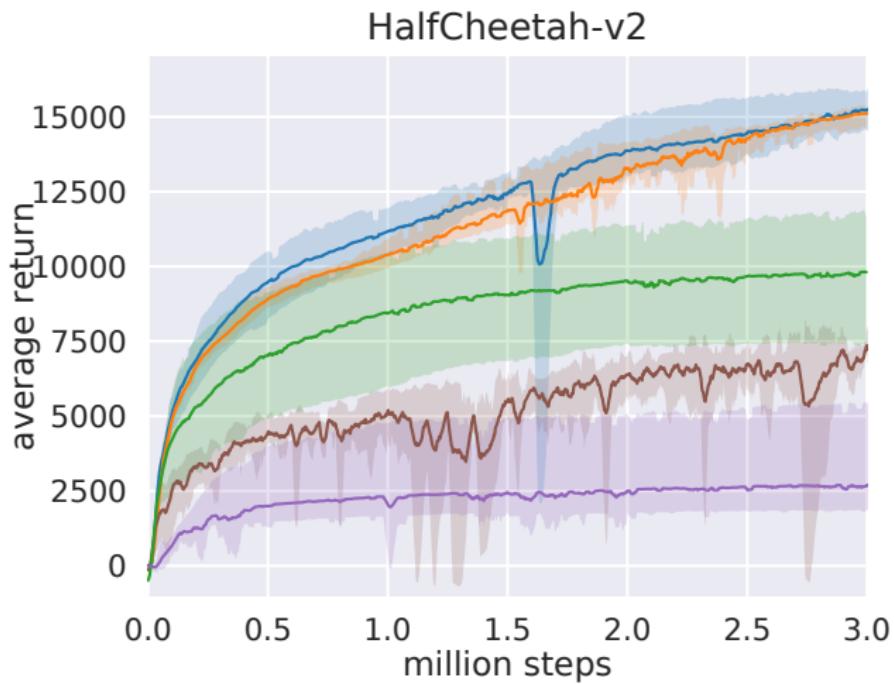
Ergebnisse



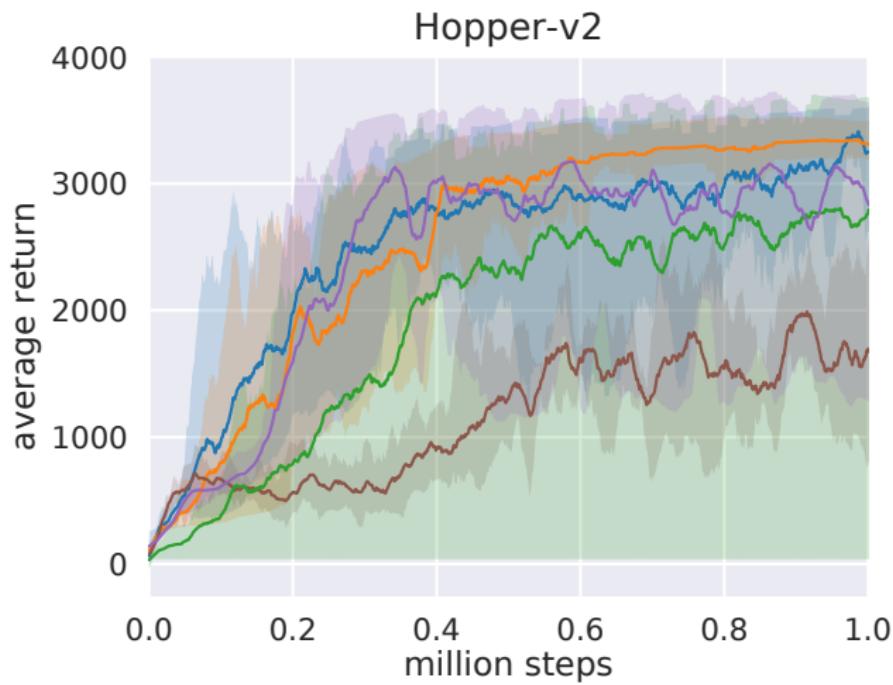
Ergebnisse



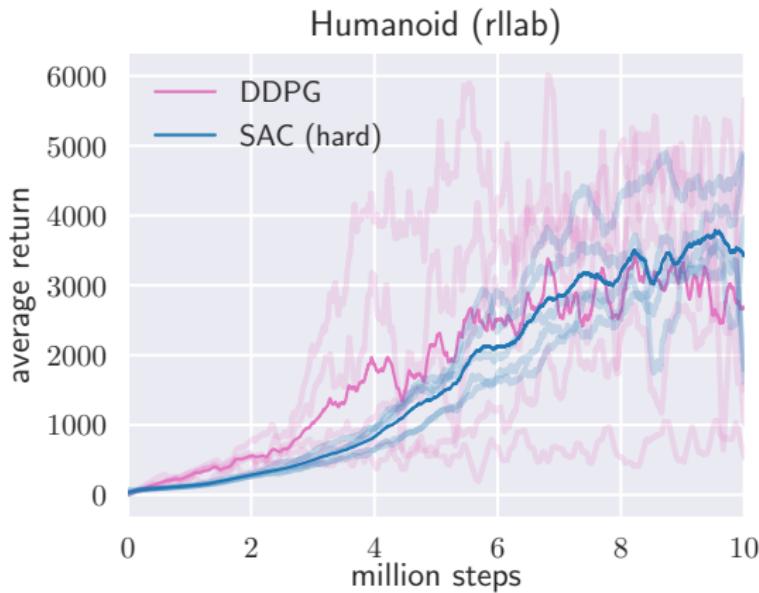
Ergebnisse



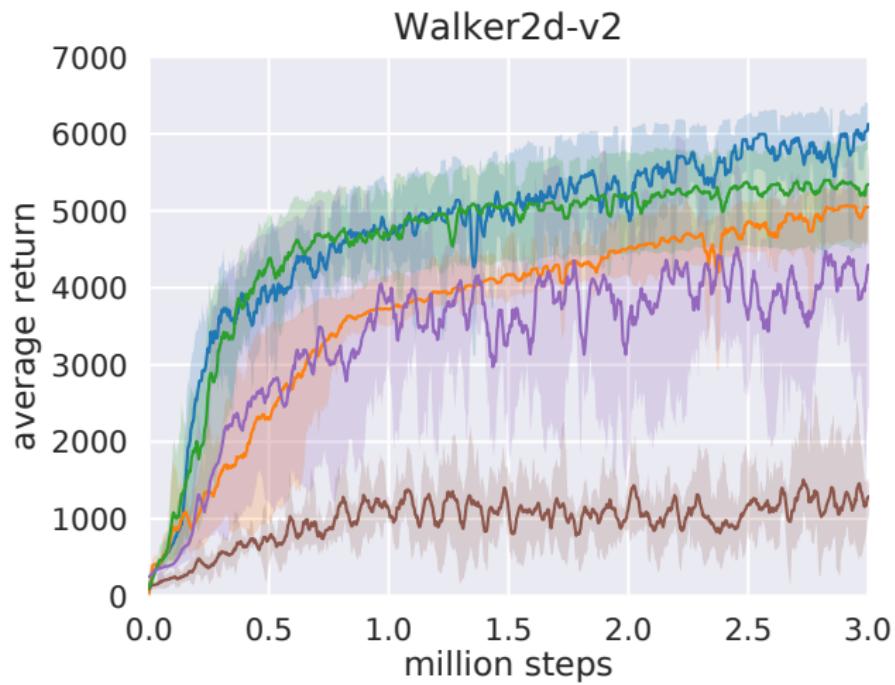
Ergebnisse



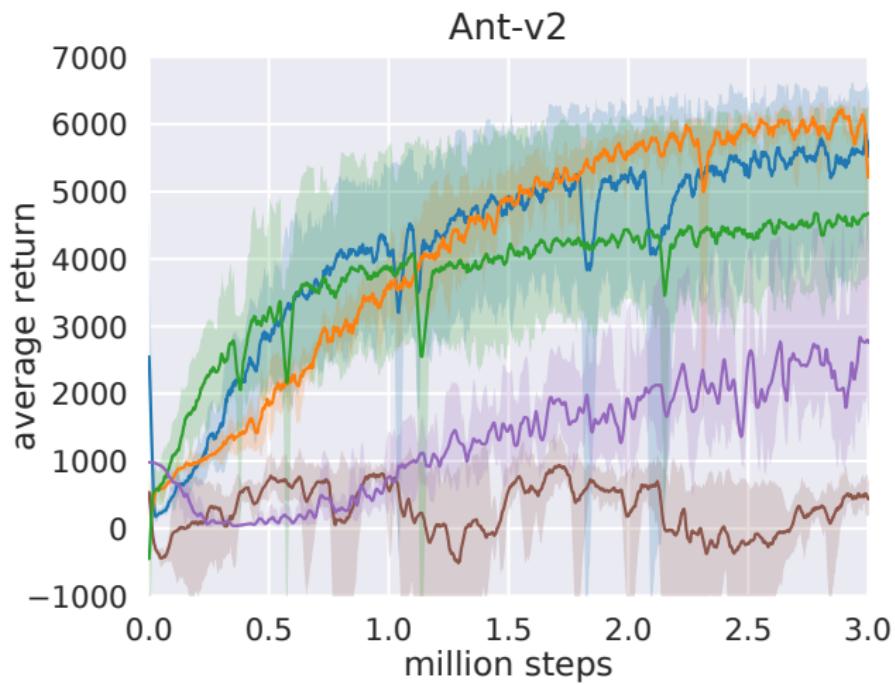
Ergebnisse



Ergebnisse



Ergebnisse



Zusammenfassung

- soft actor critic vorgestellt
 - off policy
 - Entropiemaximierung verbessert Stabilität
 - Besser als state of the art Algorithmen