Universidad del Valle de Guatemala
Deep Learning
Javier Prado - 21486
Sebastian Juarez - 21471
Bryan España - 21550

Laboratorio 7

Task 2 - Teoría

Defina en qué consiste y en qué clase de problemas se pueden usar cada uno de los siguientes acercamientos en Deep Reinforcement Learning

1. Proximal Policy Optimization
   Es un algoritmo on-policy basado en gradientes de política en el cual este usa un clippinig en la funcion objetivo para evitar actualizaciones demasiado grandes en el cual estas actualizaciones harían inestable el aprendizaje, y es más eficiente y fácil de implementar. Este proximal policy optimization se usa en problemas de decisión secuencial con observaciones de alta dimensión (imágenes, sensores) y acciones discretas o continuas, donde se requiere estabilidad en el entrenamiento sin complicar la implementación. Es ideal en control robótico (torques/posiciones continuas), videojuegos y simuladores (Atari, MuJoCo, Unity), navegación y tareas de locomoción, así como en optimización de políticas en entornos estocásticos con recompensas ruidosas.
2. Deep Deterministic Policy Gradients (DDPG)
   Es un algoritmo **off-policy** de **actor-critic** para espacios de **acciones continuas**. Aprende una **política determinista** $\pi\theta(s)$ (actor) que mapea estados a acciones y un **crítico** $Q\phi(s,a)$ que estima el valor de tomar una acción en un estado. Para estabilizar, usa **replay buffer** (reutilizar transiciones des-correlacionadas) y **target networks** (copias lentas del actor y crítico). La **exploración** se logra inyectando **ruido** a la acción del actor durante la recolección de experiencias. Es bueno implementarlo cuando las **acciones son continuas** y de **control fino**: robótica (torques/articulaciones), locomoción y equilibrio (brazos, drones), control industrial (válvulas, motores), y subproblemas de conducción autónoma (dirección/aceleración).
3. Trust Region Policy Optimization (TRPO)
   Es un método on-policy de gradiente de políticas que actualiza la política maximizando una función objetivo sujeta a una restricción de "región de confianza" medida con divergencia KL entre la política nueva y la anterior. Esa cota limita cambios bruscos en la política y ofrece mejoras monotónicas (en teoría) y estabilidad durante el entrenamiento. En la práctica, TRPO resuelve un problema cuadrático aproximado (o equivalente) para encontrar el paso de actualización que maximiza el rendimiento sin salirse de la región de confianza, usualmente con conjugate gradient y line search. TRPO es adecuado cuando la estabilidad y la conservación de la política son críticas: control continuo de alta dimensión (robótica, locomoción), simulaciones delicadas (drones, vehículos), y entornos ruidosos o no estacionarios donde pasos grandes arruinan el aprendizaje. Se usa como baseline fuerte en benchmarks (MuJoCo, control

clásico) y en escenarios donde se valora garantía teórica sobre facilidad de implementación.

4. Asynchronous Advantage Actor-Critic (A3C)
   Es un método on-policy que entrena múltiples agentes en paralelo, cada uno interactuando con su propio entorno y actualizando de forma asincrónica una red global compartida. Combina un actor (política) y un crítico (valor) y utiliza la ventaja $A=Q-V$ para reducir varianza del gradiente. Emplea n-step returns, entropía para fomentar exploración y no requiere replay buffer; la asincronía descorrelaciona experiencias y estabiliza el aprendizaje, aprovechando bien CPU multinúcleo con baja sobrecarga de memoria. A3C destaca cuando necesitas acelerar el entrenamiento con paralelismo en simuladores (Atari, entornos 3D como VizDoom/DM Lab/Unity), en tareas de navegación, locomoción o control con acciones discretas o continuas y observaciones de alta dimensión.

Referencia

- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms*. arXiv. https://arxiv.org/abs/1707.06347
- GeeksforGeeks. (s. f.). *Asynchronous Advantage Actor-Critic (A3C) Algorithm*. Recuperado de https://www.geeksforgeeks.org/machine-learning/asynchronous-advantage-actor-critic-a3c-algorithm/#:~:text=The%20Asynchronous%20Advantage%20Actor-Critic%20%28A3C%29%20algorithm%20marked%20a,learning%20by%20introducing%20a%20unique%20parallel%20training%20strategy
- Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). *Trust Region Policy Optimization*. arXiv. https://arxiv.org/abs/1502.05477