UNIVERSIDAD DEL VALLE DE GUATEMALA

CC3092 - Deep Learning y Sistemas Inteligentes

Sección 21

Ing. Luis Alberto Suriano

Laboratorio No. 7

José Pablo Orellana	21970
Diego Alberto Leiva	21752

GUATEMALA, 29 de septiembre del 2024

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**

    - El PPO es un método de optimización que se usa para entrenar políticas en entornos de aprendizaje por refuerzo, lo que busca el POO es mejorar la estabilidad del entrenamiento al poner límite en el tamaño del cambio en la política de un paso a otro, es una mejora de otros métodos de optimización que también se basan en políticas, como Trust Region Policy Optimization, pero en este caso es más simple su implementación y computacionalmente hablando, es más eficiente.

    - El PPO es óptimo para la conducción autónoma, donde se tienen que tomar decisiones muy complejas en un espacio continuo de acciones como lo son las acciones de manejo, acelerar, girar, frenar, entre otras. El punto es que es crucial que las actualizaciones de política no sean demasiado agresivas porque podría causar un comportamiento indebido o errático.

2. **Deep Deterministic Policy Gradients (DDPG)**

    - El DDPG es un método basado en el enfoque actor-crítico, se usa principalmente para resolver problemas de acción continua. Es una combinación de Q-learning y un enfoque basado en políticas, donde la política se entrena para tomar decisiones determinísticas basadas en las observaciones del entorno y el crítico evalúa esas decisiones.

    - El DDPG se utiliza comúnmente para tareas que necesitan de movimientos suaves como el controlar brazos robóticos o bien prótesis es un espacio tridimensional. En estas aplicaciones la precisión es lo primordial y las decisiones no pueden estar restringidas a acciones discretas. Por ejemplo, la prótesis puede necesitar un ajuste en algún ángulo con precisión para agarrar un objeto sin perder el control.

3. **Trust Region Policy Optimization (TRPO)**

    - El TRPO es un algoritmo que se usa para entrenar políticas en entornos de aprendizaje por esfuerzo con una mejora segura en cada paso de la política. Este algoritmo busca realizar actualizaciones más seguras y estables de la política y así asegurar que los cambios de la misma, entre pasos estén dentro de una región de confianza limitada.

    - Esto se puede utilizar en juegos con mucha variabilidad en los estados o donde las recompensas dependen de largas secuencias de acciones, como en juegos de estrategia o simulación, TRPO puede ser valioso para estabilizar el aprendizaje. Juegos con muchos sub ojetivos y recompensas distribuidas en el tiempo son adecuados para TRPO porque la restricción de cambios e política ayuda a evitar comportamientos no deseados.

4. **Asynchronous Advantage Actor-Critic (A3C)**

    - Es un algoritmo basado en el enfoque actor-crítico que permite entrenar múltiples agentes (o hilos) de manera asíncrona en entornos de refuerzo. Este método mejora la estabilidad y eficiencia del entrenamiento al permitir que varios agentes interactúen con diferentes instancias del entorno simultáneamente.

    - Este puede ser muy útil en sistemas de control en tiempo real donde la acción debe ser decidida rápidamente, como la manipulación de robots, A3C es útil debido a su eficiencia computacional. Muchos agentes entran a la vez, lo que permite que el sistema se adapte a tiempo real y mejore su rendimiento de manera más rápida.