Introducción del algoritmo de aprendizaje por refuerzo DQN para juegos de Atari

Angel Larreategui Castro
Universidad Nacional de Ingeniería
Lima, Perú
alarreateguic@uni.pe

Fernando Zambrano Altamirano
Universidad Nacional de Ingeniería
Lima, Perú
fzambranoa@uni.pe

José Reyes Gutiérrez
Universidad Nacional de Ingeniería
Lima, Perú
jreyesg@uni.pe

***Resumen*—**

Índice de Términos— Deep Learning, DQL, Atari, Atari 2600, Deep Mind, Agent57

I. Introducción

II. Objetivo del estudio

Analizar el desempeño de la IA y ver que alcance el mayor rendimiento posible en los juegos (Constantemente superándose).

III. Problema

Ya que los videojuegos son un desafío de varias tareas y de toma rápida de decisiones, se busca que un algoritmo pueda adaptarse a esto con miras a poder adaptarse a otras multitareas con la misma eficiencia.

IV. Marco Teórico

IV-A. Deep Learning

Es un conjunto de algoritmos de aprendizaje automático que intenta modelar abstracciones de alto nivel en datos usando arquitecturas computacionales que admiten transformaciones no lineales múltiples e iterativas de datos expresados en forma matricial o tensorial. El aprendizaje profundo es parte de un conjunto más amplio de métodos de aprendizaje automático basados en asimilar representaciones de datos. Una observación (por ejemplo, una imagen) puede ser representada en algunas formas (por ejemplo, un vector de píxeles), pero algunas representaciones hacen más fácil aprender tareas de interés (por ejemplo, "¿es esta imagen una cara humana?") sobre la base de ejemplos, y la investigación en esta área intenta definir qué representaciones son mejores y cómo crear modelos para reconocer estas representaciones.

IV-B. Q-Learning

En el problema completo de aprendizaje por refuerzo, el estado cambia cada vez que ejecutamos una acción. Podemos representar el problema de la siguiente manera. El agente recibe el estado (**state**) en el que se encuentra el entorno (**environment**), el cual representaremos con la letra s (**state**). El agente ejecuta entonces la acción que elija, representada con la letra a (**action**). Al ejecutar esa acción, el entorno responde proporcionando una recompensa, representada con la letra r (**reward**), y el entorno se traslada a un nuevo estado, representado con s' (**next state**). Este ciclo se puede observar en la figura 1.

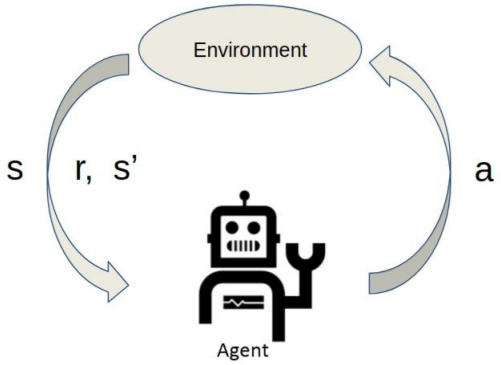

Figura 1. Representación del cambio de estado cada vez que se ejecuta una acción [1]

Por lo tanto, la acción que el agente escoja no debe sólo depender de la recompensa a que vaya a recibir a corto plazo. Debe elegir las acciones que a largo plazo le traerán la máxima recompensa (o retorno) posible en todo el episodio (**episode**). Este ciclo trae una secuencia de estados, acciones y recompensas, desde el primer paso del ciclo hasta el último: $s1, a1, r1; s2, a2, r2; \ldots; sT, aT, rT$. Aquí, T indica el fin del episodio.

IV-B1. Función de valor: Para cuantificar cuanta recompensa obtendrá el agente a largo plazo desde cada estado, introducimos la función de valor $V(s)$. Esta función produce una estimación de la recompensa que obtendrá el agente hasta el final del episodio, empezando desde el estado s. Si conseguimos estimar este valor correctamente, podremos decidir ejecutar la acción que nos lleve al estado con el valor más alto.

IV-B2. Q-Learning, resolviendo el problema: Para resolver el problema del aprendizaje por refuerzo, el agente debe aprender a escoger la mejor acción posible para ca-

da uno de los estados posibles. Para ello, el algoritmo Q-Learning intenta aprender cuanta recompensa obtendrá a largo plazo para cada pareja de estados y acciones (s,a). A esa función la llamamos la función de acción-valor (action-value function) y este algoritmo la representa como la función Q(s,a), la cual devuelve la recompensa que el agente recibirá al ejecutar la acción a desde el estado s, y asumiendo que seguirá la misma política dictada por la función Q hasta el final del episodio. Por lo tanto, si desde el estado s, tenemos dos acciones disponibles, a1 y a2, la función Q nos proporcionará los valores-Q (Q-values) de cada una de las acciones. Por ejemplo, si Q(s,a1)=1 y Q(s,a2)=4, el agente sabe que la acción a2 es mejor y le traerá mayor recompensa, por lo que será la acción que ejecutará.

IV-B3. Ecuación de Bellman: La explicación para esta ecuación es la siguiente. El valor-Q del estado s y la acción a (Q(s, a))debe ser igual a la recompensa r obtenida al ejecutar esa acción, más el valor-Q de ejecutar la mejor acción posible a' desde el próximo estado s', multiplicado por un factor de descuento (discount factor), que es un valor con rango (0, 1]. Este valor se usa para decidir cuánto peso le queremos dar a las recompensas a corto y a largo plazo, y es un hiperparámetro que debemos decidir nosotros.

IV-C. El algoritmo Deep Q-Network o DQN

Vimos que el algoritmo Q-Learning funciona muy bien cuando el entorno es simple y la función $Q(s,a)$ se puede representar como una tabla o matriz de valores. Pero cuando hay miles de millones de estados diferentes y cientos de acciones distintas, la tabla se vuelve enorme, y no es viable su utilización. Por ello, Mnih et al. [1] inventaron el algoritmo Deep Q-Network o DQN. Este algoritmo combina el algoritmo Q-learning con redes neuronales profundas (Deep Neural Networks). Como es sabido en el campo de la IA, las redes neuronales son una fantástica manera de aproximar funciones no lineales. Por lo tanto, este algoritmo usa una red neuronal para aproximar la función Q, evitando así utilizar una tabla para representar la misma. En realidad, utiliza dos redes neuronales para estabilizar el proceso de aprendizaje. La primera, la red neuronal principal (main Neural Network), representada por los parámetros θ, se utiliza para estimar los valores-Q del estado s y acción a actuales: $Q(s,a;\theta)$. La segunda, la red neuronal objetivo (target Neural Network), parametrizada por θ', tendrá la misma arquitectura que la red principal pero se usará para aproximar los valores-Q del siguiente estado s' y la siguiente acción a'. El aprendizaje ocurre en la red principal y no en la objetivo. La red objetivo se congela (sus parámetros no se cambian) durante varias iteraciones (normalmente alrededor de 10000), y después los parámetros de la red principal se copian a la red objetivo, transmitiendo así el aprendizaje de una a otra, haciendo que las estimaciones calculadas por la red objetivo sean más precisas.

V. Organización del informe (secciones)

En el presente trabajo nosotros vamos a realizar las siguientes secciones:

VI. Estado del Arte

VII. Metodología

VIII. Diseño del experimento

IX. Experimentación y resultados

X. Discusión de resultados

XI. Conclusiones

XII. Trabajos Futuros

Referencias

[1] Introduccion al aprendizaje por refuerzo q-learning Disponible en https://markelsanz14.medium.com/introducción-al-aprendizaje-por-refuerzo-parte-2-q-learning-883cd42fb48e.

[2] Agent57: Outperforming the human Atari benchmark Disponible en https://deepmind.com/blog/article/Agent57-Outperforming-the-human-Atari-benchmark.