UNIVERSIDAD DE CUENCA
desde 1867

Facultad de Ingeniería
Universidad de Cuenca
Grado en Ingeniería de Sistemas
Curso 2016-2017

Machine Learning

Aprendizaje Reforzado

Q-Learning y Deep Q-Learning

Freddy Abad, Edwin Cabrera, Daniel Campoverde
<freddy.abadl,edwin.cabrera,daniel.campoverde>@ucuenca.ec

Contenido

UNIVERSIDAD DE CUENCA
desde 1867

- Objetivos
- Descripción del problema: “La Final”
- Metodología y Métodos
  - *Q Learning*
  - *Deep Q Learning*
- Resultados
- Conclusiones
- Bibliografía

Objetivos

UNIVERSIDAD DE CUENCA
desde 1867

- Comprender y aplicar *Q-learning*
- Comprender y aplicar *Deep Q-learning*
- Ofrecer soluciones al problema *"La Final"* usando ambos algoritmos

Descripción del problema: “La Final”

UNIVERSIDAD DE CUENCA
desde 1867

- Un arquero tapa tiros de penal
- Espacio discreto de *recuadros*
- Usando *Q-Learning* y *Deep Q-Learning*

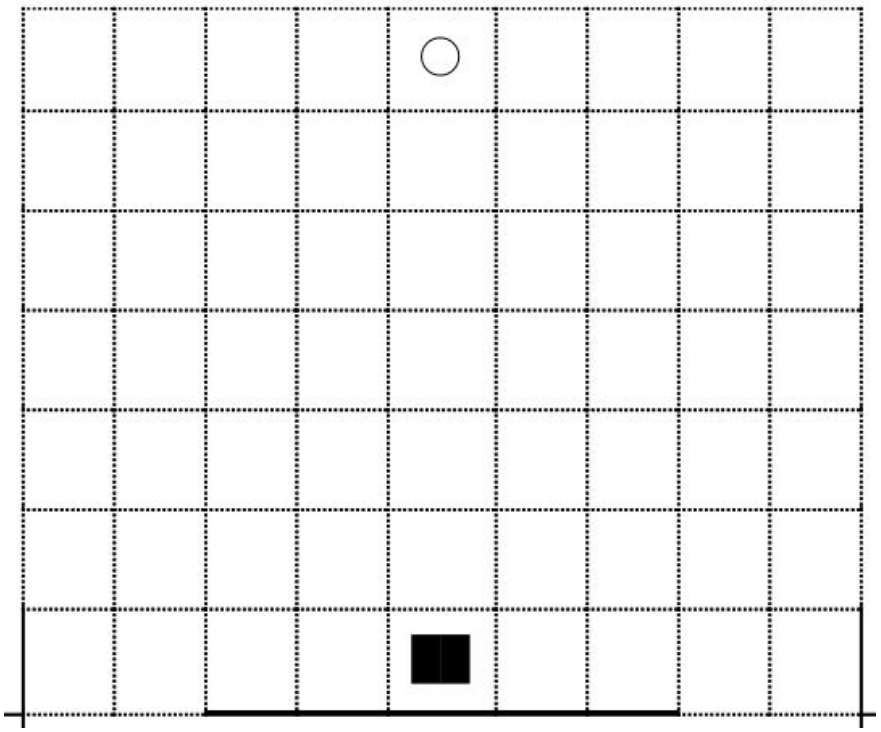


Figura 1: Problema “La Final”

Descripción del problema: “La Final”

UNIVERSIDAD DE CUENCA
desde 1867

Recompensas

- Tapar: **+2**
- Gol: **-1**
- Balón fuera
  - Arquero dentro: **+1**
  - Arquero fuera: **-1**

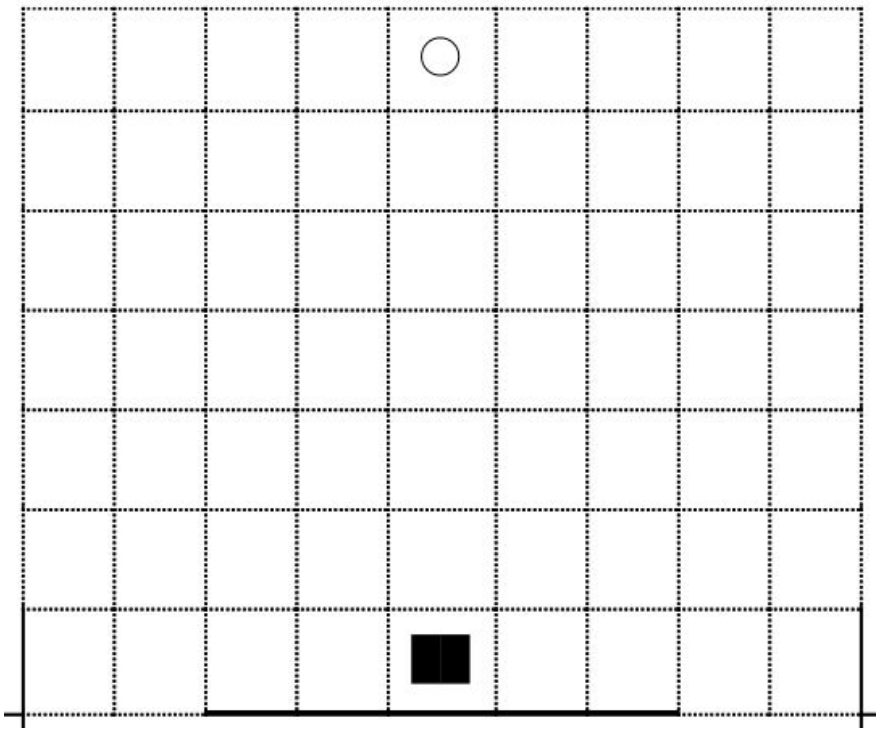


Figura 1: Problema “La Final”

Metodología y Métodos

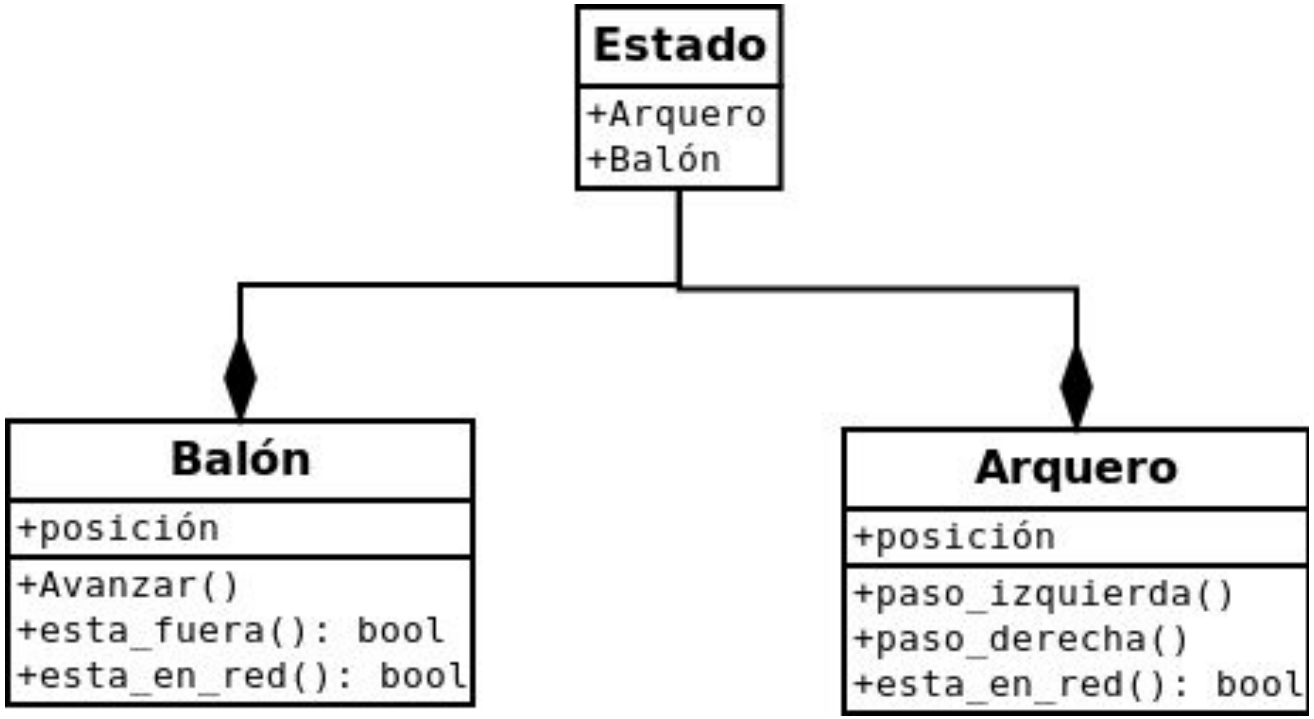


Figura 2: Diagrama de Clases

Metodología y Métodos

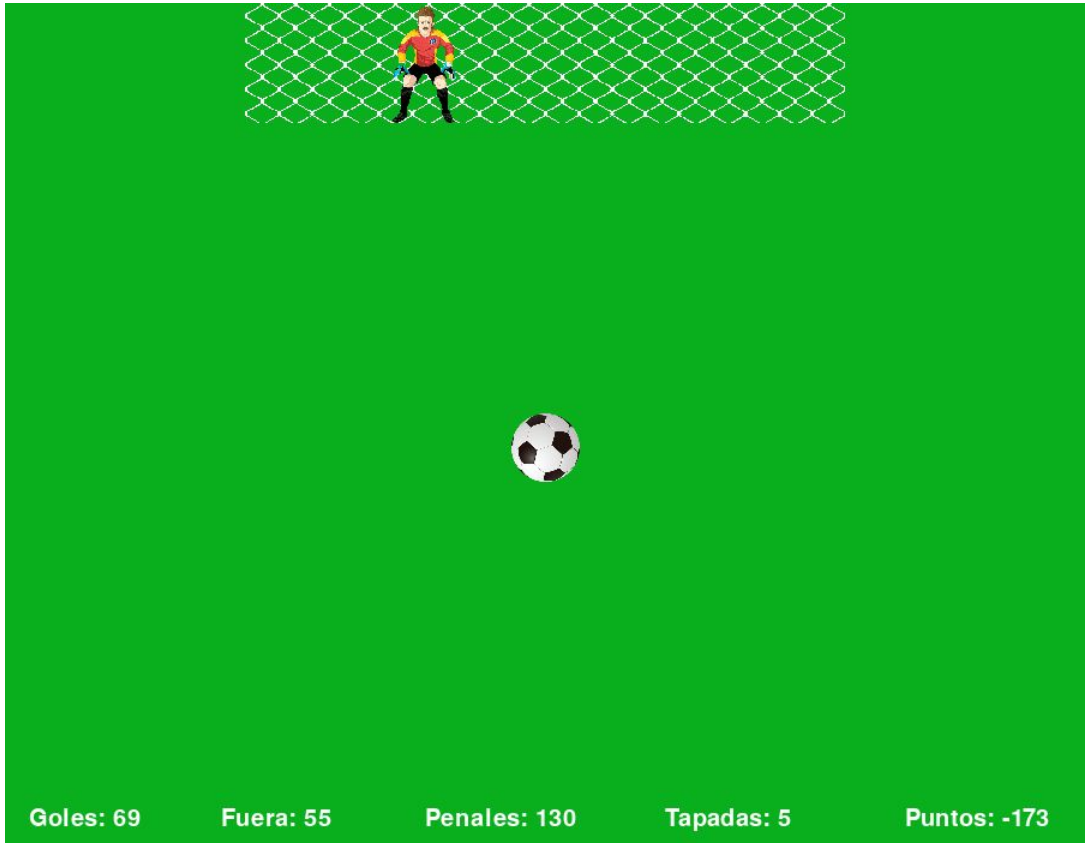


Figura 3: Interfaz del juego

Q-Learning

UNIVERSIDAD DE CUENCA
desde 1867

- Encuentra reglas de acción óptimas
- Aprende una función ***Q(s,a)*** de estados ***s*** y acciones ***a***
- Crea soluciones para tareas específicas

Q-Learning

UNIVERSIDAD DE CUENCA
desde 1867

Estado ***s***

Combinación de las posiciones en X del arquero y balón

Q-Learning

UNIVERSIDAD DE CUENCA
desde 1867

Acción ***a***

Movimiento del arquero en el eje X

Q-Learning

UNIVERSIDAD DE CUENCA
desde 1867

Q(e, a) = Q(e, a) + lr *x* (reward + gamma *x* max{a}(Q(e', a))

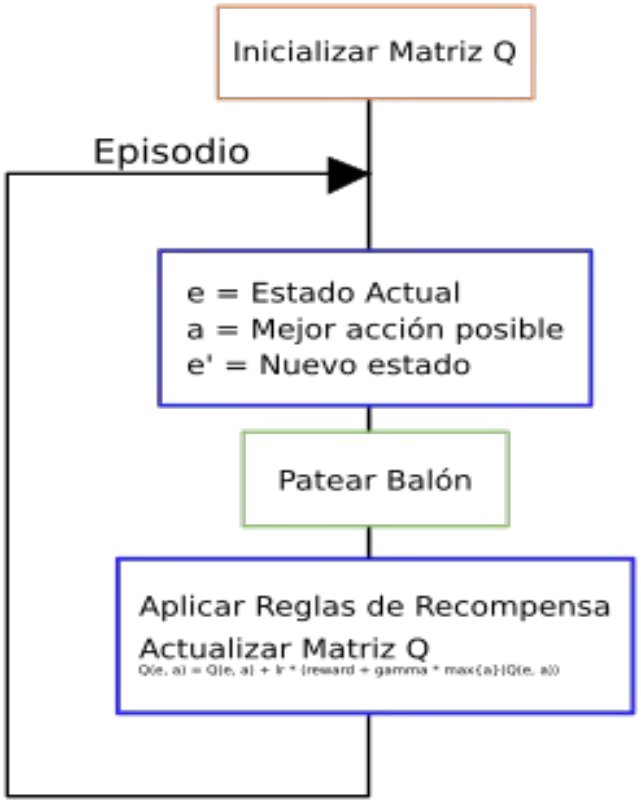


Figura 4: Algoritmo Q Learning

Optimización de Parámetros

UNIVERSIDAD DE CUENCA
desde 1867

Parámetros de *Q Learning*

- Learning Rate (lr): Sensibilidad a la retroalimentación
- Gamma: Preferencia de beneficios futuros vs inmediatos

Optimización de Parámetros

UNIVERSIDAD DE CUENCA
desde 1867

El puntaje más alto para 2000 episodios es de *26306* Ocurre

para un Q *Learning rate = 1* y *Gamma = 0*

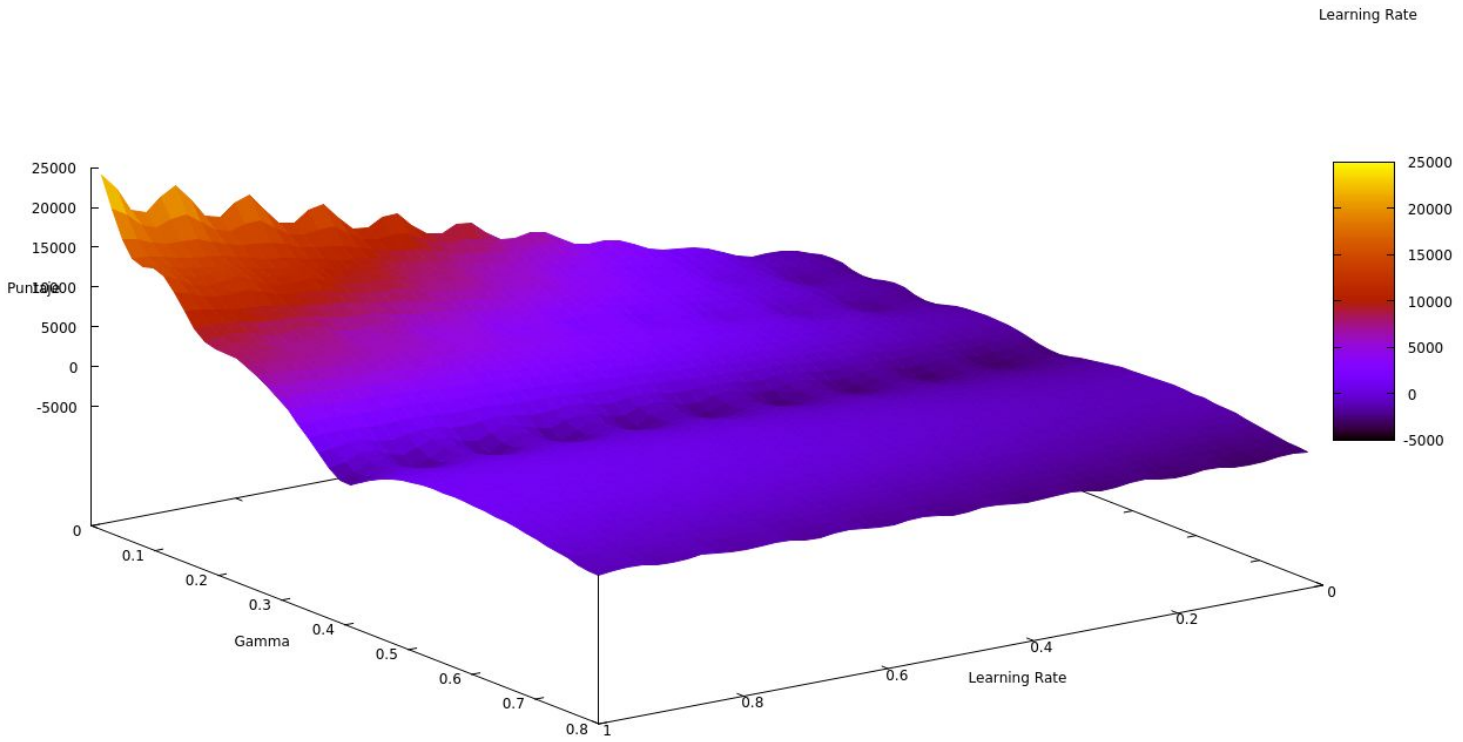


Figura 5: Superficie de optimización

Optimización de Parámetros

UNIVERSIDAD DE CUENCA
desde 1867

Número de tiros penal *vs puntaje* acumulado

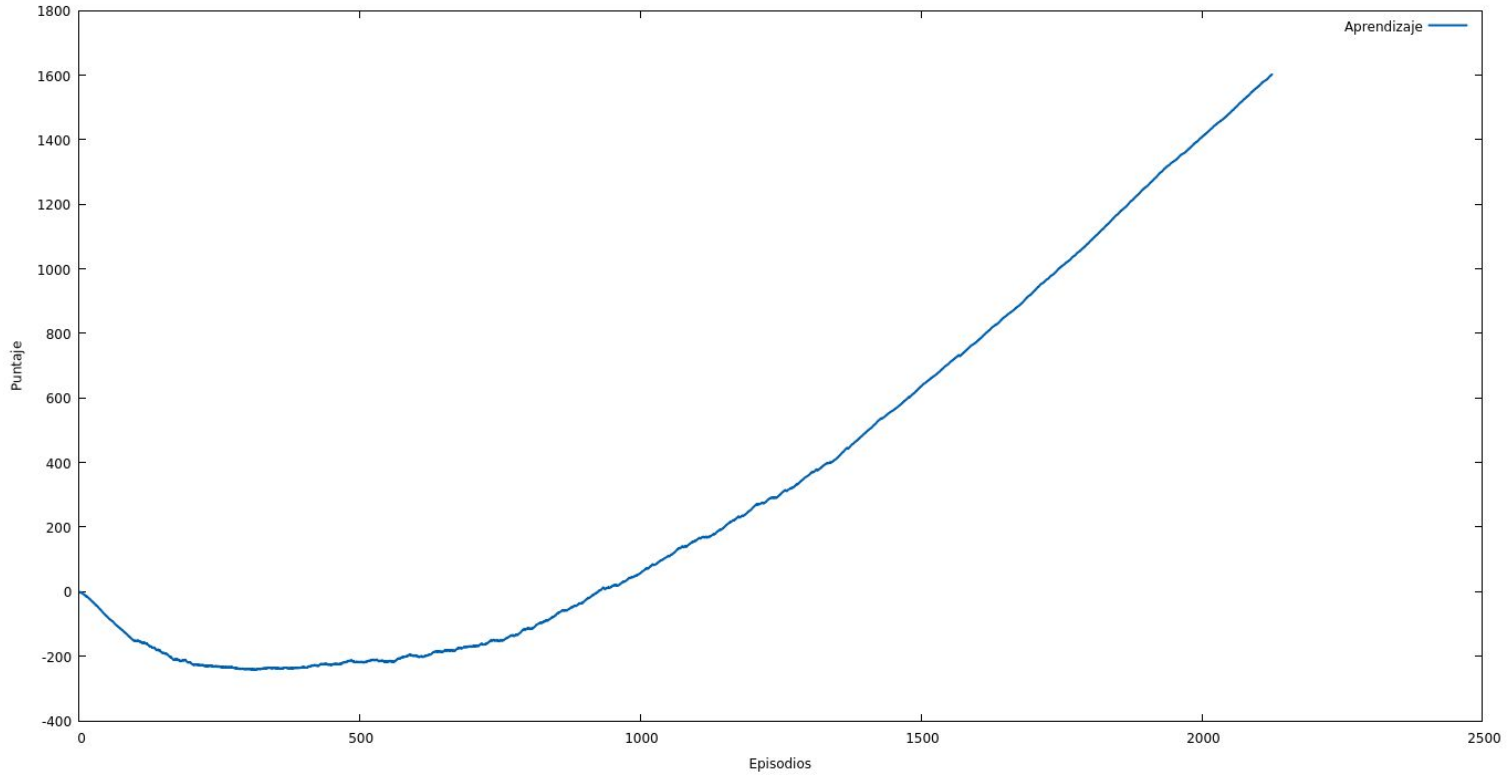


Figura 6: Curva de aprendizaje

Deep Q-learning

UNIVERSIDAD DE CUENCA
desde 1867

- Crea soluciones generalizadas que son aplicables en varias tareas diversas
- Combina *Q-Learning* con aprendizaje profundo para representar la tabla Q
- Observa los pixeles del juego para deducir el estado

Deep Q-learning

UNIVERSIDAD DE CUENCA
desde 1867

Emplea una red neuronal convolucional (una variante del perceptrón multicapa original)

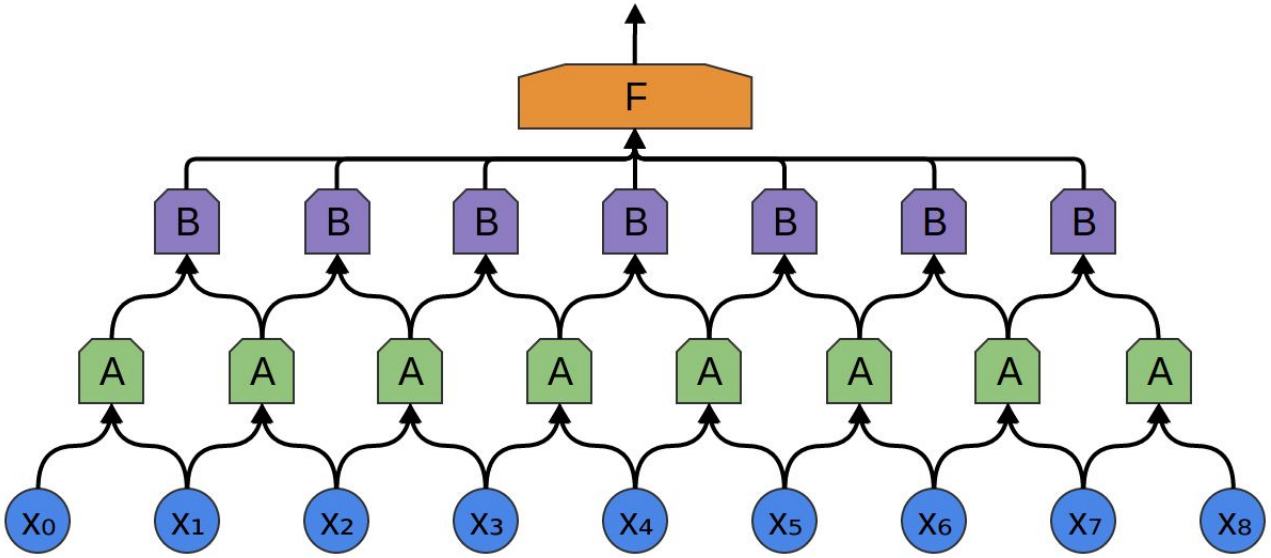


Figura 7: Estructura red neuronal convolucional

Deep Q-learning

UNIVERSIDAD DE CUENCA
desde 1867

Deep Q-Learning vs. Q-Learning

- *Ventaja*: Puede ser reutilizada para diferentes problemas
- *Desventaja*: Se requieren almacenar los estados pasados

Deep Q-learning

UNIVERSIDAD DE CUENCA
desde 1867

El número de frames capturados por segundos representan un estado del entorno en formato de imagen, es por ello que se la recopilación de dichas imágenes representan una mayor carga computacional.

Sí fps=4, por lo tanto se registran 240 estados por segundo.

Deep Q-learning

UNIVERSIDAD DE CUENCA
desde 1867

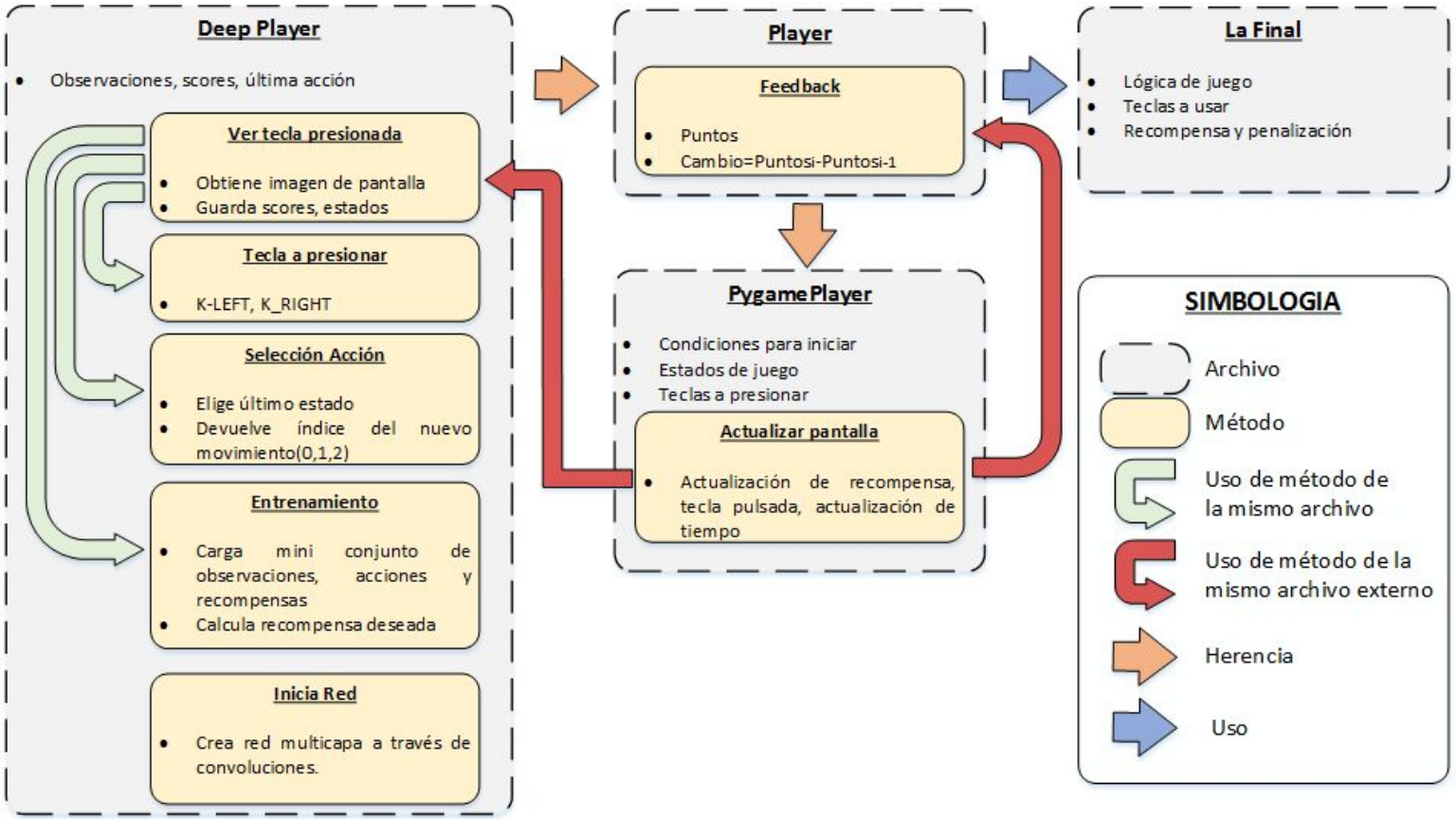


Figura 8: Implementación Deep-Q Learning

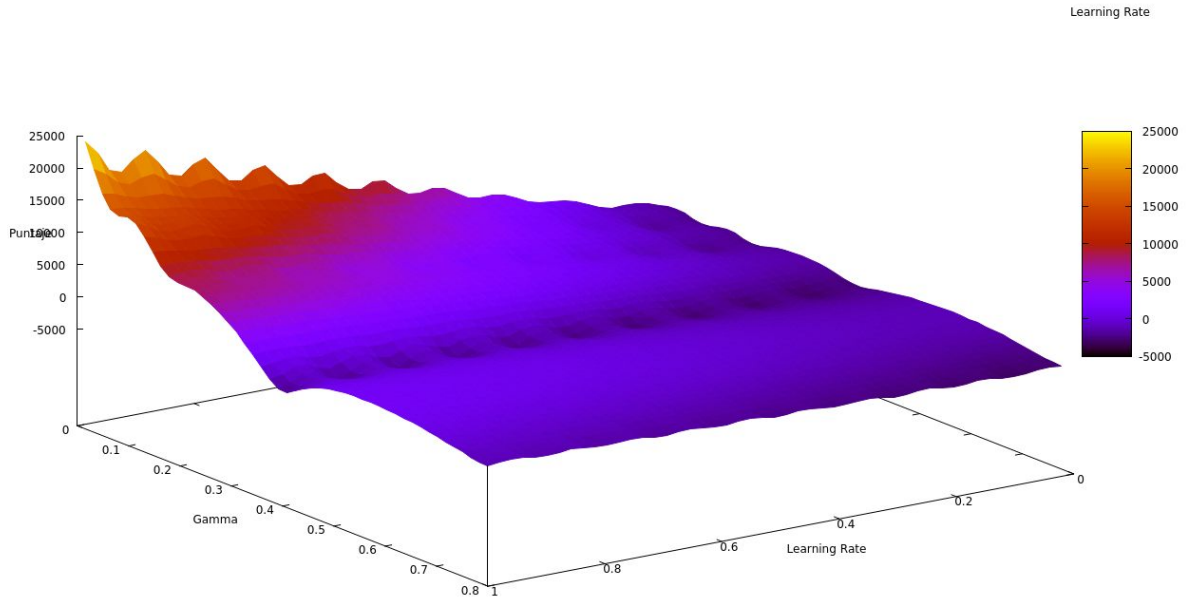
Resultados

UNIVERSIDAD DE CUENCA
desde 1867

Parámetros óptimos

Tomar en cuenta únicamente el beneficio inmediato (Gamma = 0)

Aprender siempre de la recompensa (learning rate = 1)

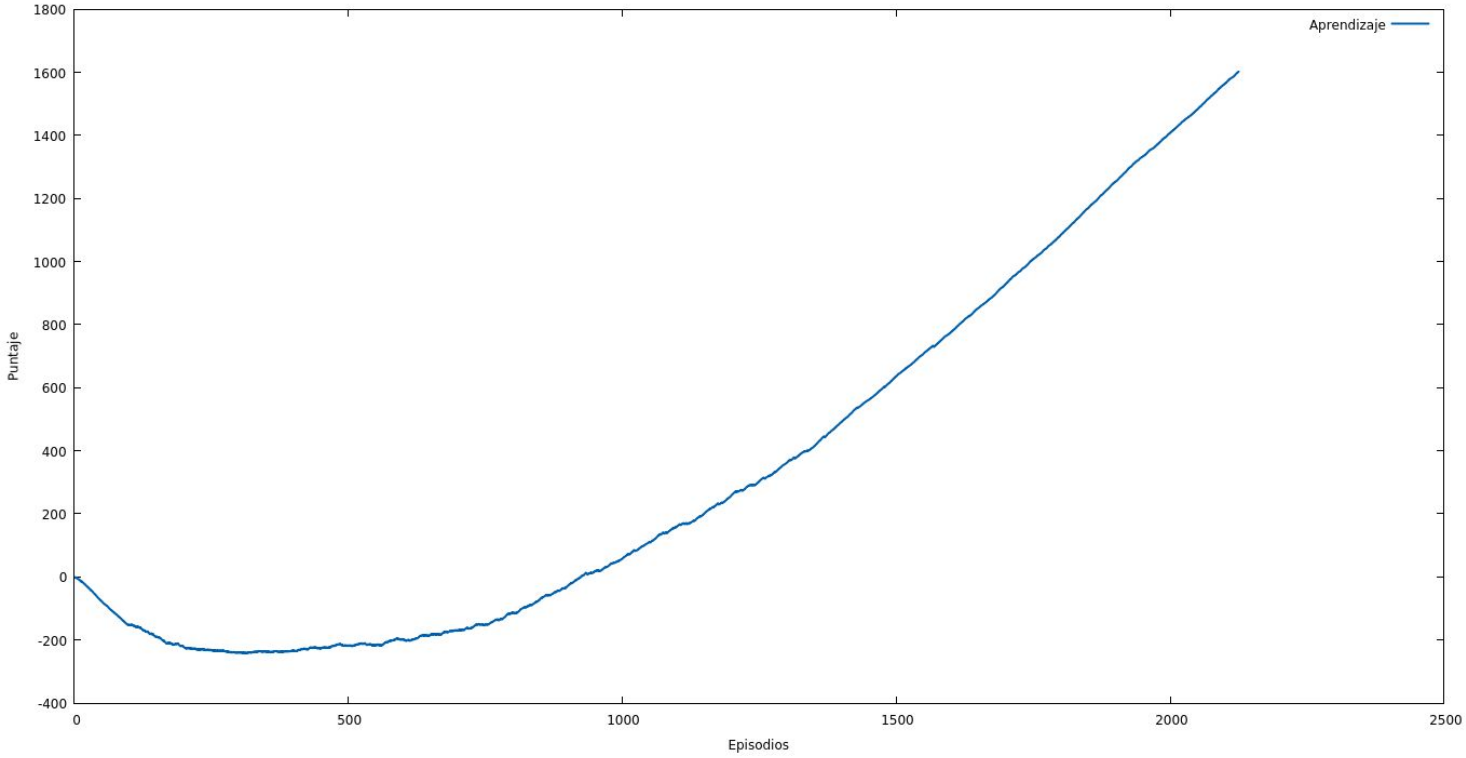


Resultados

UNIVERSIDAD DE CUENCA
desde 1867

Decisiones al azar: Descenso del puntaje durante *~300* tiros penal

Comportamiento óptimo: El arquero hace siempre lo correcto desde el tiro de penal ~1500



Resultados

UNIVERSIDAD DE CUENCA
desde 1867

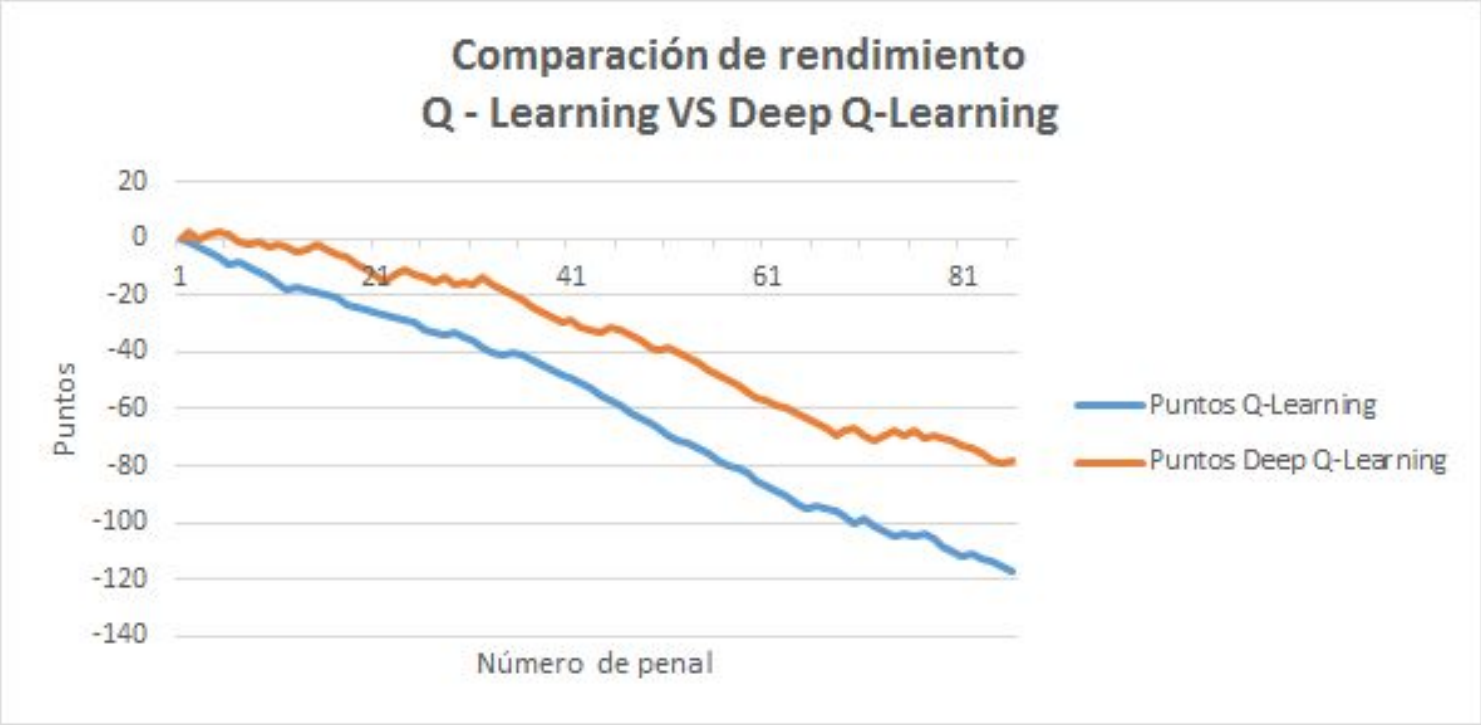
En la solución con *Deep Q-Learning* no se logró observar el momento donde el arquero deja de fallar, debido a la demanda de recursos involucrados

Prueba	Penales	Goles	Fuera	Tapadas	Puntos
1	60	32	27	1	-55
2	60	31	27	2	-49
3	60	29	29	2	-53
4	60	28	27	5	-43
5	60	31	24	5	-48
6	60	27	30	3	-34
7	60	32	25	3	-57
8	60	30	27	3	-53
9	60	32	22	6	-46
10	60	30	25	5	-41
Media	**60**	**30**	**26**	**4**	**-48**

Tabla 1: Rendimiento DQL (10 pruebas)

Resultados

Comparando las curvas de aprendizaje de ambos algoritmos se observa que *Deep Q-Learning* presenta una tendencia similar a la de *Q-Learning*



Conclusiones

UNIVERSIDAD DE CUENCA
desde 1867

- Se logró un comportamiento óptimo para el sujeto (arquero)
- Se encontraron los parámetros óptimos de los algoritmos empleados, para el problema resuelto
- Se obtuvo una solución complementaria que es potencialmente aplicable a otro problema similar usando *Deep Q Learning*

Bibliografía

- [1] Keon. 2017. Deep Q-Learning with Keras and Gym
- [2] Matiisen. 2015. Demystifying Deep Reinforcement Learning
- [3] Mnih. et. al. 2013. Playing atari with deep reinforcement Learning
- [4] Salter D. 2016. Deep Q-Learning Pong with Python & TensorFlow.