UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Estrategias de aprendizaje automático aplicadas a videojuegos

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Adrián Valero Gimeno

Tutor: Vicent Botti Navarro
Javier Palanca

Curso 2018-2019

Resum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas semper facilisis rutrum. Nam ullamcorper orci id nisl euismod facilisis. Pellentesque condimentum orci placerat, pulvinar est ut, scelerisque lacus. Pellentesque magna augue, dignissim a tempor in, tincidunt eget mauris. Nunc at sodales nulla. Maecenas congue id sem sagittis mattis. Sed a vehicula justo. Sed rhoncus rutrum ipsum a dictum. Etiam luctus sodales aliquam. Praesent nisl justo, ullamcorper et luctus ut, facilisis vel nibh. Aliquam imperdiet finibus euismod. Donec id posuere libero, eu imperdiet eros. Sed commodo egestas dolor. Ut bibendum turpis mi, vitae iaculis ligula mattis sed. Aliquam dapibus augue et felis pulvinar condimentum vel id mauris.

Paraules clau: ??????????????????

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas semper facilisis rutrum. Nam ullamcorper orci id nisl euismod facilisis. Pellentesque condimentum orci placerat, pulvinar est ut, scelerisque lacus. Pellentesque magna augue, dignissim a tempor in, tincidunt eget mauris. Nunc at sodales nulla. Maecenas congue id sem sagittis mattis. Sed a vehicula justo. Sed rhoncus rutrum ipsum a dictum. Etiam luctus sodales aliquam. Praesent nisl justo, ullamcorper et luctus ut, facilisis vel nibh. Aliquam imperdiet finibus euismod. Donec id posuere libero, eu imperdiet eros. Sed commodo egestas dolor. Ut bibendum turpis mi, vitae iaculis ligula mattis sed. Aliquam dapibus augue et felis pulvinar condimentum vel id mauris.

Palabras clave: Inteligencia artificial, aprendizaje, automatico, videojuegos, OpenAI, hiperparámetros

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas semper facilisis rutrum. Nam ullamcorper orci id nisl euismod facilisis. Pellentesque condimentum orci placerat, pulvinar est ut, scelerisque lacus. Pellentesque magna augue, dignissim a tempor in, tincidunt eget mauris. Nunc at sodales nulla. Maecenas congue id sem sagittis mattis. Sed a vehicula justo. Sed rhoncus rutrum ipsum a dictum. Etiam luctus sodales aliquam. Praesent nisl justo, ullamcorper et luctus ut, facilisis vel nibh. Aliquam imperdiet finibus euismod. Donec id posuere libero, eu imperdiet eros. Sed commodo egestas dolor. Ut bibendum turpis mi, vitae iaculis ligula mattis sed. Aliquam dapibus augue et felis pulvinar condimentum vel id mauris.

Key words: ?????, ????? ?????, ?????????????

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

El aprendizaje automático o "Machine Learning" ha sido durante los últimos años el foco de muchísima investigación, debido a su gran potencial en la aplicación a problemas del mundo moderno. En los últimos años, proyectos como AlphaGo o la supercomputadora de google Deep Mind han conseguido hacer grandes avances en juegos de gran dificultad, siendo los agentes desarrollados capaces de competir contra las mentes más experimentadas del tradicional juego de mesa Go.

Hoy en día se están consiguiendo hacer grandes avances en el campo, debido a la investigación en nuevas técnicas como la combinación de redes neuronales tradicionales con otros métodos como el Deep Learning. Por el momento, Google está liderando este nuevo movimiento desde la compra de la empresa DeepMind. Esta empresa fue capaz de desarrollar AlphaGo, una inteligencia artificial capaz de ganar a las mentes más experimentadas del juego tradicional chino Go. Éste juego, considerado uno de los más difíciles del mundo, se calcula que tiene sobre 10^{172} configuraciones posibles de las piezas sobre el tablero - un número superior al número estimado de átomos existentes en el universo -, haciéndolo extraordinariamente más complejo que juegos como el ajedrez. Recientemente, DeepMind también desveló AlphaStar, una nueva inteligencia artificial que por el momento es capaz de competir al mismo nivel que algunos de los mejores jugadores de StarCraft del mundo.

1.1 Motivación

Se ha elegido un trabajo de estas características debido a una motivación personal hacia los campos de la inteligencia artificial, así como una manera de extender el conocimiento en ciertos ámbitos del mundo de la computación los cuales no se encuentran necesariamente presentes en un plan de estudios tradicional en el campo de la Ingeniería Informática. De esta forma se busca además un desarrollo personal y profesional en el campo de la computación, que permita el acceso a un futuro laboral en éste campo.

Se partía además de una necesidad personal de realizar un trabajo propio y, de cierta manera, novedoso con respecto a lo que podría conllevar la realización de un proyecto de fin de carrera típico. En definitiva, realizar un trabajo unos objetivos y una metodología a seguir bien definidos y, por encima de todo, que contribuyera al desarrollo de las competencias necesarias para la formación en el campo de la investigación y el desarrollo de sistemas de estas características.

El campo de la inteligencia artificial está creciendo exponencialmente, y no dejará de hacerlo en los próximos años. De esta manera, se cree que el aprendizaje por refuerzo puede suponer una nueva revolución en la disciplina del aprendizaje automático, cosa

que ya se puede observar en los avances realizados por empresas como DeepMind. De la misma manera, aplicar estos nuevos avances a los videojuegos supone un desafío muy alentador, que además se podría extrapolar más adelante a otros ámbitos como la robótica.

1.2 Objetivos

El propósito de este TFG consistirá en el estudio de las diferentes técnicas existentes de aprendizaje automático aplicadas a sencillos videojuegos en 2D, los cuales tendrán un número limitado de acciones a realizar. Se pretende además, realizar implementaciones propias de las diferentes técnicas descritas durante el resto del documento.

Otro de los objetivos consistirá en el estudio y realización de pruebas sobre distintos entornos predefinidos de juegos arcade, haciendo uso de librerías que permiten el acceso a dicho tipo de juegos como la herramienta OpenAI desarrollada por Google, o las distintas librerías de Python relacionadas con la creación de entornos de redes neuronales y, más concretamente, la manipulación de hiperparámetros para encontrar los mejores resultados para cada uno de los entornos estudiados.

Por último, resultaría interesante aplicar estas nuevas técnicas estudiadas en la resolución de algún videojuego de mayor complejidad. Se plantea, por lo tanto, un estudio de los algoritmos más destacables estudiados en algún juego complejo como puede ser Montezuma's Revenge, un juego de la atari 2600 estrenado en el año 1983.

1.3 Metodología

Para la realización de este trabajo se pretende llevar a cabo una investigación en el campo de la inteligencia artificial, concretamente en el desarrollo de algoritmos de aprendizaje aplicados a entornos estocásticos. Para ello, se investigarán técnicas como el aprendizaje por refuerzo o Q-Learning, algoritmos evolutivos, o el uso de redes neuronales. Adicionalmente, se ha seguido un curso externo enfocado a la teoría de inteligencia artificial aplicada a distintos videojuegos, en los cuales se introducen los conceptos principales en los cuales se basa todo el campo del aprendizaje por refuerzo, desde la ecuación de Bellman hasta el diseño de redes neuronales convolucionales como método para acceder a la información de nuestros entornos.

Para complementar nuestra base de conocimientos, nos basaremos en artículos de investigación publicados por entidades como Google DeepMind, o el Massachussets Institute of Technology, en busca de ideas y nuevas técnicas para su posterior aplicación en nuestro trabajo.

1.4 Estructura de la memoria

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 2

Estado del arte. La situación del aprendizaje automático en la actualidad

2.1 Introducción al Q-Learning

El aprendizaje por refuerzo es un principio altamente utilizado en la actualidad en la investigación del campo de la Inteligencia Artificial. Este enfoque se inspira en el campo de la psicología de comportamiento y pone el énfasis en cómo un agente independiente será capaz de tomar las acciones pertinentes basándose en un sistema de recompensas que le son suministradas en función de las acciones que toma en cada momento. En este sentido, se intenta emular la capacidad de los seres vivos de hacer predicciones basándose en la información que se percibe del entorno. Estas predicciones [8] proporcionan a los animales tiempo para preparar reacciones de comportamiento, que en un futuro pueden mejorar las opciones de las que éste disponga en el futuro. De esta manera, se habla de recompensaçomo un concepto que describe un valor positivo que una criatura le atribuye a un objeto, patrón de comportamiento o estado físico interno. Éstas recompensas se manifiestan en una gran variedad de animales a través de un neurotransmisor llamado dopamina. En estos sistemas, la dopamina se produce como respuesta a una recompensa inesperada que, si se produce de manera repetida a lo largo del tiempo, puede condicionar el comportamiento del organismo. Frecuentemente, se relaciona la dopamina con funciones vitales tan importantes como la motivación, la regulación de la memoria o en la toma de decisiones. Sirve, por lo tanto, como un método de refuerzo positivo, en situaciones en las que estas respuestas constituyan un patrón de comportamiento recurrente que se mantenga a lo largo del tiempo.

En los últimos años, se ha podido observar un crecimiento notable en el uso de sistemas de inteligencia artificial en empresas de internet hasta el punto en el que se han convertido en un requisito casi indispensable para ser capaz de ofrecer un servicio específico; por ejemplo, los sistemas de recomendación en servicios de plataformas de Streaming o la publicidad personalizada que muchas páginas de compras por internet ofrecen basándose en las compras o búsquedas realizadas anteriormente. Estos sistemas se basan en el entrenamiento de redes neuronales que permiten crear una relación entre las selecciones anteriores y productos afines. En las próximas secciones haremos un repaso de las técnicas más conocidas de aprendizaje por refuerzo, así como repasar ciertos conceptos esenciales para una comprensión correcta de lo presentado.

2.1.1. Los elementos del Aprendizaje por refuerzo

Como se ha mencionado anteriormente, el aprendizaje por refuerzo se basa en ideas de la neurociencia y la psicología, específicamente del campo de la psicología del comportamiento.

Para plantear el problema sobre un entorno sencillo, se suele modelar el problema del aprendizaje por refuerzo como un conjunto de elementos:

El conjunto de estados. Se trata de una colección de todos los estados posibles en los que un agente se puede encontrar, teniendo en cuenta toda la información del entorno.

El conjunto de acciones. Como su nombre indica, recoge todas las acciones posibles que un agente puede realizar en un entorno determinado.

Reglas de transición. Normalmente se componen de una serie de valores escalares que determinan la **recompensa** resultante de ejecutar una acción sobre un estado determinado.

Función de recompensa. Definirá el objetivo del problema. Consiste en una colección de relaciones estado-acción a las que se les asignará un valor numérico individual, dependiendo de si la acción tomada en un estado dado es favorable o no para la consecución del objetivo. El objetivo del agente será siempre intentar maximizar la recompensa total obtenida en un episodio[1].

En resumidas cuentas se podría decir que el aprendizaje por refuerzo consiste en introducir a un agente independiente una serie de directivas que le permitan atribuir un valor sobre una o un conjunto de acciones que ejecuta en un entorno. De esta manera, el agente podrá ser capaz de determinar una serie de acciones que le ofrezcan un resultado favorable. El agente aprenderá a medida que ejecute acciones (las cuales en un primer momento serán escogidas aleatoriamente) y sopese las recompensas obtenidas, que el entorno le devolverá en función de las acciones que vaya tomando.

El aprendizaje que el agente experimenta puede variar dependiendo de las directrices (hiperparámetros) que se le introduzcan o de la aleatoriedad que el propio entorno puede presentar. Asimismo, existen una gran variedad de algoritmos de aprendizaje por refuerzo que se pueden aplicar. En la sección **POR COMPLETAR** mencionaremos algunos junto con su funcionamiento.

Además de lo explicado anteriormente, vamos a introducir una serie de conceptos frecuentemente usados cuando se habla de sistemas de aprendizaje por refuerzo. Estos elementos son: *política*, *modelo de entorno* y lo que llamaremos *función de valía*.

Una *política* definirá el comportamiento de el agente en cuestión en un momento determinado. Consistirá pues, en un mapeado de los estados que se han percibido hasta el momento en relación a las acciones tomadas en el pasado. En psicología, se correspondería a lo que es llamado una serie de reglas de estímulo respuesta. Esta política podrá venir dada de diferentes maneras, desde una simple tabla de resultados a una serie de computaciones llevadas a cabo en una red neuronal.

Mientras que la función de recompensa nos indicará la recompensa inmediata que un agente recibirá al realizar una determinada acción, la *función de valía* o Q-Value nos especificará el valor esperado de recompensa que se conseguirá al seguir una serie de

[1] Al hablar de episodio, nos referimos a una determinada secuencia de estados dentro de un entorno hasta dar con un estado final.

acciones que, a priori, podrían no parecer las óptimas, pero pueden ser ventajosas a medio o largo plazo. Éste valor puede resultar muy complicado de calcular, y requiere una capacidad de cómputo enorme que aumenta exponencialmente en entornos con un gran número de dimensiones. Esto se traduce en una dificultad real - y que todavía no se ha sido capaz de solventar - de los agentes de planear acciones a medio o largo plazo.

El *modelo del entorno* será el que contenga toda la información referente al problema que se pretende solucionar, como el conjunto de estados, acciones o las funciones de recompensa.

La ecuación de Bellman

Richard Bellman fue un investigador estadounidense que formuló lo que también se conoce como la **ecuación de la programación dinámica**, que describe una condición necesaria para la optimalidad en la resolución de problemas Markovianos. Esta ecuación constituye un elemento esencial para la comprensión de los algoritmos de aprendizaje por refuerzo, y aplica muchos de los conceptos introducidos en secciones anteriores.

2.1.2. Cadenas de Markov

Una cadena de Markov se podría entender como un grafo dirigido en el cual cada nodo (estado) cuenta con una serie de aristas (acciones posibles) con un valor asociado correspondiente a la probabilidad de que esta acción se ejecute. Éste tipo de sistemas han de ser **deteriministas**, es decir, se compondrán de una sucesión de estados conocidos a priori. La probabilidad de que cierta acción se ejecute dependerá solamente del evento inmediatamente anterior.

En nuestro trabajo, utilizaremos un enfoque que se basa en la consideración del problema a resolver como un proceso de decisión de Markov con recompensa (MRP). Esto será una tupla (S, P, R, γ) donde S_n denota el estado posible **S** en el que un agente se puede encontrar, P_n una función de transición **P** y R_n es la recompensa que un agente conseguirá al realizar una determinada acción. Por último, γ denotará el factor de descuento donde $\gamma \in [0, 1]$. Éste parámetro indica al agente cuanto peso debe darle a las recompensas que recibe para aplicar estos conocimientos en momentos posteriores del aprendizaje. Se ha observado en diferentes investigaciones que introducir un factor de descuento de 1 no garantizará necesariamente la convergencia, por lo que resulta recomendable dotar al agente de cierta autonomía para la exploración, que podría resultar en el descubrimiento de acciones más ventajosas a la hora de resolver un problema determinado. De esta manera, la noción en la que el agente percibirá la señal de recompensa $R(s, a)$ vendrá dada por:

$$R(s,a) = R_{t+1} + \gamma^2 R_{t+2} + ... + \gamma^{n-t} R_n = \sum_{k=0}^{n} \gamma^k R_{t+k+1}$$

El objetivo de nuestro agente será en todo momento conseguir maximizar la recompensa total obtenida en un episodio, basándose en el principio de la ecuación anterior para determinar los posibles caminos a seguir. De esta manera, la función de recompensa final podrá ser expresada como:

$$V(s) = max_a[R(s,a) + \gamma \sum s' P(s,a,s') V(s')]$$

2.1.3. Un pequeño ejemplo ilustrativo

Para terminar la sección, en la figura X se muestra un sencillo ejemplo de entorno de aprendizaje por refuerzo. En este ejemplo, el robot ganará la partida en el momento que se encuentre en la casilla superior derecha, y morirá en el caso de pisar la casilla en llamas. Además, para cada movimiento que el agente realice, hay una probabilidad de 0.1 de que el agente se mueva a uno de los lados

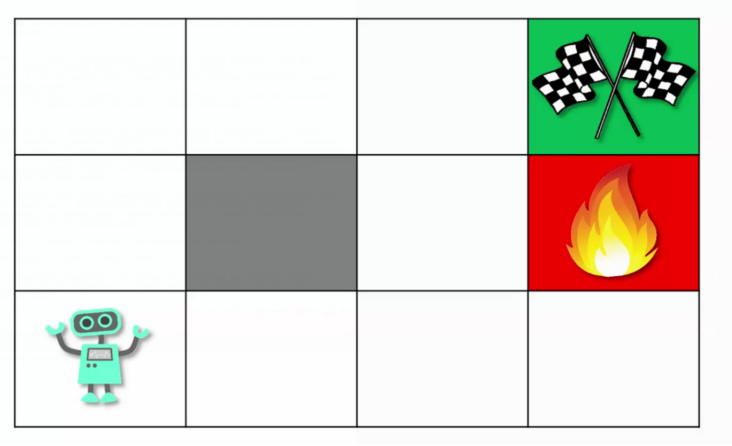

Figura 2.1: Ejemplo de entorno de aprendizaje por refuerzo

2.1.4. Redes neuronales

El concepto de red neuronal nace de la idea de conseguir un comportamiento autónomo similar al de un ser humano. De esta manera, las redes neuronales intentan replicar el razonamiento humano creando una relación entre lo que el agente observa y las acciones que ha de realizar. Dependiendo del problema a abarcar, estas redes se pueden ver reformuladas para ajustarse a una serie de requerimientos especiales. Por ejemplo, las redes neuronales convolucionales hacen uso de una serie de capas previas que extraen características de las imágenes para ser posteriormente introducidas en una estructura de red neuronal clásica y realizar así la tarea de clasificación.

Una red neuronal se compone esencialmente de tres componentes interconectados: Una capa de entrada, una o varias capas ocultas, y una capa de salida.

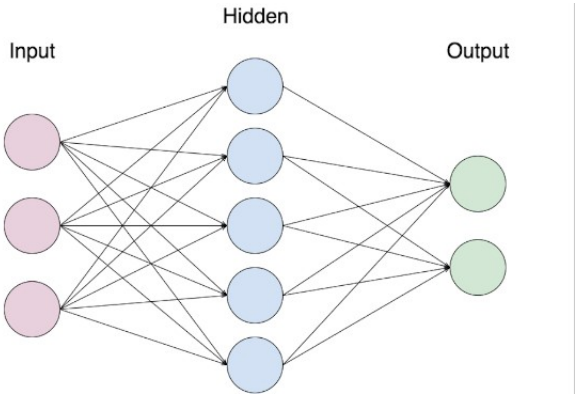

Figura 2.2: Ejemplo de una red neuronal con una capa oculta de 5 nodos (neuronas).

Con respecto a las neuronas, su valor se extraerá de realizar una suma de las señales que recibe, para ser pasadas posteriormente por una función de activación. Normalmen-

te, a este sumatorio se le agrega un valor llamado prejuicio (*bias*). Se ha demostrado que esta adición puede reducir el tiempo necesario de entrenamiento de una red neuronal.

Este sistema toma una serie señales introducidas a través de la capa de entrada, normalizadas a unos pesos independientes entre si con valores entre 0 y 1, que a su vez emitirán estas señales a cada una de las neuronas de la primera capa oculta. Aquí, las señales tendrán que pasar por una **función de activación** que determinará si esa neurona se activa o no, para después pasar la respectiva señal a la siguiente capa. Finalmente, se obtendrá un valor de salida de las neuronas de la última capa correspondiente a 0 o 1, dependiendo de la clasificación que la red haya determinado. Este concepto se conoce como *forward-propagation*.

Dado que nuestro entorno se trata de un entrono de **aprendizaje supervisado**, se puede determinar el valor correcto que las capas exteriores deben tomar. El concepto del *backpropagation* utiliza esta ventaja que nos ofrecen los entornos deterministas para realizar así una correción de cómo el sistema maneja las señales de entrada. Así, la red neuronal es capaz de ajustar sus funciones de activación para poder realizar una mejor clasificación en pruebas posteriores.

2.1.5. El concepto de Q-Learning

2.1.6. Redes convolucionales (CNNs)

Las redes neuronales clásicas no están preparadas para soportar inputs de imágenes, ya que las señales de entrada necesitan tener ya unos valores numéricos normalizados. Por ello es necesario que las imágenes pasen por un tratamiento previo a su introducción en una red neuronal clásica. Este tipo de sistemas se llaman redes neuronales convolucionales.

El preprocesamiento de imágenes llevado a cabo en este tipo de imágenes viene ilustrado en la siguiente figura:

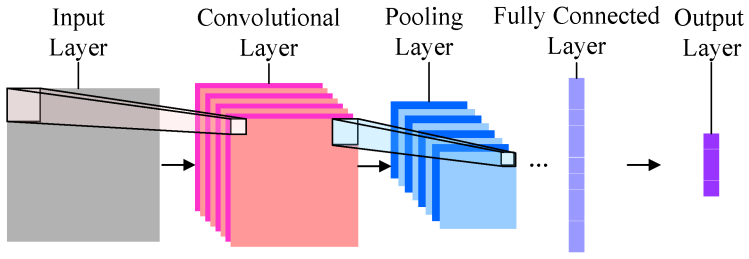

Figura 2.3: Ejemplo de red convolucional

Como se puede observar, la red se puede dividir en diferentes capas:

Capa convolucional o detector de características. Se utiliza para filtrar la imagen de entrada de manera que el resultado sea una matriz de menor orden conteniendo las características de la imagen que sean útiles para la clasificación. Normalmente se aplican diferentes filtros en búsqueda de diversas características, que luego se agrupan y se pasan a la siguiente capa.

Capa de agrupamiento. Reduce de nuevo el tamaño de la matriz de características, extrayendo de ella la información más remarcable de cada grupo de píxeles. De esta manera, se puede reducir el tamaño de la imagen hasta en un 75 %.

Capa de aplanamiento. Adicionalmente, se utiliza una capa extra que toma los mapas de características resultantes de la capa de agrupamiento y crea con ellos una matriz de tamaño n x 1, que a su vez será la capa de entrada de la red neuronal.

2.1.7. El problema de la memoria. El *experience replay*

Normalmente, para una correcto aprendizaje en entornos estocásticos[2], no basta con la implementación de una red neuronal. En su definición más simple, los algoritmos de aprendizaje descartan la información respectiva a experiencias anteriores inmediatamente después de decidir la acción a tomar.

Resulta útil un mecanismo que sea capaz de almacenar y usar de nuevo las experiencias pasadas del agente, ya que soluciona distintos problemas que pueden surgir si no se sigue este enfoque. Por ejemplo, es interesante el tener constancia de estados poco comunes que requieran unas acciones específicas para conseguir un resultado favorable en posteriores iteraciones. Este mecanismo es el denominado *experience replay*.

El resultado de realizar una implementación de estas características constituye en una menor necesidad de los agentes de experiencias para converger, a cambio de poder de cómputo y almacenamiento de datos - recursos accesibles con un menor coste temporal.

Adicionalmente, [7] argumenta que priorizar las experiencias que el agente accede en memoria puede tener un efecto positivo en la rápida convergencia del aprendizaje, partiendo de la idea de que existirán experiencias de las que el agente podrá aprender de manera más efectiva que otras, o bien existir algunas que puedan no ser útiles para un estado determinado y que más tarde pasen a ser vitales para la correcta ejecución de cierta secuencia de acciones. El principal problema de esta ordenación por prioridad será entonces determinar el criterio por el cual se mide la importancia de cada transición de estados. Idealmente, esto se realizaría tomando como criterio la cantidad de información que el agente puede aprender de la transición desde el estado actual. Sin embargo, este no es un factor que podamos calcular. En su lugar, una aproximación adecuada podría ser tomar el error TD de una transición δ, lo cual nos daría una noción de cómo de 'sorprendente' o innovadora determinada acción resultará.

Esta solución no viene sin problemas, ya que este algoritmo de priorización acabará centrándose en un pequeño grupo de experiencias que se repetirán de manera frecuente. Además el error se reducirá lentamente con el paso de cada iteración, lo cual podría desembocar a su vez en problemas de sobreajuste[3].

[7] concluye después de realizar experimentaciones en el entorno de Atari 2600, que una correcta implementación del *experience replay* basado en prioridades podría aumentar el aprendizaje conseguido por un factor de 2.

[2] Un entorno estocástico es aquel con un número finito de estados definidos.

[3] El sobreajuste o *overfitting* en inglés, es el efecto de sobreentrenar un algoritmo de aprendizaje para encontrar soluciones que ya se conocen.

2.2 Historia de la evolución tecnológica

2.3 Algoritmos existentes en la actualidad

2.4 Aplicación en videojuegos

Por norma general, los videojuegos constituyen un marco idóneo para la investigación en los campos del aprendizaje por refuerzo por diversas razones. En primer lugar, suelen ser productos inicialmente diseñados para ser controlados por jugadores humanos, por lo que pueden constituir un buen marco de referencia para la valoración del aprendizaje que un agente lleva a cabo. Además, estos sistemas proveen un entorno con unas reglas ya marcadas, en los que la información que se provee al jugador suele ser suficiente para valorar si se están tomando o no las decisiones correctas, bien a través de una puntuación atribuida a la partida, la duración de la misma o la consecución de una serie de objetivos previamente definidos. Por último, el reducido número de acciones que el jugador puede realizar (dimensionalidad) permite crear correlaciones mucho más exactas entre los pares acción recompensa, y facilita enormemente la tarea de aprendizaje.

En este documento, nos centraremos en la aplicación de los algoritmos previamente definidos a videojuegos de la videoconsola Atari 2600, comercializada en 1977 y con un extenso catálogo consistente en famosos juegos arcade como Pacman, Pong, Pinball o Space Invaders. Estos juegos fueron inicialmente diseñados para resultar desafiantes y difíciles de dominar por jugadores humanos, pero dado su baja dimensionalidad, constituyen entornos de estudio muy interesantes.

En estudios recientes [9], se demostró la capacidad de una variante del algoritmo DQN de superar exponencialmente el rendimiento de un jugador humano sobre los mismos entornos. En este estudio se hizo uso de una red convolucional que aplicaba los principio de la diferencia temporal para determinar factores como la trayectoria de los objetos o la diferencia de puntuación, atribuyendo así una correlación entre las acciones llevadas a cabo por el agente y su rendimiento a corto plazo. Adicionalmente, en estas implementaciones se hizo uso de un mecanismo de *experience replay* que aseguró que el agente fuera capaz de usar las experiencias del pasado para mejorar su rendimiento, y una función que actualizara los valores de acción (*Q-values*) de manera periódica en lugar de instantánea para reducir las correlaciones entre acciones específicas y su resultado inmediato. La combinación de estas dos técnicas resultó en una mejora de hasta un 2800 % sobre el rendimiento máximo de un jugador humano en el entorno de Video Pinball.

CAPÍTULO 3
??? ???? ??????

????? ????????????? ????????????? ????????????? ????????????? ?????????????

3.1 ?? ???? ???? ? ?? ??

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 4
Conclusions

????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

[1] Jennifer S. Light. When computers were women. *Technology and Culture*, 40:3:455–483, juliol, 1999.

[2] Georges Ifrah. *Historia universal de las cifras*. Espasa Calpe, S.A., Madrid, sisena edició, 2008.

[3] Comunicat de premsa del Departament de la Guerra, emés el 16 de febrer de 1946. Consultat a http://americanhistory.si.edu/comphist/pr1.pdf.

[4] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning, febrero de 2016. *Google DeepMind, Montreal Institute for Learning Algorithms (MILA), University of Montreal*

[5] Michel Tokic Adaptive ε-greedy Exploration in Reinforcement Learning Based on Value Differences. *Institute of Applied Research, University of Applied Sciences Ravensburg-Weingarten, 88241 Weingarten, Germany*

[6] Jianxin Wu Introduction to Convolutional Neural Networks, mayo de 2017. *LAMDA Group National Key Lab for Novel Software Technology. Nanjing University, China*

[7] Tom Schaul, John Quan, Ioannis Antonoglou, David Silver Prioritized Experience Replay, febrero de 2016. *Google DeepMind*

[8] Wolfram Schultz, Peter Dayan, P. Read Montague A Neural Substrate of Prediction and Reward Science 275, 1593-1599 (1997)

[9] Varios autores Human-level control through deep reinforcement learning *Nature*, vol 518, 26 de febrero de 2015

APÉNDICE A

Configuració del sistema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.1 Fase d'inicialització

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.2 Identificació de dispositius

????? ????????????? ????????????? ????????????? ????????????? ?????????????

APÉNDICE B
??? ????????????? ????

????? ????????????? ????????????? ????????????? ????????????? ?????????????