Plan de investigación

JUSTIFICACIÓN DEL TEMA DE ESTUDIO

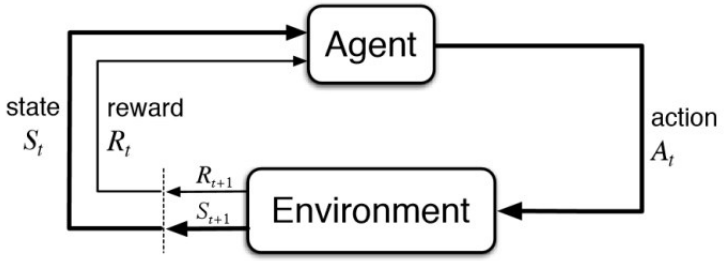
El estado del arte define la inteligencia como la capacidad de adaptarse a un conjunto diverso de entornos complejos de manera eficiente [1]. Esta definición sugiere que la capacidad máxima o el límite máximo de la inteligencia viene determinado por el entorno en el que un agente desea desenvolverse. Los humanos, por ejemplo, con el paso de los años vamos adquiriendo las competencias necesarias para desenvolvernos de manera satisfactoria de forma progresiva a lo largo de la vida, cada vez de manera más eficiente, según vamos aprendiendo de nuestros propios errores. Además, algunos autores [2] también afirman que entornos de diversidad más y más complejos proveen a los agentes que interactúan en ellos de un nivel de inteligencia aún mayor.

La inteligencia artificial trata de proveer a las máquinas esta capacidad de aprendizaje que poseemos los seres humanos de forma latente. Sin embargo, esta tarea no es nada trivial ya que las máquinas no poseen ningún conocimiento del entorno. En este sentido, una gran fuente para el desarrollo de la inteligencia artificial ha sido la **imitación** del proceso de aprendizaje humano. Uno de los ejemplos más claros de este proceso cognitivo son las redes neuronales artificiales, las cuales simulan matemáticamente el proceso biológico que acontece en las activaciones de las neuronas del cerebro humano [3]. Otra fuente de inspiración ha sido el de imitar el proceso incremental de aprendizaje de las personas, en particular, el aprendizaje por refuerzo.

El aprendizaje por refuerzo (*reinforcement learning)* premia o castiga en función de si el agente sujeto de estudio se ha comportado de la forma deseada, incentivando las “buenas” acciones y desalentando las “malas”. En base a esta idea se pueden diseñar algoritmos que sean inteligentes o al menos **racionales** en función de un objetivo determinado [4].

En el aprendizaje por refuerzo de un **único agente** el algoritmo o la función de aprendizaje sólo debe considerar las acciones o decisiones a realizar por parte de un individuo que interactúa en un entorno dado. En este escenario la complejidad del aprendizaje viene determinada tanto por el entorno como por la variedad de las acciones disponibles [4]. Entre los grandes hitos de este campo se encuentran la victoria de las máquinas frente a los

humanos en el juego del ajedrez [5] o el Go! [6]. A pesar de estos éxitos, y del gran potencial de aprendizaje que ha demostrado el aprendizaje por refuerzo, en un entorno más realista es difícil encontrar un entorno donde opere un único agente inteligente.



Esquema del aprendizaje por refuerzo.Fuente: https://www.kdnuggets.com/images/reinforcement-learning-fig1-700.jpg

En el mundo en el que vivimos es más común encontrar un entorno donde operan **multitud de agentes que interactúan**, determinan y coaccionan el ambiente en el que a su vez el resto de agentes se deben adaptar. Esta adaptación respecto al resto de agentes a la vez que la propia necesidad de interactuar y compartir recursos con terceros es una gran fuente de complejidad y a su vez de aprendizaje que los algoritmos de aprendizaje por refuerzo del estado del arte no son capaces de cubrir. El aprendizaje multi agente (*Multi-agent reinforcement learning*) es precisamente el área en la que se estudia el desarrollo y análisis de formas de aprendizaje y algoritmos que descubran estrategias cooperativas-competitivas efectivas en entornos de múltiples agentes. Esta disciplina está cosechando grandes éxitos como en el videojuego Strarcraft [8], aunque todavía tiene un gran potencial por explotar.

OBJETIVO PRINCIPAL

El objetivo de esta tesis es precisamente el de profundizar en esta línea de investigación con el fin de avanzar en el **estado del arte referente al aprendizaje por refuerzo en entornos cooperativos-competitivos**. Esta línea de investigación tiene multitud de aplicaciones reales, como pudieran ser: la conducción autónoma, la robótica en el contexto del internet de las cosas, o la automatización industrial. Todas estas aplicaciones tienen en común que cuentan con un entorno complejo conformado por un gran número de agentes que deben tomar decisiones independientes tratando de maximizar un objetivo común, incluyendo casos en los que un agente determinado deba optar por tomar decisiones que lo perjudiquen a favor de una recompensa global de valor más elevado.

Marco conceptual

Para estudiar el marco conceptual del aprendizaje por refuerzo multi agente se ha dividido la tarea en tres fases:

1. En primer lugar, se comenzará por un estudio del marco general de aprendizaje por refuerzo, del aprendizaje profundo y del curriculum learning.
2. En segundo lugar, se hará un estudio para superar las dificultades que surgen al considerar múltiples agentes, en particular, cómo se deben comunicar, colaborar y reciprocar.
3. Por último, se aplicarán los algoritmos desarrollados en entornos realistas siendo una línea muy interesante la robótica colaborativa.

En una primera fase, se estudiará el problema del aprendizaje por refuerzo individual y colaborativo en el estado del arte. Con el fin de evaluar los resultados de la investigación, utilizaremos como base y referencia las tres principales vertientes de investigación relacionadas con el aprendizaje por refuerzo hasta el momento, que son:

1. Algoritmos tradicionales de aprendizaje por refuerzo
2. Deep Learning y algoritmos basados en la evolución (computación evolutiva)
3. Aprendizaje por curriculum

Principalmente se utilizará una metodología empírica que permitirá evaluar las distintas propuestas a desarrollar en este proyecto de tesis contra las propuestas ya conocidas en el estado del arte. A continuación se describen brevemente las ideas más relevantes a analizar dentro de estas áreas.

Algoritmos tradicionales de aprendizaje por refuerzo

Durante el aprendizaje por refuerzo el agente interactúa secuencialmente con un entorno. Para ello es necesario definir una función de transición o **política de comportamiento** que especifica la probabilidad de tomar una decisión dentro de un espacio de opciones. Es por lo tanto un proceso de Markov pues la probabilidad del siguiente estado depende total o parcialmente del estado actual. También es necesario definir una función de recompensa que asigna un valor de retorno a cada acción tomada. El objetivo del agente será actualizar su estrategia o política de comportamiento para maximizar esta recompensa.

Creemos que en una fase inicial de la tesis necesitamos analizar y estudiar conceptos provenientes de la Teoría de Juegos [9] como el equilibrio de Nash [10], para posteriormente ser capaces de diseñar algoritmos que vayan más allá del estado del arte.

Deep Learning

La función de transición y la actualización de la política de comportamiento y su relación con la obtención de las recompensas puede ser modelizado mediante redes neuronales profundas. Este tipo de algoritmos han demostrado su eficacia con otros problemas de visión, procesamiento del lenguaje natural y el aprendizaje por refuerzo. Dentro de la primera fase de la tesis, consideramos muy necesario **explorar métodos de aprendizaje por refuerzo** como el Q-learning [11], como algoritmos que han sido capaces de superar a los algoritmos tradicionales de aprendizaje por refuerzo. La exploración de este tipo de técnicas nos permitirá conocer las librerías y utilidades actuales para la implementación de sistemas, como: Pytorch, Tensorflow, Keras, OpenAI Gym y otros.

Dentro del análisis de algoritmos complejos, no descartamos explorar también algoritmos basados en la optimización combinatoria, debido al reciente éxito que han cosechado, como los algoritmos de búsquedas locales utilizados en AlphaGo (Monte Carlo Tree Search) o los conocidos algoritmos evolutivos que se basan en la evolución Darwiniana de poblaciones generacionales [12].

Aprendizaje por curriculum

Los humanos al igual que el resto de los animales aprendemos de manera más eficiente cuando los ejemplos de aprendizaje se nos presentan de manera organizada, con un orden significativo que ilustran gradualmente conceptos más variados y complejos. En [13] los autores formalizan la descrita estrategia de capacitación en el contexto del aprendizaje automático, y lo describen como "aprendizaje curricular". La idea clave de esta nueva estrategia de aprendizaje gradual reside en explotar conceptos previamente aprendidos para **facilitar el aprendizaje de nuevas abstracciones**. Para ello resulta imprescindible la correcta selección de qué ejemplos de aprendizaje y en qué orden se van a presentar al sistema. De esta manera, el objetivo principal es el de guiar el aprendizaje del sistema para optimizar el rendimiento y los resultados. A su vez, el uso del

aprendizaje por currículum es una vertiente que no ha sido extensamente utilizada junto a las técnicas mencionadas anteriormente, pero que podría dar lugar a algoritmos novedosos. Consideramos que la exploración conjunta de estas líneas de investigación es especialmente relevante en el contexto de esta tesis.

En una segunda fase, se estudiarán mediante técnicas de inteligencia artificial las estrategias para superar las barreras existentes en los modelos del estado del arte, y explotar las posibilidades de optimizar un aprendizaje colaborativo entre agentes. En un principio pretendemos explorar las siguientes dimensiones:

Dimensión 1: Aprender a colaborar

Consiste en la exploración de métodos para evaluar la contribución que ha realizado cada agente en un juego cooperativo con respecto al éxito del conjunto cuando la recompensa es global [14]. Aunque sea posible deducir funciones que permiten deducir la contribución individual de cada agente, determinar la contribución global de un agente es algo que va más allá del estado del arte actual.

Dimensión 2: Aprender a comunicar

La comunicación entre agentes puede suponer una ventaja fundamental a la hora de afrontar ciertos entornos. Por lo tanto, es necesario explorar técnicas de comunicación o protocolos de comunicación cooperativos.

Dimensión 3: Aprender a reciprocar

Cuando los agentes persiguen objetivos diferentes pueden provocar conflictos y resultados indeseados. Un ejemplo claro de esto es la conducción autónoma en la que cada coche cooperará con otros agentes (coches, peatones, …). En este contexto, cada agente tiene objetivos individuales que en situaciones extremas pueden entrar en conflicto. Sin embargo, la reciprocidad entre agentes es un aspecto fundamental para obtener un bien general.

En una tercera fase, se analizará la generalización del sistema desarrollado con el fin de aplicarlo en un entorno más realista en colaboración con el **Immersive Lab de Deusto**.

Imagen del Deusto Immersive Lab. Fuente:
https://agenda.deusto.es/wp-content/uploads/2019/11/7-Deusto-141-Laboratorio-Virtualware.jpg

El Immersive Lab de Deusto es un laboratorio fruto de la colaboración entre la empresa vasca Virtualware y la Universidad de Deusto. Este laboratorio permitirá la realización de experimentos enmarcados en la **robótica colaborativa**, en la que un usuario interactúa con uno o más agentes. Los casos de uso, y por lo tanto el usuario y el entorno virtual, pueden ser muy variados, desde un entorno industrial en el que se simula un operario interactuando con una máquina hasta un entorno doméstico en el que un robot asiste o acompaña a una persona de la tercera edad. La principal virtud de realizar experimentos en un laboratorio virtual consiste en la seguridad de las personas, aunque también hay que considerar que las posibles mejoras en eficiencia y coste que puede suponer en función del caso de uso. Esta seguridad se explica por ejemplo en el caso de la industria si la máquina está operando con materiales peligrosos.

Además cabe mencionar que no sería necesaria centrarse únicamente en una línea de investigación. Una de las características fundamentales de la inteligencia artificial y en particular del aprendizaje por refuerzo es la capacidad de aplicarse a diversas problemáticas. Es decir, un algoritmo que funcione para controlar un brazo robot puede servir para un robot asistencial. Por lo tanto, se destacan dos líneas de investigación adicionales que se consideran relevantes por su interés científico y por su relación a proyectos en los que participa la Universidad de Deusto.

1. Inteligencia ambiental y su uso para la **creación de espacios asistenciales inteligentes**. La importancia de la creación de agentes que ayuden a las personas en posibilitar una vida buena y digna es cada vez más notoria, destacando los grupos de riesgo como la tercera edad o personas que padecen enfermedades. Estos grupos tienen necesidades y capacidades concretas y es necesario que los agentes que interactúan con estos usuarios sean capaces de adaptarse. Esta línea de investigación está particularmente bien relacionada con la colaboración con el Immersive Lab.
2. Aplicaciones a las **ciudades inteligentes**. Las ciudades inteligentes o *smart cities* son una área de investigación cada vez más relevante. En esta área surgen multitud de problemáticas diferentes como la gestión de la energía, la optimización de rutas de robots autónomos urbanos de limpieza u otras aplicaciones. Se ha comenzado a realizar avances en estas áreas, [15] y [16] por ejemplo, pero queda mucho por profundizar.

Hipótesis

El proyecto de tesis se podría resumir en la siguiente hipótesis:
"El desarrollo de algoritmos de aprendizaje por refuerzo cooperativo-competitivo optimizará las capacidades de aprendizaje de agentes inteligentes en entornos cooperativos-colaborativos para la resolución conjunta de objetivos"

Metodología

Para desarrollar la tesis se ha seguido y se va seguir una metodología de investigación inspirada en la investigación en acción donde la experiencia guiará la pregunta de investigación. En resumen, se seguirán los siguientes pasos:

- Identificación del problema. Este paso consiste en diagnosticar un problema cuya solución supondrá un avance en la línea de investigación o que se podrá emplear de manera positiva.
- Planificación. Se realizará un estudio detallado de la solución para trazar el conjunto de acciones que se pueden realizar. Para valorarlas habrá que profundizar en la solución, determinado a que situación se quiere llegar o qué restricciones puede tener.

- Puesta en marcha. En base a este análisis se escogerá el plan a ejecutar.
- Evaluación y *feedback*. Tras la ejecución se evaluarán los resultados obtenidos. En la medida de los posible, se buscará feedback externo de expertos en conferencias científicas para así poder evaluar los resultados de manera más crítica y objetiva.
- Reflexión. En esta última fase se reflexionará para destilar las lecciones aprendidas para poder mejorar y compartirlas con la comunidad científica.

Esta secuencia de pasos no es necesariamente secuencial. Por ejemplo, la retroalimentación de expertos puede modificar el diagnóstico del problema o la acción escogida.

Bibliografía

[1] Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. arXiv preprint arXiv:1903.00742.

[2] Wu, Y., Wu, Y., Gkioxari, G., & Tian, Y. (2018). Building generalizable agents with a realistic and rich 3d environment. arXiv preprint arXiv:1801.02209.

[3] Chua, L. O., & Yang, L. (1988). Cellular neural networks: Theory. IEEE Transactions on circuits and systems, 35(10), 1257-1272.

[4] Russel, S., & Norvig, P. (2013). Artificial intelligence: a modern approach. Pearson Education Limited.

[5] Newborn, M. (2012). Kasparov versus Deep Blue: Computer chess comes of age. Springer Science & Business Media.

[6] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484.

[7] Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. arXiv preprint arXiv:1903.00742.

[8] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Oh, J. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), 350-354.

[9] Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games, 2(28), 307-317.

[10] Nash, J. F. (1950). Equilibrium points in n-person games. Proceedings of the national academy of sciences, 36(1), 48-49.

[11] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.
[12] Bull, L. (1998, March). Evolutionary computing in multi-agent environments: Operators. In International conference on evolutionary programming (pp. 43-52). Springer, Berlin, Heidelberg.
[13] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41-48).
[14] Chang, Y. H., Ho, T., & Kaelbling, L. P. (2004). All learning is local: Multi-agent learning in global reward games. In Advances in neural information processing systems (pp. 807-814).
[15] Hirata, T., Malla, D. B., Sakamoto, K., Yamaguchi, K., Okada, Y., & Sogabe, T. (2019). Smart Grid Optimization by Deep Reinforcement Learning over Discrete and Continuous Action Space. *Bulletin of Networking, Computing, Systems, and Software*, *8*(1), 19-22.
[16] Dowling, J., Curran, E., Cunningham, R., & Cahill, V. (2005). Using feedback in collaborative reinforcement learning to adaptively optimize MANET routing. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *35*(3), 360-372.

Medios y recursos materiales disponibles

La tesis doctoral se enmarca dentro de proyectos activos relacionados con el aprendizaje por refuerzo que darán soporte a la misma entre los que habría que destacar:

VRAIGYM: Ai-powered Virtual Training Environment For Collaborative Robotics (Convocatoria del Programa de Ayudas a la Investigación Colaborativa en áreas estratégicas 2020). El objetivo del proyecto VRAIGYM es explorar los fundamentos del diseño de entornos de entrenamiento de realidad virtual con sistemas de inteligencia artificial integrados en los que operarios humanos y robots colaborativos puedan llevar a cabo sesiones de trabajo y entrenamiento conjunto de manera virtual, segura, económica y con ágil configuración y despliegue. El objetivo I2 del proyecto es la “Investigación en técnicas de Inteligencia Artificial para el aprendizaje automático de robots colaborativos en entornos virtuales” que se alinea con la presente propuesta.

ACROBA: AI-driven Cognitive Robotic Platform For Agile Production Environments (en preparación para call ICT-46 H2020). El objetivo del proyecto es diseñar mecanismos de aprendizaje por refuerzo single-agent y multi-agent para la creación de entornos de producción basados en robots autónomos y robots colaborativos que permitan una ágil reconfiguración que satisfaga la variabilidad en la demanda en procesos de producción.

Además se estudiará la relación con **empresas y centros de investigación** como el Instituto Ibermática en Innovación (I3B), Virtualware, Ikerlan y Tecnalia, con los que el grupo de investigación viene colaborando desde hace años tanto en proyectos como en desarrollo de tesis doctorales, incluyendo la posibilidad de que el doctorando realice estancias en estos centros de investigación aplicada, dado el interés y la estrecha relación entre empresa, doctorando y contenidos de la tesis.

- Instituto Ibermática de Innovación (I3B). Es la unidad de I+D+i empresarial asociada a la corporación Ibermática, en la que se desarrollan proyectos de investigación y exploración tecnológica para evaluar su futura incorporación a líneas de negocio. El contacto es Aitor Moreno, Responsable del Área de Inteligencia Artificial

- Virtualware 2007, S.A. Uno de los líderes europeos en simulaciones de realidad virtual para propósitos industriales, con los que se mantiene una estrecha colaboración en proyectos de I+D para añadir capacidades a sus sistemas, incluyendo nuevas técnicas de Inteligencia Artificial para la simulación de escenarios con robots inteligentes. El contacto es Sergio Barrera, CTO (Chief Technical Officer).
- Ikerlan. Centro de investigación tecnológica de la Corporación Mondragón con el que hay una estrecha colaboración tanto en proyectos de investigación como en la codirección de tesis doctorales. El contacto es Josu Bilbao, Head of Research Department - ICT (IoT Digital Platforms, Data Analytics & Artificial Intelligence).
- Tecnalia. Centro de investigación tecnológica con el que hay una estrecha colaboración tanto en proyectos de investigación como en la codirección de tesis doctorales. El contacto es Ana Ayerbe, Director of the IT Competitiveness area.

Planificación temporal hasta la conclusión de la tesis

Este apartado ofrece un avance del plan de trabajo que se seguirá para el proyecto de tesis doctoral propuesto. La investigación está planificada para desarrollarse durante 6

semestres. El plan se presenta en diferentes etapas ligadas a los objetivos principales de la investigación:

1. Elaborar un **estudio profundo y completo** del estado del arte que pueda servir de referencia para el diseño de nuevos algoritmos más allá del estado del arte.
Semestres 1,2,4,5
2. Elaborar un **análisis de los recursos disponibles** y de las medidas de evaluación de los agentes, incluyendo la réplica, reproducción y puesta en marcha de estos recursos.
Semestre 2.
3. Partiendo del conocimiento adquirido durante la primera y segunda etapa, **elaborar una definición** concreta de nuevos modelos para el aprendizaje colaborativo de sistemas multiagente, estableciendo una clara relación respecto a las técnicas de referencia analizadas y el problema a abordar.
Semestre 2.
4. **Realizar un estudio experimental** basado en sistemas del estado del arte en aprendizaje por refuerzo y estudiar su tecnología concreta de implementación.
Semestre 3.
5. **Elaborar una definición y un plan concreto de implementación** que permita realizar una evaluación de los nuevos algoritmos desarrollados contra los modelos analizados en el estado del arte.
Semestre 3.
6. **Implementar y poner a prueba el modelo diseñado**. Semestres 3,4,5.
7. **Evaluación, documentación y retroalimentación** del ciclo del proyecto con el objetivo de analizar el cumplimiento y grado de satisfacción de los objetivos inicialmente definidos.
Semestre 5.
8. **Escritura** de la documentación de la tesis. Semestres 5,6.

En formato gráfico, la distribución de las tareas en los meses previstos sería la siguiente, donde la intensidad del color marca el esfuerzo relativo que se le dedicará a la tarea en un semestre determinado:

	Año 1		Año 2		Año 3	
	Sem 1	Sem 2	Sem 3	Sem 4	Sem 5	Sem 6
Tarea 1						
Tarea 2						
Tarea 3						
Tarea 4						
Tarea 5						
Tarea 6						
Tarea 7						
Tarea 8						

Previsión de movilidad para la mención internacional

Para obtener la mención internacional de la tesis y para que la experiencia doctoral sea más diversa y completa se prevé una estancia internacional en una universidad o centro de investigación extranjero. Se estima que la oportunidad principal para definir la estancia nacerá del proyecto ACROBA pues participan diversos centros europeos como el “BREMER INSTITUT FUER PRODUKTION UND LOGISTIK GMBH” en Alemania o “CLERMONT AUVERGNE” en Francia. Esta estancia tendrá como objetivos, entre otros, profundizar en las técnicas de aprendizaje por refuerzo desarrolladas, conocer y trabajar con investigadores europeos y crear relaciones para futuros proyectos de investigación.