APRENDIZAJE POR **REFUERZO EXPLICATIVO** BASADO EN **MEMORIA EPISÓDICA**

Francisco Cruz
Doctor en Ciencias de la Computación.

El aprendizaje por refuerzo (reinforcement learning, RL) es un enfoque de aprendizaje basado en la psicología conductual, utilizada por agentes artificiales para aprender de forma autónoma mediante interacción con su entorno. Como en muchas otras técnicas de aprendizaje automático, RL se considera una caja negra desde la perspectiva del modelo y, por lo tanto, un problema aún abierto es la falta de visibilidad y comprensión para los usuarios finales en términos de las decisiones tomadas por un agente durante el proceso de aprendizaje. Una forma de superar este problema es proporcionar al agente la capacidad de explicar en términos simples por qué se toma una acción particular en una situación particular. En esta investigación, proponemos un enfoque de aprendizaje por refuerzo explicativo basado en memoria (memory-based explainable reinforcement learning, MXRL). Por medio del uso de una memoria de transiciones, el agente puede explicar sus decisiones utilizando la probabilidad de éxito y el número de transacciones para alcanzar el estado objetivo. El escenario experimental considera dos escenarios simulados: una grilla sin límites con regiones aversivas y una grilla limitada. Los resultados obtenidos muestran que el agente, utilizando información extraída de la memoria, puede explicar su comportamiento de manera comprensible para usuarios finales no expertos en cualquier momento durante su operación.

Introducción

El objetivo del aprendizaje por refuerzo (RL) [1] es proporcionar a un agente autónomo la capacidad de aprender nuevas habilidades solamente interactuando con su entorno. RL ha demostrado ser un enfoque de aprendizaje efectivo, sin embargo, un problema aún abierto es la falta de un mecanismo que permita comunicar claramente las razones por las que se eligen ciertas acciones dado un estado particular.

Pese a que normalmente los agentes artificiales se consideran cajas negras, a menudo, es posible proporcionar interpretaciones técnicas del por qué se deciden ciertas acciones, por ejemplo, un agente de RL podría explicar su comportamiento en términos de valores Q y recompensas futuras [2]. Sin embargo, este tipo de explicación no tiene mucho sentido para usuarios no expertos a quienes se les debe proporcionar explicaciones utilizando un lenguaje similar al dominio del problema para permitirles comprender completamente el comportamiento del agente.

En este trabajo, proponemos un enfoque de aprendizaje por refuerzo explicable basado en memoria (memory-based explainable reinforcement learning, MXRL), que permita a un agente explicar en el lenguaje del dominio la decisión de seleccionar una acción sobre las otras posibles. En nuestro enfoque, se dan explicaciones utilizando la probabilidad de éxito y la cantidad de transiciones necesarias para alcanzar el estado objetivo. Por lo tanto, un agente puede explicar su comportamiento no sólo en términos de valores Q ni en la probabilidad de seleccionar una acción, sino más bien, en términos de la necesidad de completar la tarea prevista.

Inteligencia Artificial Explicable

La inteligencia artificial (IA) está recibiendo cada día más atención en diferentes áreas de nuestra vida diaria. Las aplicaciones en campos como robótica, conducción autónoma de automóviles, asistencia en casa, videojuegos, entre otros, se muestran a diario en los medios de comunicación [3]. En los últimos años, la inteligencia artificial explicable (explainable artificial intelligence, XAI) ha surgido como un área de investigación prominente que tiene como objetivo proporcionar a los sistemas basados en IA la capacidad de dar explicaciones sencillas usuarios finales no expertos [4].

La idea detrás de XAI no sólo tiene como objetivo proporcionar explicaciones, sino también permitir que un sistema de IA: justifique sus decisiones y resultados, controle y prevenga problemas, mejore su comportamiento y descubra nuevo conocimiento [5]. La necesidad de XAI está motivada principalmente por la necesidad de confianza, interacción y transparencia entre usuarios finales y sistemas basados en IA.

XAI es un campo amplio, como la misma IA, con aplicaciones en áreas como transporte, finanzas, medicina y militares, entre otras [3]. En el campo de la interacción humano-robot (human-robot interaction, HRI), el término de agente explicativo se ha utilizado para referirse a robots dedicados a responder preguntas sobre sus razones durante el proceso de toma de decisiones. Langley et al. [6] proponen los elementos de agentes explicativos como el contenido que respalda las explicaciones, una memoria episódica para registrar estados y acciones, además de acceso a su experiencia previa.

Aprendizaje por Refuerzo Explicativo Basado en Memoria

El comportamiento de un agente de RL podría explicarse técnicamente en términos de los valores Q o también en términos algorítmicos. Sin embargo, en este trabajo, buscamos explicaciones que tengan sentido para todo tipo de posible usuario final y no sólo para aquellos que pueden comprender el proceso de aprendizaje subyacente detrás de un agente artificial. En este sentido, buscamos obtener explicaciones de manera similar a como se realizaría en interacción de agentes cognitivos, es decir, mediante el uso de un lenguaje relativo al dominio.

Desde una perspectiva del usuario final, podemos considerar las preguntas más relevantes como: ¿por qué? y ¿por qué no? [7]. Por ejemplo, las siguientes preguntas pueden hacerse a un agente artificial para comprender mejor su comportamiento:

- ¿Por qué diste un paso hacia adelante en el último movimiento?
- ¿Por qué no giraste a la derecha en esta situación?

Por lo tanto, para responder estas preguntas usando un lenguaje comprensible, nuestras explicaciones tienen la intención de determinar:

- La probabilidad de éxito del agente artificial.
- El número de transiciones para alcanzar el estado objetivo, para terminar la tarea o finalizarla dentro de un marco de tiempo.

Para lo anterior, proponemos un enfoque de aprendizaje por refuerzo explicable basado en memoria (MXRL) para calcular la probabilidad de éxito Ps y las transiciones a la meta Nt. Nuestro enfoque consiste de un agente de RL con memoria de transiciones. Para esto implementamos una lista con pares de estado-acción TList que comprende las transiciones que el agente realizó durante su proceso de aprendizaje.

Por un lado, para calcular la probabilidad de éxito Ps, calculamos previamente el número total de transiciones Tt y el número de transiciones involucradas en una secuencia exitosa Ts. Para obtener Ts, utilizamos las transiciones previamente guardadas en la lista Tlist. Cada vez que el agente alcanza el estado final, calculamos la probabilidad Ps = Ts / Tt considerando las transiciones involucradas en el camino hacia el estado objetivo. Por otro lado, las transiciones hasta el estado objetivo Nt se calculan cada vez que termina un episodio. Para cada estado, Nt está determinado por la posición en la lista TList ya que todas las transiciones se han guardado previamente allí. Por lo tanto, cada estado está tan lejos de la meta como su posición en la lista, es decir, índice + 1.

Como también queremos comparar la probabilidad de elegir una acción, calculada a partir de los valores Q, con la probabilidad de tener éxito, se ha implementado el método SARSA y el método de selección de acción softmax. El Algoritmo 1 muestra el enfo-

---

1. Initialize $Q(s,a)$, T_t, T_s, P_s, N_t
2. **for** each episode **do**
3. Initialize T_{List} []
4. Choose an action using $a_t \leftarrow$ SELECTACTION (S_t)
5. **repeat**
6. Take action a_t
7. Save state-action trasnition T_{list}.add (s, a)
8. T_t [s][a] $\leftarrow$ T_t [s][a] + 1
9. Observe reward r_t+1 and next state s_t+1
10. Choose next action a_t+1 using softmax action selection method
11. $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\,[r_t+1 + \gamma Q(s_{t+1},a_{t+1}) + Q(s_t,a_t)]$
12. $s_t \leftarrow s_t$+1; $a_t \leftarrow a_t$+1
13. **until** s is terminal (goal or aversive state)
14. **if** s is goal state **then**
15. **for** each s,a $\in T_{List}$ **do**
16. T_s [s][a] $\leftarrow$ T_s[s][a]+1
17. **end for**
18. **end if**
19. Compute $P_s \leftarrow T_s/T_t$
20. Compute N_t for each s $\in T_{List}$ as pos(s, T_{List}) + 1
21. **end for**

Algoritmo 1

Enfoque de aprendizaje de refuerzo explicable basado en memoria con el método SARSA para calcular la probabilidad de éxito y el número de transiciones hasta el estado objetivo.

que MXRL para entrenar agentes usando memoria de transiciones. Mientras que en la línea 7 cada par estado-acción ejecutado se guarda en la memoria, las líneas 19 y 20 calculan las probabilidades finales de éxito Ps y el número de transiciones hasta el estado objetivo Nt para cada episodio respectivamente.

Escenario Experimental

Para producir explicaciones, implementamos un escenario de grilla en dos versiones: acotada y sin límites. En otras palabras, el mismo par de estado-acción puede conducir a características diferentes para la explicación según el contexto. Utilizamos una grilla de 3x4, como se muestra en la Fig. 1. En la figura, es posible observar los 12 estados en los que puede estar el agente. El estado objetivo se muestra con un círculo verde en la parte inferior derecha. El círculo gris representa el agente que necesita encontrar el camino hacia el estado final. En cada episodio, el agente se ubica en una posición inicialmente aleatoria dentro de la grilla. Durante los episodios de entrenamiento, el agente tiene que aprender una política adecuada para llegar a la posición deseada. Hay cuatro acciones posibles en este escenario: abajo, arriba, derecha e izquierda.

En principio, consideramos una grilla ilimitada, es decir, el agente puede salir de la grilla y entrar en regiones aversivas que lleven a finalizar el episodio de aprendizaje actual y reiniciar uno nuevo. Las regiones aversivas se muestran en amarillo en la Fig. 1. En este caso, la probabilidad de tener éxito se calcula después de cada episodio de aprendizaje y depende de la experiencia de cada agente para alcanzar el estado final. Además, también hemos considerado una grilla delimitada, es decir, el agente no puede salir de ella. En otras palabras, cada vez que el agente intenta salir de la grilla, el estado actual no se actualiza, manteniendo la posición como estaba antes de seleccionar esa acción. En este contexto, el agente tiene una probabilidad de éxito constante igual a 1, ya que siempre puede completar la tarea, sin embargo, los movimientos necesarios para alcanzar la posición deseada son diferentes para cada estado alcanzado después de ejecutar una acción.

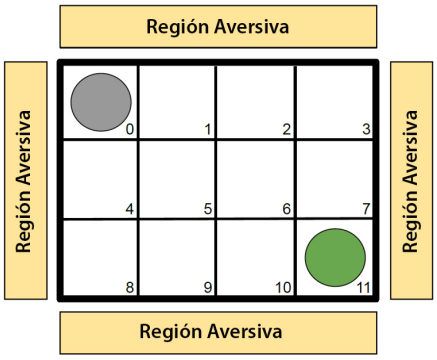


Figura 1

Grilla de 3x4 rodeada de regiones aversivas. El agente puede moverse en cuatro direcciones: abajo, arriba, derecha e izquierda. El círculo verde muestra el estado objetivo. Si el agente alcanza la región aversiva, el episodio de aprendizaje finaliza y se comienza uno nuevo. En grilla acotada, el agente no puede ingresar a las regiones aversivas.

Resultados Experimentales

Para el proceso de aprendizaje, la función de recompensa retorna un valor positivo igual a 1 cuando el agente alcanza el estado final y una recompensa negativa igual a -1 en caso de que el agente entre en la región aversiva. Todos los experimentos se han realizado utilizando el algoritmo de aprendizaje SARSA y el método de selección de acción softmax para el entrenamiento de 100 agentes. Los parámetros utilizados para el entrenamiento son: tasa de aprendizaje $\alpha = 0.3$, factor de descuento $\gamma = 0.9$ y temperatura softmax $\tau = 0.25$, todos ellos fueron determinados experimentalmente y relacionados con nuestro escenario. Los parámetros anteriores se mencionan aquí solo como referencia, pero no son relevantes para este trabajo. Estos parámetros ciertamente afectan la capacidad de los agentes para aprender una solución, sin embargo, estamos interesados en comprender la decisión tomada por un agente en lugar de la velocidad o la capacidad de aprender por parte de los agentes.

Grilla sin límites

Como ha sido mencionado, en la grilla ilimitada, el agente puede salir del escenario hacia la región aversiva. La figura 2 muestra los valores Q obtenidos, la probabilidad de elegir una acción, la probabilidad de éxito y el número de transiciones hacia el estado objetivo.

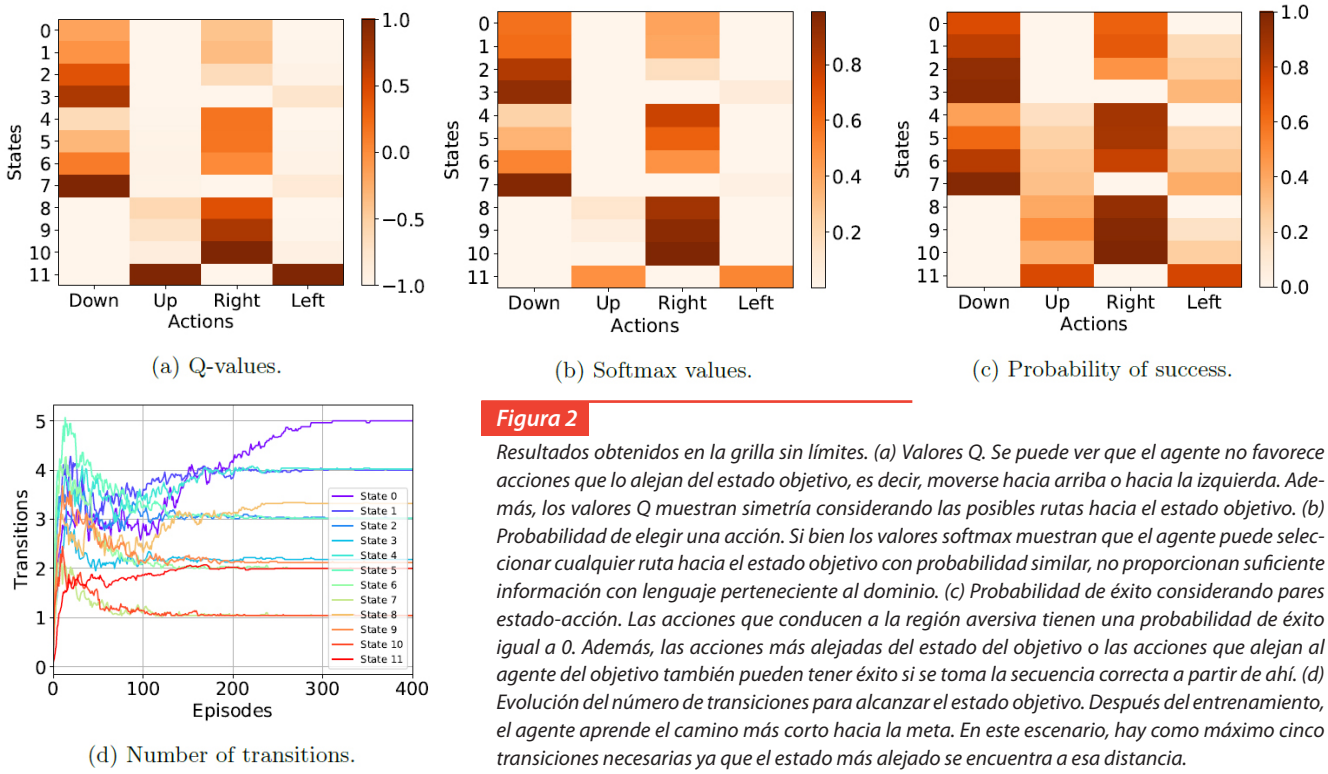
Después de completar el entrenamiento, los valores Q promedio se muestran en la Fig. 2a. Se puede observar que el agente no prefiere acciones como subir o bajar, ya que, independientemente del estado actual, siempre representan alejarse del estado objetivo. En términos generales, los valores Q muestran valores simétricos, lo que significa que el agente puede seleccionar cualquier ruta hacia la meta siempre que sus movimientos sean hacia abajo o hacia la derecha. Por supuesto, cuanto más cerca de la meta mayor la recompensa, por ejemplo, en los estados 7 y 10 ejecutando las acciones abajo o derecha respectivamente, ambos estados siendo adyacentes al estado final. Hay algunas excepciones que presentan un valor Q bajo al desplazarse hacia abajo (estados 8, 9 y 10) y hacia la derecha (estados 3 y 7) que representan el hecho de salir de la cuadrícula hacia la región aversiva.

Fig. 2b muestra la probabilidad softmax de elegir una acción de cada estado después del aprendizaje. Aunque las probabilidades de elegir una acción están conectadas con los valores Q, en términos de las diferentes rutas posibles hacia el estado objetivo, sólo explican qué tan probable es seleccionar una acción en lugar de cuán exitoso sería el agente al seleccionarla. Por lo tanto, todavía no se puede explicar claramente a un usuario final sin experiencia por qué un agente de RL favorecería una de esas acciones.

Fig. 2c muestra la probabilidad de éxito para cada par estado-acción después del proceso de aprendizaje. Las probabilidades se calculan después de cada episodio utilizando la memoria de transiciones. Como se discutió anteriormente, esto es una manera más transparente de explicar a un usuario final no experto las razones por las cuales un agente favorece acciones específicas de estados específicos. En la 2c, por ejemplo, es claro ver las acciones que conducen a la región aversiva ya que muestran una probabilidad igual a 0. Además, se muestra que incluso las acciones que alejan al agente del estado del objetivo pueden ser exitosas, o que los estados ubicados lejos del objetivo también pueden ser altamente exitosos si se toma la secuencia adecuada de acciones.

Además, la Fig. 2d muestra el número de transiciones necesarias para alcanzar la posición final desde cada estado durante los episodios de aprendizaje. El número de acciones ejecutadas en este caso no es mayor a 5 para todos los estados durante el aprendizaje, que es el camino más corto posible desde el estado más alejado.

En este contexto, una posible pregunta para el agente artificial es: ¿Por qué elegiste la acción desplazarse a la derecha en el estado 0? Intentar explicar esto en términos de valores Q significa mostrarle al usuario final la siguiente información:



(a) Q-values. (b) Softmax values. (c) Probability of success.

(d) Number of transitions.

Figura 2

Resultados obtenidos en la grilla sin límites. (a) Valores Q. Se puede ver que el agente no favorece acciones que lo alejan del estado objetivo, es decir, moverse hacia arriba o hacia la izquierda. Además, los valores Q muestran simetría considerando las posibles rutas hacia el estado objetivo. (b) Probabilidad de elegir una acción. Si bien los valores softmax muestran que el agente puede seleccionar cualquier ruta hacia el estado objetivo con probabilidad similar, no proporcionan suficiente información con lenguaje perteneciente al dominio. (c) Probabilidad de éxito considerando pares estado-acción. Las acciones que conducen a la región aversiva tienen una probabilidad de éxito igual a 0. Además, las acciones más alejadas del estado del objetivo o las acciones que alejan al agente del objetivo también pueden tener éxito si se toma la secuencia correcta a partir de ahí. (d) Evolución del número de transiciones para alcanzar el estado objetivo. Después del entrenamiento, el agente aprende el camino más corto hacia la meta. En este escenario, hay como máximo cinco transiciones necesarias ya que el estado más alejado se encuentra a esa distancia.

- → Q (s = 0, a = abajo) = -0.18061778,
- → Q (s = 0, a = arriba) = -0.99816993,
- → Q (s = 0, a = derecha) = -0.41132827,
- → Q (s = 0, a = izquierda) = -0.99830604,

lo cual no tiene sentido para un usuario no experto. Sin embargo, si usamos la probabilidad de éxito, podemos decir que:

- → Ps(s = 0, a = abajo) = 0.73613932,
- → Ps(s = 0, a = arriba) = 0,
- → Ps(s = 0, a = derecha) = 0.65609104,
- → Ps(s = 0, a = izquierda) = 0.

En otras palabras, le estamos diciendo al usuario final que elegir acciones hacia abajo o hacia la derecha tiene 73.61% y 65.61% de probabilidad de terminar en el estado objetivo, y por lo tanto, aunque elegir abajo tiene una mayor probabilidad de éxito, no hay mucho diferencia entre ellos. Además, Ps muestra claramente que elegir acciones hacia arriba o hacia la izquierda desde el estado 0 conlleva con una probabilidad de éxito igual a cero ya que el agente entra en la región aversiva.

Grilla limitada

Como se mencionó anteriormente, la grilla limitada es un escenario siempre exitoso ya que el agente no puede salir del entorno hacia la región aversiva y, por lo tanto, finalmente siempre alcanzará el estado objetivo. La figura 3 muestra los valores Q obtenidos, la probabilidad de elegir una acción y el número de acciones para el estado final.

En la Fig. 3a, los valores Q obtenidos presentan una distribución similar a la del caso anterior sin límite, es decir, las acciones que desplazan al agente hacia arriba y hacia la izquierda tienen valores más bajos en comparación con abajo y a la derecha que acercan al agente a la posición objetivo.

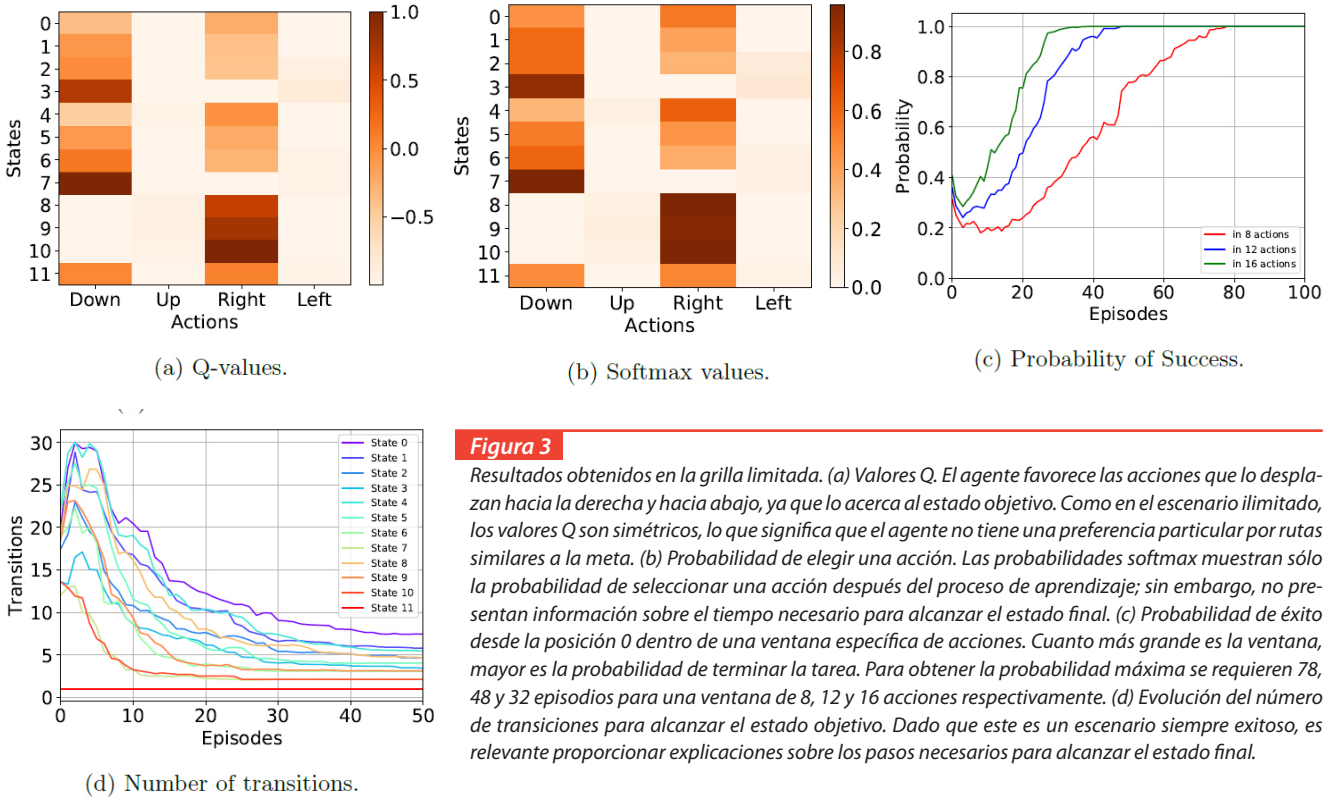
En este caso, la probabilidad de elegir una acción también está relacionada con los valores Q, como se muestra en la Fig. 3b. Sin embargo, esta probabilidad no proporciona suficiente información para comprender y explicar la decisión de selección de acción del agente, especialmente teniendo en cuenta que, en este caso, el agente nunca falla la tarea. Por lo tanto, en este escenario, es importante calcular el número de transiciones necesaria para alcanzar el objetivo y la probabilidad de éxito dentro de una ventana de tiempo. Así, un agente puede responder más claramente a las preguntas de por qué se prefiere una acción particular sobre otras, a partir de un estado específico, refiriéndose a la cantidad de pasos necesarios para alcanzar el estado final.

Fig. 3c muestra la evolución de la probabilidad de éxito durante los episodios de aprendizaje con el agente que comenzando en la posición 0. Se consideran tres ventanas de tiempo diferentes como ejemplos, es decir, la probabilidad de alcanzar la meta en 8, 12 y 16 acciones. En la Fig. 3d se muestra el número de transiciones a partir de cada estado para alcanzar el estado final. El agente puede usar esta información para responder si una acción tomada conlleva a otro estado, desde donde es más rápido alcanzar el estado final.

En este problema, una posible pregunta para el agente podría ser: ¿Cuál es la probabilidad de terminar la tarea en 8 movimientos a partir del estado 0? Una vez más, si queremos responder esta pregunta al usuario final en términos de los valores Q, deberíamos mostrar lo siguiente:

- → Q (s = 0, a = abajo) = -0.36809672,
- → Q (s = 0, a = arriba) = -0.99281156,
- → Q (s = 0, a = derecha) = -0.24326064,
- → Q (s = 0, a = izquierda) = -0.99446782,

que no tiene ningún significado para un usuario final no experto en el área. Sin embargo, si nos referimos a la gráfica 3c, podemos observar claramente la probabilidad de terminar la tarea en 8 movimientos a partir del estado 0.



(a) Q-values.

(b) Softmax values.

(c) Probability of Success.

(d) Number of transitions.

Figura 3

Resultados obtenidos en la grilla limitada. (a) Valores Q. El agente favorece las acciones que lo desplazan hacia la derecha y hacia abajo, ya que lo acerca al estado objetivo. Como en el escenario ilimitado, los valores Q son simétricos, lo que significa que el agente no tiene una preferencia particular por rutas similares a la meta. (b) Probabilidad de elegir una acción. Las probabilidades softmax muestran sólo la probabilidad de seleccionar una acción después del proceso de aprendizaje; sin embargo, no presentan información sobre el tiempo necesario para alcanzar el estado final. (c) Probabilidad de éxito desde la posición 0 dentro de una ventana específica de acciones. Cuanto más grande es la ventana, mayor es la probabilidad de terminar la tarea. Para obtener la probabilidad máxima se requieren 78, 48 y 32 episodios para una ventana de 8, 12 y 16 acciones respectivamente. (d) Evolución del número de transiciones para alcanzar el estado objetivo. Dado que este es un escenario siempre exitoso, es relevante proporcionar explicaciones sobre los pasos necesarios para alcanzar el estado final.

Conclusiones

En este trabajo, hemos presentado un enfoque MXRL con el objetivo de que un agente pueda explicar a usuarios finales no expertos las razones por las que toma sus decisiones en determinadas situaciones. Con este fin, utilizando una memoria de transiciones, hemos calculado la probabilidad de éxito y el número de pasos para hasta estado objetivo, lo que permite proporcionar explicaciones utilizando un lenguaje relacionado al dominio. Nuestros experimentos se han realizado en un escenario con dos variaciones, una grilla ilimitada y otra acotada. Los resultados obtenidos muestran que el agente, utilizando la memoria de transiciones, puede encontrar explicaciones claras para los usuarios finales sin conocimientos previos de técnicas de aprendizaje automático.

Las explicaciones que se muestran en este trabajo son ejemplos de posibles respuestas obtenidas a partir de la probabilidad resultante de éxito y el número de transiciones hacia el estado final. Por supuesto, generar automáticamente una explicación es una alternativa plausible. Sin embargo, esto sigue siendo un problema importante aún no resuelto, ya que todavía no está completamente establecido qué constituye una buena explicación. A menudo, se argumenta que un sistema explicable no debe centrarse en explicar situaciones evidentes. Sin embargo, para diferentes usuarios finales, diferentes situaciones parecen ser evidentes o no, y pueden requerir más explicaciones en algunos casos. Por lo tanto, encontrar una buena métrica para explicar diversas situaciones también es un problema importante sin resolver todavía.

Actualmente, nuestro método presenta algunas limitaciones como el uso de memoria en grandes espacios de solución. Además, hasta este punto en este trabajo, solo hemos considerado una tarea episódica discreta. En este sentido, los resultados obtenidos motivan el trabajo futuro en varias direcciones posibles. Por ejemplo, estamos planeando extender nuestro enfoque para calcular la probabilidad de éxito y el número de transiciones a la meta mediante el uso de otro método más general, por ejemplo, a través de un aproximador de funciones, métodos bayesianos o relaciones fenomenológicas a partir de los valores Q. Mediante el uso de un método de estimación más general, nuestro enfoque podría ampliarse a escenarios más complejos como problemas sin estado final, es decir, que necesitan operar continuamente, o problemas con representaciones de estado-acción continuas. Otra alternativa interesante para una extensión es la verbalización, es decir, que la generación de explicaciones puede ser aprendida por el agente inteligente por medio de interacción con el usuario final. Además, esto permitiría personalizar las explicaciones según las características de cada usuario.

Bibliografía

[1] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. Cambridge, MA, USA: Bradford Book (1998).

[2] Cruz, F., Magg, S., Nagai, Y., Wermter, S.: Improving interactive reinforcement learning: What makes a good teacher? Connection Science 30(3), 306-325 (2018).

[3] Gunning, D.: Explainable artificial intelligence (XAI). • Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[4] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1-38 (2018).

[5] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138-52160 (2018).

[6] Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: Twenty-Ninth IAAI Conference. pp. 4762-4763 (2017).

[7] Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. arXiv preprint arXiv:1905.10958 (2019).