

**A4. RL Vision**

Elías Joaquín Yáñez Huerta A01028482

08 de Noviembre del 2024

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)

TC3007C.501

### Interpretabilidad y Atribución en Modelos de Aprendizaje por Refuerzo

El artículo de Distill examina la interpretabilidad en modelos de aprendizaje por refuerzo (RL) a través de una investigación exhaustiva sobre las señales visuales que guían las decisiones de los agentes en entornos complejos. Este estudio profundiza en cómo un agente de RL percibe y prioriza los elementos visuales críticos de su entorno al tomar decisiones. Para clarificar el funcionamiento interno de la "visión" del agente, los investigadores emplean técnicas de visualización, como mapas de saliencia y manipulación de características. Estas herramientas permiten identificar qué partes del entorno influyen más en las decisiones del agente, revelando sesgos o puntos ciegos en su percepción.

Mediante el uso de mapas de saliencia, los autores muestran cómo ciertas áreas de una imagen captan la atención del modelo en momentos clave. Superponiendo estos mapas a los fotogramas de entrada, se proporciona una representación visual de las zonas importantes, lo que permite a los investigadores evaluar si la atención del agente está alineada con las partes relevantes del entorno, tales como obstáculos o recompensas. Gracias a estas visualizaciones, se puede verificar si el agente sigue patrones lógicos o simplemente responde a pistas visuales superficiales, lo cual podría llevar a un comportamiento inconsistente en distintos contextos.

Técnicas de Interpretabilidad y Atribución en el Aprendizaje Automático

En el ámbito del aprendizaje automático, diversas técnicas de interpretabilidad ayudan a esclarecer el funcionamiento de los modelos. Los mapas de saliencia, por ejemplo, destacan las áreas de las imágenes de entrada que más influyen en las predicciones del modelo, ofreciendo una visión clara de las señales clave para la toma de decisiones. Herramientas como los gradientes integrados y la propagación de relevancia por capas (LRP) permiten atribuir importancia a píxeles o características individuales, ayudando a los investigadores a identificar qué entradas son más determinantes en las decisiones del modelo. Además, el método SHAP (SHapley Additive exPlanations) ofrece una manera de puntuar la importancia de cada característica basándose en principios cooperativos de teoría de juegos, lo que permite una comprensión detallada del comportamiento del modelo.

Aplicadas en el contexto de los modelos de RL, estas técnicas proporcionan una visión más profunda de qué elementos de un entorno complejo el agente considera relevantes o irrelevantes. Esto facilita a los investigadores el ajuste de datos de entrenamiento o la arquitectura del modelo, permitiendo una calibración más precisa de la "atención" del modelo para evitar el sobreajuste a características irrelevantes y asegurar que no se ignoren factores ambientales críticos. La interpretabilidad en los modelos de RL no solo mejora la transparencia, sino que también contribuye a resolver problemas éticos en la toma de decisiones autónomas, especialmente en aplicaciones donde los agentes de RL operan de manera independiente.

El estudio de Distill ofrece un avance significativo en la comprensión de los procesos de toma de decisiones en los modelos de RL a través de técnicas de interpretabilidad y atribución visual. Este conocimiento es fundamental para el desarrollo de modelos de aprendizaje automático que sean no solo más efectivos, sino también más seguros y éticamente responsables en su interacción con el entorno real.