

Para entrenar el manager, se emplea la recompensa obtenida, y emplea la similitud coseno entre la dirección en la que se movió el worker y la compara con el objetivo establecido, empleando la similitud coseno como función de

pérdida. Esta pérdida incentiva al Manager a emitir objetivos que maximicen el progreso hacia estados ventajosos.

El vector de objetivos se envía al worker sin propagar gradientes, esto garantiza que los objetivos mantengan un significado semántico independiente, en lugar de ser simples variables latentes optimizadas de manera conjunta.

En el caso del worker, también se emplea una red LSTM para mantener un estado interno y poder capturar información relevante, pero en este caso, el worker recibe tanto la percepción del entorno como el objetivo del manager. El worker emplea esta información para predecir la acción que debe realizar para alcanzar el objetivo. La acción se predice en el espacio de acciones, y se emplea

Definición

Estado del arte

Evaluación práctica

$Q_{Learningv2}$

Debido a la disipación de las paredes que se experimentaba al realizar las abstracciones del mapa de estados a medida que se disminuía el nivel del gerente, el aprendizaje resultaba altamente perjudicado llegando transmitirse ordenes imposibles de realizar debida a la naturaleza cerrada del mapa del problema.

Para enfrentar dicha adversidad se probó a disminuir el número de gerentes/subgerentes, limitándose al uso del gerente de menor nivel de abstracción; aquel que veía el mapa de estados en la escala original. Por lo tanto, se realizó una disminución de agentes para quedarse con únicamente 2: un gerente encargado de determinar la siguiente zona destino y otro agente encargado en navegar hasta dicha posición de interés. Dicha estructura contaba con las mismas limitaciones y políticas de recompensa definidas para el proceso anterior.

Para facilitar el arranque y el uso del mánager desde el principio, se facilitaba al mánager una pista de la localización de las monedas que debían ser recogidas para que guiara al trabajador (o worker) hacía dichas zonas sin necesidad de un proceso inicial donde las directrices del worker podían ser “ignoradas” debida a su aleatoriedad.

El trabajador a su vez realizaba un proceso de exploración que se veía recompensado cuando alcanzaba el punto objetivo determinado por el mánager. Mediante estas exploraciones, iba creando sus caminos o matrices Q que iba actualizando y afinando con el fin de encontrar el camino óptimo a partir de las recompensas obtenidas por llegar a los destinos asignados por el manager.

En cuanto al proceso de aprendizaje, se incluyeron diferentes implementaciones para ver la influencia que estas variaciones o perturbaciones tenían en el proceso de aprendizaje del super-agente. Las perturbaciones más comunes consistían en limitar el conocimiento que tenía el teacher (dándole solo el conocimiento de la solución empezando desde las coordenadas (1,1) o el conocimiento completo) o variando la posición de comienzo del agente.

En cuanto a los resultados se observó como sin ayudas externas el proceso de aprendizaje era más lento al tener que explorar muchos estados y no contar con una ayuda externa en caso de caer en bucles debido a los estrechos caminos que

componían el mapa. Cuando se le introducía la ayuda externa, teacher, dicho proceso de aprendizaje se veía acortado siendo capaz de definir la ruta optima con mayor rapidez. Sin embargo, la influencia de un teacher con conocimiento de un solo camino ante el que siempre tenía la solución optima suponía una gran diferencia también en los momentos en los que se partía de zonas inexploradas. Esto se debía a que el agente tenía que lograr llegar hasta el punto más cercano de conocimiento del teacher para poder obtener su ayuda, lo que lo ponía en similar situación que cuando no contaba con ayuda externa.

En conclusión, esta nueva estructura feudal, de un noble (mánager) y un subordinado (worker), supuso una mejora en comparación con el nivel anterior debida a la simplicidad y mayor énfasis en las limitaciones del mapa de estados que permitía definir. Mediante los experimentos que se realizaron se observó no solo una mejora en el proceso de aprendizaje si no que una capacidad de llegar a soluciones en mapas de 20x20 tras 9000 iteraciones en caso de no incluir la ayuda del teacher o 1000 iteraciones cuando se le facilitaba un teacher para poder preguntar (limitando el número de consultas a realizar en $3 \times P$; siendo P los pasos del camino más corto para solucionar el problema).

Conclusiones

References

Science, T. D. 2024. Hierarchical Reinforcement Learning: Feudal Networks.