bien cargado el café wow

0:15

quizás recuerden que hace unas semanas subió un vídeo de noticias bueno realmente hace un mes y pico

0:20

subí un vídeo de noticias en el que por primera vez les propuse una cosa y era que aquellos que fueran patrios del

0:25

canal aquellos que están apoyando financieramente este proyecto el dedo csv pues un poco como recompensa a su

0:32

aporte tenían la oportunidad de elegir de todas las noticias que presenté en ese vídeo en el vídeo de noticias de

0:38

septiembre octubre noviembre y que eligieron aquellas que le parecieran más interesante entre todas las noticias

0:43

propuestas fueron ustedes los patrones los que decidieron que la noticia que tenemos que tratar hoy en este data coffee era la de go explorer que es esa

0:51

inteligencia artificial que ha creado über lapse y que ha supuesto un hito en el campo del aprendizaje reforzado al

0:57

conseguir una puntuación récord por encima de lo que se había conseguido hasta el momento en el juego montes o

1:03

más revenge hoy les recomiendo que se preparen un café bastante fuerte porque

1:09

viene un tema muy interesante comenzamos en esa lista de futuros retos se encontraba un juego llamado monte sumas

1:16

revenge uno de los juegos más complicados del atari y en el que sólo un algoritmo había capaz de superar el primer nivel y donde

1:23

nunca nadie antes había superado la puntuación de 17.500 puntos ahora un

1:29

nuevo algoritmo desarrollado por uber lab y denominado go explorer ha conseguido mejorar esto mejorarlo por

1:35

mucho superando el segundo tercer y cuarto a

1:41

nivel y el décimo y el número 58 y de darle la

1:48

vuelta al marcador con más de un millón de puntos y de dársela a una segunda vez

1:53

con más de dos millones de puntos vamos sin duda de dominar completamente este juego esta nueva arquitectura

1:59

desarrollada supone un nuevo hito en el campo del aprendizaje reforzado y promete allanar el camino en la

2:05

resolución de problemas complejos donde la recompensa es escasa y dispersa en el árbol de búsqueda algo que tiene su

2:11

utilidad por ejemplo en su uso para resolver problemas de aprendizaje con robots como vemos hoy vamos a hablar de

2:16

go explorer un algoritmo que da un salto cualitativo a la investigación del campo de aprendizaje reforzado al haber

2:23

superado el récord en este juego lo que vamos a hacer va a ser desgranar el cómo lo ha hecho y también vamos a ver un

2:29

poco de controversia que ha ido desarrollándose durante las últimas semanas y también motivos por los cuales

2:34

me ha costado tanto sacar este vídeo adelante y es que es un tema que ha ido evolucionando semana tras semana con

2:40

diferentes actualizaciones por parte de los autores con parte de quejas de la comunidad científica todo eso lo vamos a

2:46

ver en el vídeo de hoy así que vamos a comenzar el primer punto en el que habría que parar es entender realmente qué es eso de

2:52

aprendizaje reforzado si recuerdan en el canal hace ya un año y pico madre mía

2:57

sobre un vídeo en el que contaba cuáles eran los diferentes paradigmas del aprendizaje dentro del machine learning

3:02

es decir las diferentes maneras que tenemos para clasificar nuestros algoritmos a la hora de de la manera en

3:09

la que aprenden vale estos aprendizajes eran aprendizaje supervisado o aprendizaje no supervisado y aprendizaje

3:16

reforzado estos no son los tres únicos paradigmas pero sí son los tres principales si no recuerdan bien el

3:21

contenido de ese vídeo pues les recomiendo que vayan a verlo hasta por aquí arriba el link pero si no también vamos a hacer un recordatorio rápido el

3:28

aprendizaje supervisado es este aprendizaje donde tú le muestras al algoritmo ejemplos de lo que quieres

3:33

conseguir es decir tú tienes datos de entrada y dos de salida y tú lo que quieres es que el algoritmo aprenda la relación entre el proyecto tanto técnico

3:41

por ejemplo tú puedes tener imágenes de gatos y de perros y puedes tener etiquetas que digan que esta foto de

3:47

aquí son gatos y perros con el paso del tiempo el algoritmo deberá de aprender a asociar estas dos fuentes de datos por

3:53

otro lado la forma antagónica de este tipo de aprendizaje no supervisado que obviamente es aquel

3:59

en el que no le mostramos ningún ejemplo de salida de lo que queremos conseguir esto a priori puede parecer extraño pero

4:05

realmente créeme que se pueden hacer muchísimas cosas solamente con los datos de entrada por ejemplo podrías intentar

4:11

buscar similaridades entre los datos hacer clusterización e intentar comprimir tus datos

4:17

todo esto lo puedes hacer sin necesidad de tener una señal de salida y es de lo que se ocupa el campo o el paradigma del

4:23

aprendizaje no supervisado si se dan cuenta en este canal casi siempre nos hemos centrado en hablar del aprendizaje

4:28

supervisado y en parte en alguna ocasión del aprendizaje en la supervisado

4:33

pero incluso en aquel vídeo donde hablaba de los tres paradigmas nunca llegué a hablar sobre lo que era el

4:38

aprendizaje reforzado con suerte este año podré centrarme más en esto y tratarlo en más profundidad porque es un

4:44

campo bastante interesante y que puede ser la llave del futuro del machine learning en este caso en el aprendizaje

4:50

reforzado el output que queremos que prediga a nuestro algoritmo de machine learning o nuestro agente en este caso

4:56

es una secuencia de acción es una secuencia de acciones que serán ejecutadas dentro de un entorno de

5:02

simulación con el objetivo de que se tenga que realizar una determinada tarea por la

5:07

cual nuestro agente va a ser recompensado en caso de que haga la tarea correctamente o penalizado en caso

5:13

de que la haga mal esto si lo piensas se parece bastante a cómo aprendemos nosotros en la naturaleza cuando eres

5:19

pequeño la forma en la que aprendes a caminar por ejemplo es pues que gateando y cayendo te muchas veces hasta que poco

5:26

a poco vamos entendiendo de alguna manera y nos vamos reconfigurando internamente para poder ir haciendo esa tarea cada vez mejor y

5:32

claro en ese caso la señal de recompensa que serían nuestros padres aplaudiendo y diciéndonos muy bien muy bien harán que

5:38

poco a poco nosotros pues vayamos levantando nos y empezando a dar nuestros primeros pasos esto es muy

5:45

interesante porque lo que estamos hablando aquí es que ya no tenemos que diseñar un experimento donde tenemos que

5:50

seleccionar datos de entrada y datos de salida sino que lo único que hacemos es definir una tarea resolver dentro de un

5:56

framework dentro de un espacio una simulación y nuestra gente va a empezar a desenvolverse a explorar y a explotar

6:04

su conocimiento para intentar cada vez hacer mejor la tarea y creo que queda

6:09

muy claro porque esto es muy interesante dentro del campo del machine learning claro de esta forma de aprendizaje nos plantea una gente

6:15

inteligente un entorno de simulación un objetivo a cumplir recompensas y

6:20

rápidamente la cabeza nos empieza a pensar que estos elementos son los que también nos encontramos en los videojuegos utilizamos estos videojuegos

6:28

como si fueran una especie de reto a resolver de un entorno de simulación donde nuestros agentes inteligentes

6:34

tienen que resolver esa tarea determinada y es por eso por lo que con frecuencia se utilizan estos videojuegos

6:40

para establecer pues nuevos hitos dentro del campo de aprendizaje reforzado que

6:45

son los que llegan luego a la prensa escrita a la prensa mainstream y que todos conocemos por eso vemos que inteligencias artificiales que consiguen

6:51

un rendimiento súper o humano jugando al ajedrez jugando al go o jugando al starcraft 2 y es por eso que en ese

6:58

sentido también los grandes laboratorios de inteligencia artificial como open ya jodi mind han sacado sus propias

7:04

plataformas para que nosotros seamos los que desarrollamos pues nuestros propios algoritmos de aprendizajes reforzados

7:10

sobre una gran variedad de juegos que ellos aportan de todos modos no te creas que todo esto trata sobre resolver juegos lo que

7:17

estamos haciendo aquí es utilizar estos juegos para luego desarrollar algoritmos que podamos traspasar al mundo real a

7:23

resolver problemas más importantes como por ejemplo el campo de la robótica el campo de la robótica es uno de esos

7:29

campos que se benefician bastante de los avances en aprendizaje reforzado ya que lo que queremos es que nuestros robots

7:35

aprendan a resolver una tarea determinada simplemente en base a ensayo y error de una manera adaptativa la

7:41

buena noticia es que durante los últimos años este es un campo que ha ido ganando tracción y momentum y donde se han ido

7:47

viviendo una serie de avances que también han sido muy interesantes al igual que el campo del machine learning ha evolucionado para integrar todo lo

7:53

que está relacionado con redes neuronales que sería el paso del machine learning al deep learning

7:59

el campo de aprendizaje reforzado también ha visto pues esa integración de las redes neuronales dentro de sus

8:04

arquitecturas en lo que se conocería como el deep rainford cement learning o aprendizaje

8:09

reforzado profundo y fue en febrero de 2015 cuando una empresa poco conocida como the mind publicó en una revista

8:16

poco conocida como neytiri un artículo que fue esto dentro del aprendizaje reforzado este

8:22

presentaba una nueva arquitectura un nuevo algoritmo de aprendizaje reforzado que era capaz de aprender a resolver

8:28

múltiples juegos del atari 2600 con un rendimiento superman o quizás muchos

8:34

hayan visto este ejemplo de este juego de aquí donde la gente inteligente la pala en este caso es capaz de aprender

8:39

estrategias avanzadas que incluso nosotros los humanos solemos utilizar como la típica de hacerte el huequito en

8:45

el lateral entre el bloque y la pared para intentar colar la pelota y que vaya rebotando contra el techo y maximizar

8:50

así la recompensa pero como digo este algoritmo no sólo era capaz de resolver este juego de aquí sino que era capaz de

8:55

resolver múltiples juegos de hecho en el artículo original mostraban esta gráfica donde podemos ver como todos estos

9:03

juegos de aquí eran resueltos por este algoritmo con un rendimiento súper mano tenemos el break out que acabamos de ver

9:09

también tenemos el pong tenemos el space invaders todos ellos por encima del rendimiento super humano sin embargo lo

9:15

interesante viene cuando empezamos a bajar por esta lista de juegos y nos vamos fijando en cuáles son aquellos que le cuesta más resolver adecúen al final

9:23

de esta lista hay un juego para el cual este algoritmo ha sido incapaz de ganar un punto porcentual y este juego es

9:29

monte sumas revenge

9:35

podría sentarme aquí a explicar porque creo que este juego es bastante complicado de ser resuelto pero creo que

9:41

la mejor manera de entenderlo es empatizando con estos algoritmos y es por ello que les quiero enseñar cuál fue mi primera experiencia jugando a este

9:48

juego vale vamos a hacer la primera prueba el juego lo pueden encontrar en esta página web donde tienen bastantes

9:54

juegos de la tarea donde pueden probar y vamos a poner esto en pantalla completa

9:59

la calidad es un poco pésima pero bueno no pasa nada vale famoso a montes o más

10:05

revenge creo que me pongo él con las teclas vale perfecto vale bien primera muerte

10:12

vale puedo bajar por la escalera

10:17

vale esto se mueve solo o sea si dejo de pulsar esto se movería solo aquí ya veo

10:23

que bueno hay daño por caída si me cayera aquí seguramente también moriría imagínate que tengo que saltar aquí el

10:29

allianz a vale no se salta con la tecla para arriba buenísima pero vamos de nuevo si yo fuera un algoritmo ahora de

10:35

aprendizaje reforzado pues en base a este ensayo y error pues yo estaría aprendiendo cómo son los controles de

10:41

este juego ahora yo lo tengo un poco fácil porque tengo un cierto conocimiento a priori de cómo funcionan

10:46

los juegos por ejemplo sé que pulsando espacio efectivamente voy a poder saltar entonces vamos a acertar aliana no se

10:53

siente que para se coge solo perfecto puedo subir y bajar por la liana le damos a espacio un nuevo salto para acá

10:59

vale vamos a bajar por la escalera también entiendo que esto es una escalera por mi conocimiento a priori de

11:05

otros juegos con lo cual también sé que puedo moverme de abajo a arriba es otra facilidad que yo tendría valor hay que

11:11

saltar esto sin ok buenísima a mira vale por ejemplo cuando cuando matas a un

11:18

enemigo pues el enemigo desaparece esto es un conocimiento que acabo de adquirir ahora

11:23

a la experiencia vale esa no la quise acceder al se me olvidó que es con

11:29

espacio para saltar quiero a ver si podemos para acá vale esta es una de las puertas entiendo que es abrirá con la

11:35

llave que tengo ahí abajo voy a bajar para acá saltamos la liana bajamos la escalera

11:41

corremos para acá ya no está el enemigo escogemos la llave y aquí por primera

11:48

vez veo que he tenido una recompensa de 100 puntos vale no ha sido este momento hasta que el juego me ha recompensado

11:54

con 100 puntos todo lo que estaba haciendo antes lo estaba haciendo pues un poco por mi conocimiento de cómo

12:00

funcionaría los juegos convencionales pero un algoritmo de aprendizaje cada

12:06

uno quería un aprendizaje reforzado no tendría por qué saber todo este conocimiento avale mirar la sesión vale

12:12

entiendo que estoy perdiendo vidas y que dentro de poco me van a reiniciar el estado vale efectivamente también para

12:17

abrir una puerta pues me dan 300 puntos

12:24

vale esto es una escalera no tengo muy claro hacia dónde tendría

12:29

que moverme y estoy moviéndome por habitaciones esto también me da recompensa parece que es bueno y aquí me

12:35

encuentro con un muro no sé si hay alguna tecla especial que pueda pulsar

12:41

como ven se trata de un juego bastante variado donde tienes que ir haciendo mucho ok

12:46

genial vale creo que acabo de morir y una vez muero que pasa vale se reinicia

12:52

el juego entero y se reinicia toda la sesión con todos los enemigos y todos como pueden ver se trata de un juego muy

12:59

complejo con una gran variedad de dinámicas donde muchas de las decisiones que he tomado creo que han sido en base

13:04

al conocimiento a priori que yo tenía de mi experiencia jugando a videojuegos lo que quiere decir que si tú quieres

13:09

aprender a jugar este juego bien requerirán múltiples sesiones de ensayo y error donde posiblemente mueras

13:15

muchísimas veces si lo pensamos desde un punto de vista del aprendizaje reforzado el motivo por el cual este juego de aquí

13:21

es fácil de resolver y este de aquí no lo es es porque éste tiene una capacidad de recompensar a la gente mucho más

13:28

rápida y directa sobre las acciones que él realiza aquí cada vez que la gente golpea bien la bola y éste rompe un

13:34

bloque y la gente recibe una señal de recompensa que le indica que lo que ha hecho está bien por el contrario en el

13:40

caso de montezuma solamente en la primera pantalla para no cagarla tienes que bajar una escalera no quedarte quieto saltaron a liam a bajar

13:47

a moverte a la izquierda saltar una calavera que se mueve subir una escalera una llave y será en ese momento

13:53

cuando tú recibas una señal de recompensa que te indique que todo lo que has hecho está bien claro para un

13:59

algoritmo que se basa en ensayo y error y tenga que explorar todas las posibles opciones que puedes realizar en ese entorno esto hace el juego bastante

14:06

complicado este problema que nos estamos refiriendo se conoce como hard exploration problema o problema de

14:12

exploración compleja y básicamente se produce porque aquí la recompensa es bastante escasa o está bastante dispersa

14:19

este concepto de la recompensa escasa o dispersa es un concepto que se investiga bastante dentro del área del aprendizaje

14:25

reforzado porque realmente la gran parte de problemas que nos encontramos en el mundo real se basan en esto es decir

14:32

muchas de las decisiones que tomamos en nuestro día a día no tienen como resultado una recompensa inmediata sino

14:37

que tenemos que esperar a que se produzca una concatenación de eventos hasta que esa recompensa nos llegue y

14:43

sepamos si lo que hemos hecho está bien o mal es por eso que este concepto tenga tanta importancia dentro del área del

14:48

aprendizaje reforzado y que este juego se vuelva como un desafío muy interesante a resolver

14:53

y es por eso que ahora que entendemos todo el contexto que envuelve a este juego que se hace interesante

14:59

preguntarse cómo ha sido que google labs con su algoritmo go explore ha resuelto un juego que en 2018 solamente había

15:05

sido superado el primer nivel por un único algoritmo con 17.500 puntos ahora

15:11

gracias a go explorer hemos llegado a alcanzar hasta el nivel 159 y hemos

15:17

alcanzado 2 millones de puntos así que la pregunta es cómo lo han hecho

15:22

[Música] vale

15:27

para entenderla mejor introducida en este trabajo lo primero que tenemos que hacer es volver al concepto de recompensa dispersa que hemos mencionado

15:33

previamente en el campo de aprendizaje reforzado es habitual intentar resolver este problema mediante un mecanismo

15:39

denominado recompensa intrínseca es decir que la gente inteligente no sólo se vea recompensado cada vez que el

15:46

entorno de simulación lo considere porque como hemos visto en muchos casos esta recompensa puede ser muy dispersa

15:52

sino que la gente también pueda experimentar algún tipo de recompensa por el simple hecho de lanzarse a

15:58

explorar o tener curiosidad o encontrar nuevos estados dentro de la simulación imagínate por ejemplo una versión del

16:04

juego de pac man donde pacman sólo consigue recompensa es decir que sólo suban puntos a su marcador cada vez que

16:10

se consiga comer una fruta en este caso un mecanismo de recompensa intrínseca sería establecer en todo el mapa estas

16:17

bolas de puntos que normalmente nos encontramos pacman así se verá motivado para explorar todo el mapa y así ir comiendo

16:24

poco a poco toda la bola de punto y será un tipo de motivación intrínseca que le permitirá moverse y no quedarse quieto

16:30

en su sitio como digo este tipo de mecanismos de motivación intrínseca son utilizados en set apps de aprendizaje

16:36

reforzado y es el punto de partida para los investigadores de uber lapse que se plantea en el siguiente problema

16:42

imagínate que el entorno de simulación se representa por el siguiente mapa donde tu agente se encuentra justo en el medio de estos dos laberintos para

16:48

favorecer la exploración de la gente pues podemos introducir recompensas intrínsecas a lo largo de todas las zonas no exploradas representado aquí en

16:56

verde de manera aleatoria la gente decidirá ir por el laberinto de la izquierda y poco a poco irá explorando y

17:01

recibiendo recompensa por su curiosidad esto será así hasta que llegados a un punto puesto que la gente al final toma

17:06

decisiones de manera aleatoria quizás acabe dándose la vuelta y volviendo al punto inicial y encontrándose con el

17:12

laberinto de la derecha donde podrá continuar su exploración alimentándose de esa señal de motivación intrínseca que hay en el laberinto y aquí es donde

17:19

ellos ven el gran problema y punto de partida de su hipótesis y es que un agente inteligente como tú por ejemplo

17:24

si estuvieras haciendo este tipo de exploración en un juego una vez acabes con esta zona recordarás que antes

17:29

dejaste es una zona potencialmente interesante sin explorar y por tanto volverías a ella sin embargo en este

17:35

caso este agente de aquí ya no tendría una motivación para querer explorar estas zonas puesto que ya no encuentran

17:41

señal de recompensa que le motiva a llegar a ese punto es decir el problema el de hitachi en lo que nos presenta es

17:46

una situación en la que zonas con posibilidad de exploración se vuelven potencialmente poco interesantes para la

17:51

gente inteligente que habrá olvidado que ahí había algo interesante que explorar y al final es la solución a este

17:57

problema la principal mejora introducida por este trabajo como lo hacen pues de

18:02

una manera muy intuitiva y bastante parecida a ese símil que habíamos planteado de un jugador que recuerda que

18:08

ha dejado una zona sin explorar y que tiene que volver a visitarla en este caso lo que hace la gente inteligente es

18:14

mantener un archivo de aquellos estados de la simulación es decir aquellas zonas que ha explorado previamente y que va a

18:20

utilizar para seleccionar aleatoriamente y dando prioridad a aquellas nuevas una nueva zona donde moverse y poder seguir

18:26

explorando y descubriendo cosas pero claro antes cuando hablamos de recordar

18:31

que tenías que volver una zona del juego estamos hablando de la representación mental que tú como humano te haces de

18:37

ese juego es como si te digo que recuerdes cómo era pueblo paleta en el primer pokemon la casa de as que tube tu

18:45

si lo intentará pensar ahora segura sería una representación difusa de la zona pero sabrías a lo que te estás

18:51

refiriendo cómo podemos conseguir representar ese estado dentro de nuestra simulación para que la gente pueda

18:57

entender que tiene que volver a una determinada zona aquí en este caso para hacer esta representación del juego

19:03

muchos algoritmos de aprendizaje reforzado utilizan métodos que pueden ir hasta cosas complejas como utilizar

19:08

redes neuronales convolución ales que aprenden a representar en un estado latente el estado actual del juego a

19:14

partir de los píxeles recibidos por una imagen pero en este caso en el go

19:19

explorer del método que utilizan es bastante rudimentario y bastante simple lo único que están haciendo aquí es

19:25

la imagen que tenemos en pantalla y hacer un downgrade es decir bajarle la calidad de píxeles hasta generar un mapa

19:33

de bits de 11 x 8 píxeles con una escala de grises de 8 bits y será solamente con este mapa de píxeles que han generado

19:40

con el cual van a representar todos los estados diferentes del juego lo cual es bastante sorprendente que funcione pero

19:46

a priori parece que lo hace aunque este es un punto en el que vamos a volver dentro de poco complementario a esto

19:52

también introduce en una segunda denominada fase de robusti ficación que lo que busca es hacer el sistema más robusto frente a posibles perturbaciones

19:59

que se puedan dar en el entorno de simulación de todos modos ya esto no nos vamos a meter a explicarlo porque sino

20:05

el vídeo se va a quedar más largo de lo que está quedando por tanto al final lo que parece que están proponiendo en este trabajo es una especie de exploración

20:12

exhaustiva dentro del árbol de búsqueda en el que estamos dando posibilidad a la gente de saltar de una rama a otra según

20:19

su conveniencia y esto aunque sea muy rudimentario parece que funciona bien y

20:25

parece que consigue un récord y un estado del arte muy por encima de lo que se había conseguido con otras investigaciones

20:32

[Música]

20:40

y es quizás por esto mismo por esta simpleza en las técnicas utilizadas y

20:45

por lo sorprendente de los resultados que mucha gente se ha sentido interesada por este trabajo y le han puesto el ojo

20:52

encima encontrando algunos puntos que generan controversia por ejemplo uno de los puntos que se le ha criticado ha

20:58

sido el sistema que ha utilizado para codificar los estados dentro de la simulación haciendo el down sampling

21:03

porque claro puede que dentro de un juego del atari 2600 este sistema sí te

21:08

sirva es decir reducir todo el contenido de tu pantalla a un mapa de 11 x 8 a lo mejor podría ser factible dentro de

21:15

estos juegos donde en pantalla realmente no hay una gran variación de las intensidades de los píxeles y hay muy

21:21

pocos elementos en pantalla pero el problema es que esto cuando te lo llevas a otros sistemas como por ejemplo una

21:27

simulación en 3d o simulador es más complejos puede que no sea un sistema tan generalizable y esto es algo que se

21:34

le ha criticado y quizás este punto no sea tan problemático porque incluso ellos mismos han reconocido que el

21:39

sistema que utilizan es bastante y que otros más avanzados se podrían utilizar que puede que sean más

21:44

generalizables lo que sí puede ser criticable de este aspecto es que ellos han realizado varios experimentos en

21:51

algunos de ellos no solamente han codificado la imagen que tenían en pantalla sino que también han

21:56

introducido alguna información extra que permitía que el sistema pudiera aprender mejor ejemplos de la información que han

22:03

codificado ha sido por ejemplo la posición en coordenadas xe y del personaje principal o el número de

22:10

llaves que portaba o el número de habitaciones en la que se encontraba o el número del nivel en el que se

22:15

encontraba claro si tú le facilita a la gente esta información esto va a permitir que él pueda tener un

22:21

rendimiento mejor pero qué pasa que esta información solamente es útil para este

22:26

juego y por tanto no va a ser generalizable para otros casos es decir es como que le estamos dando pistas de

22:32

lo que necesitará utilizar para poder aprender a desenvolverse pero no es algo que él haya aprendido solo con lo cual

22:38

cuando nos llegamos a este sistema a otro juego diferente necesitaremos que haya un investigador que introduzca esta

22:43

información para que la gente pueda desarrollarse bien y claro esto tampoco tiene que ser un punto problemático porque al final bueno han

22:50

hecho varios experimentos y efectivamente aún así ha sido con información complementaria y otros en información complementaria como es obvio

22:57

aquellos con información complementaria van a tener un resultado muy muy por encima de los que no tengan información

23:02

complementaria pero claro el problema no es ese el problema es que ellos para publicitar su

23:08

trabajo han utilizado los resultados con información complementaria y eso sea algo que se le puede criticar cuando

23:14

comparamos ambos resultados pasamos de tener una puntuación máxima de 2 millones de puntos que era la que nos

23:19

estamos refiriendo en un principio a tener solamente una puntuación máxima de 35 mil puntos que aún así sigue siendo

23:26

una puntuación por encima del estado del arte que se había conseguido hasta el momento pero claro no son 2 millones de

23:31

puntos la otra cosa que nos queda los investigadores es la forma en la que el sistema hace para volver a aquellos estados en los que la gente ha estado

23:38

previamente si recuerdas antes te explicado que la gente puede volver a un punto previamente visto pero ningún

23:44

momento te explicado cómo lo hace imagínate que yo ahora soy un agente y que me muevo del estado a al estado ve

23:51

ahí si quisiera volver de nuevo al estado a porque quiero retomar la exploración a partir de ese punto pues

23:57

lo puedo hacer de tres maneras diferentes una de ellas sería simplemente memorizar y rehacer aquellos

24:02

movimientos que te han llevado de un estado al otro y claro si esto es una opción pues también otra cosa que se podría hacer es

24:09

directamente resetear todo el entorno de simulación al estado que justamente quieres alcanzar esto lo que te va a

24:15

permitir volver al estado previo que quisieras sin toda la sobrecarga de tener que simular todo el recorrido de

24:21

vuelta esto es como cuando en el pasado guardaban la partida antes de enfrentarte al boss final y cagarla efectivamente lo mismo claro estas dos

24:28

opciones son solamente factibles si te encuentras dentro de un entorno de simulación determinista es decir un

24:33

entorno de simulación donde tú sepas que las acciones que vas a realizar te van a llevar con seguridad al estado

24:39

que tú quieres en este ejemplo si nos imaginamos que en el proceso de movernos de un estado a otro hay algún fenómeno

24:44

aleatorio por ejemplo una compuerta en el suelo que se abre y se cierra de manera aleatoria no podemos asegurarnos que al repetir

24:50

los movimientos hacia atrás éstos nos vayan a llevar al mismo estado aquí la

24:56

solución sería pues entrenar al sistema para que también pueda aprender a volver a los estados que queramos y claro la

25:02

cosa de este asunto es que en este trabajo solamente se han limitado a hacer pruebas en los dos primeros casos optando realmente por la opción de

25:09

resetear el entorno algo que tiene su controversia la crítica esto es que entonces a priori el

25:15

sistema que ellos están presentando sólo funcionaría de momento en entornos de simulación que sean no estocástico es

25:22

decir que sean deterministas y esto pues nos aleja bastante de las situaciones reales que nos encontramos en el día a

25:28

día piensa por ejemplo un robot que tenga que aprender a realizar una tarea en un entorno donde exista simplemente

25:34

viento esta componente ya va a ser una componente estocástica de la cual no tenemos la seguridad que este sistema go

25:40

explore puede aprender a resolver este problema no sólo se limita a que no puede haber una transferencia de

25:46

conocimiento de un entorno determinista a un entorno estocástico sino que también se conoce que cuando tú entrena

25:52

a un agente inteligente en un entorno determinista éste puede desarrollar un comportamiento que exploté en su

25:57

beneficio este determinismo para obtener resultados que realmente son muy buenos pero que en realidad son irreales y es

26:04

por eso que desde 2017 se viene utilizando una serie de técnicas que lo que buscan es introducir esa componente

26:10

estocástico dentro de tu simulación determinista algo que se conoce como sticky actions

26:16

y aquí el nombre es perfecto porque sticky actions acciones pegajosas viene a ser un mecanismo que te introduce una

26:22

pequeña probabilidad aleatoria de que tu agente inteligente vuelva a repetir la última acción que ha hecho haciendo que

26:29

su comportamiento pues ya no sea determinista claro esto se asemeja bastante a cuando tienes un joystick

26:34

antiguo y se te queda atascado en uno de los movimientos en esta situación pues ya te puedes imaginar que un agente

26:40

inteligente que dé un paso hacia adelante tiene una pequeña probabilidad de dar otro paso más con suerte espero

26:46

que no sea hacia un precipicio con esta simple técnica porque tenemos un sistema bastante sencillo que nos permite

26:51

introducir aleatoriedad en el entorno de simulación que hace que el sistema sea no determinista y que por tanto la gente

26:56

no se pueda beneficiar de esto el problema aquí es que ellos no han utilizado este mecanismo y esto hace que

27:02

la comparación que ellos realizan con el resto de algoritmos no sea justa desde un principio a todas estas críticas los

27:10

autores han ido actualizando su propio artículo en las semanas posteriores se han ido presentando resultados pues

27:15

utilizando este que actions y haciendo una especie de reflexión debate sobre el uso determinismo

27:22

en cualquier caso todo esto está basado en un artículo de un blog y no en un paper científico es decir que ésta no ha

27:28

pasado el filtro de la comunidad científica y por tanto hay que mantenerlo un poco en cuarentena yo

27:33

simplemente hoy pues quería comentar todo este sistema porque además de que me lo habéis pedido ustedes los patriots

27:40

y por eso me lo ha querido preparar también creo que era un tema interesante tratar en el canal porque nos introduce

27:45

a conceptos que a lo mejor no son familiares para muchos de nosotros que nos dedicamos a pues al tema del buy in

27:51

learning más tradicional por así decirlo al análisis supervisado y no supervisado

27:56

en ese sentido pues si les ha interesado el tema y no quieren mantenerse en este

28:03

en este algoritmo en el go explore hasta que ellos saquen una publicación final

28:09

pues les invito a consultar el paper publicado en creo que fue julio del año

28:14

pasado por open y hai que sería el estado del arte previo al de go explore y es un artículo que podrán encontrar

28:21

aquí abajo en la descripción que les he puesto todos los links de las cosas que hablando y que es un artículo pues que también

28:27

resuelve el monte sumas revenge de una manera muy interesante eso si quieren mirar hacia atrás si quieren mirar hacia

28:34

adelante les recomiendo que visiten otro link que van a encontrar en la cajita de descripción que es un sistema que ha

28:41

implementado la gente de juniti planteando un reto un chile ntx de un juego que está basado en el monte so más

28:47

revenge pero una versión tridimensional que ahora mismo no recuerdo cómo lo han llamado pero bueno estarán viendo en

28:54

pantalla esas imágenes y pues pues ya está que más quieres ya está el vídeo ya

29:01

se ha acabado ya ya ni más ya creo que hemos sacado todo el aspecto técnico posible ha habido y por haber de este

29:07

artículo ya es que no me queda ni café que es que más quieres qué más quieres

29:12

nada más chico es de me señal positiva de recompensa dándole a like a este

29:17

vídeo si te ha gustado que sepan que pueden apoyar el contenido de este canal posiblemente de los pocos en youtube que

29:23

se metan a hablar de aspectos técnicos de inteligencia artificial es con este nivel de profundidad en español

29:29

y eso pues si lo consideran interesante pues lo pueden apoyar a través de patrón y así podrán también poder votar los

29:36

futuros temas que quieren que traten que trate y yo en el canal así que nada tenéis abajo todos los links a mis redes

29:42

sociales patrio en twitter tweet también muchas muchas muchas cosas más

29:48

suscríbete si te ha gustado y bueno hasta el siguiente vídeo adiós zuma