



viu

**Universidad
Internacional
de Valencia**

Aprendizaje por refuerzo aplicado a tareas de control

Titulación:
Máster de Inteligencia
Artificial

Curso académico
2021-2022

Alumno/a: Werner Seoane
Lucas Ezequiel
D.N.I: 39459365J

Director/a de TFM:
Gabriel Enrique Muñoz

Convocatoria:
Tercera

*The true sign of intelligence is not knowledge but
imagination.*

Albert Einstein

Agradecimientos

A mis padres y a mis hermanos. Sin ellos esto no sería posible.

A Agustina, por ser mi apoyo todos los días.

Índice general

Índice de figuras	III
Índice de tablas	V
Índice de algoritmos	VI
Resumen	1
1. Introducción	3
1.1. Acotación del problema	4
1.2. Dispositivo utilizado	5
1.3. Marco teórico	6
1.3.1. Aprendizaje por refuerzo	6
1.3.2. Algoritmos <i>Policy Gradient</i>	7
1.3.3. Algoritmo Actor-Critic	7
1.3.4. Transformers y Vision Transformers	7
2. Objetivos	9
3. Metodología	11
3.1. Obtención de los datos	11
3.2. Preprocesamiento de los datos	12
3.3. Análisis de la solución e implementación	14
3.3.1. Elección de los algoritmos utilizados por el agente	15
3.3.2. Definición del estado	15
3.3.3. Definición de la recompensa	17
3.3.4. Definición del entorno	18
3.3.5. Definición del agente	19
3.3.6. Entrenamiento	20
3.3.7. Evaluación	21
4. Experimentación	23

4.1. Introducción	23
4.2. Policy Gradient	24
4.2.1. Experimento 1	24
4.2.2. Experimento 2	25
4.2.3. Conclusiones	26
4.3. Actor-Critic	27
4.3.1. Experimento 1	27
4.3.2. Experimento 2	29
4.3.3. Experimento 3	31
4.3.4. Experimento 4 - Vision Transformers	34
4.3.5. Conclusiones	36
5. Resultados	37
6. Conclusiones	40
7. Limitaciones y Perspectivas de Futuro	43
A. Apéndize A	46
Bibliografía	48



Índice de figuras

1.1.	A la izquierda, el pipeline utilizado durante active tracking. A la derecha, el pipeline utilizado en passive tracking.	4
1.2.	DJI Tello. Dispositivo utilizado durante el proyecto	5
1.3.	Componentes del aprendizaje por refuerzo y la interacción entre ellos.	6
1.4.	Imagen tomada durante una de las partidas entre <i>AlphaGO</i> y Lee Sedol.	7
1.5.	Imágenes de las arquitecturas (a) <i>Transformer</i> y (b) <i>Vision Transformer</i>	8
3.1.	Imágenes de ejemplo del conjunto final de datos	11
3.2.	Pipeline del procesamiento del conjunto de datos	12
3.3.	Visualización de la salida de la red YOLO	13
3.4.	Arquitectura de red propuesta por Luo et al. (2019)	16
3.5.	Comparación de recompensa sobre la misma imagen dados dos puntos de vista distinto.	18
3.6.	Interacción agente-entorno.	20
3.7.	Imagen ejecutada en dos experimentos distintos, a la misma distancia del punto objetivo, con diferente número de acciones.	22
4.1.	Experimento Policy Gradient 1 - Recompensa media en el conjunto de entrenamiento	24
4.2.	Experimento Policy Gradient 1 - Testing reward mean	25
4.3.	Experimento Policy Gradient 1 - Duración de los episodios	25
4.4.	Experimento Policy Gradient 2 - Recompensa media en el conjunto de entrenamiento	25
4.5.	Experimento Policy Gradient 2 - Testing reward mean	26
4.6.	Experimento Policy Gradient 2 - Duración de los episodios	26
4.7.	Experimento Actor Critic 1 - Recompensa media en el conjunto de entrenamiento .	27
4.8.	Experimento Actor Critic 1 - Testing reward mean	28
4.9.	Experimento Actor Critic 1 - Duración de los episodios	28
4.10.	Experimento Actor Critic 1 - Train loss	28
4.11.	Experimento Actor Critic 2 - Recompensa media en el conjunto de entrenamiento .	29
4.12.	Experimento Actor Critic 2 - Testing reward mean	29
4.13.	Experimento Actor Critic 2 - Duración de los episodios	30
4.14.	Experimento Actor Critic 2 - Duración de los episodios	30

4.15. Experimento Actor Critic 2 - Duración de los episodios	30
4.16. Experimento Actor Critic 3 - Recompensa media en el conjunto de entrenamiento	32
4.17. Experimento Actor Critic 3 - Función de perdida	32
4.18. Experimento Actor Critic 3 - Número de acciones media por imagen	32
4.19. Experimento Actor Critic 3 - Duración de los episodios	33
4.20. Experimento <i>Vision Transformer</i> - Recompensa media en el conjunto de entrenamiento	35
4.21. Experimento <i>Vision Transformer</i> - Función de perdida en el conjunto de entrenamiento	35
4.22. Experimento <i>Vision Transformer</i> - Testing reward mean	35
4.23. Experimento <i>Vision Transformer</i> - Duración de los episodios	36
5.1. Imágenes obtenidas en los experimentos usando <i>Policy Gradient</i>	37
5.2. Imágenes obtenidas con <i>Actor Critic</i> sin incentivar al agente a moverse	38
5.3. Misma imagen en diferentes etapas del entrenamiento usando <i>Actor Critic</i>	39

Índice de tablas

4.1. Resultados Actor Critic - Acciones escogidas por el agente durante el entrenamiento	33
5.1. Resultados Actor Critic - Media de acciones tomadas por el agente en el conjunto de test	38



Índice de algoritmos

1.	Algoritmo REINFORCE <i>Policy Gradient</i>	15
2.	Algoritmo <i>Actor Critic</i>	16
3.	Algoritmo de recompensa inicial	17
4.	Ejecutar acción en el entorno	19

Resumen

El siguiente trabajo presenta una alternativa mediante el uso de técnicas de aprendizaje por refuerzo al algoritmo de seguimiento de una persona en tiempo real por parte de un dron. A diferencia de estudios ya realizados anteriormente y los cuales iremos detallando a lo largo del siguiente documento, no utilizaremos simuladores para la obtención de los datos, sino que utilizaremos la propia cámara del dispositivo para obtener estos, de manera que el entorno final sea lo más parecido al entorno de entrenamiento.

Construiremos en primer lugar una base teórica, en la que explicaremos brevemente los conceptos que involucra el siguiente trabajo y que da pie a entender las diferentes decisiones tomadas durante su ejecución.

Luego realizaremos un análisis de nuestros objetivos, tomaremos la definición de nuestro problema, y lo estructuraremos para adaptarlo al modelo de problema de aprendizaje por refuerzo. Además iremos comentando los problemas que cada uno de estos pasos conlleva y su evolución y distintas variantes a lo largo de los diferentes experimentos, incluyendo decisiones en el entrenamiento de nuestros agentes, así como la redefinición y adaptación mediante análisis de resultados, pero también mediante prueba y error de cada una de las piezas.

Finalmente haremos un análisis de los algoritmos utilizados y trataremos los diferentes problemas que estos fueron ocasionando durante el transcurso del entrenamiento.

Introducción

1

Cuando pensamos en drones, estamos pensando en dispositivos que podemos pilotar con un *smartphone* y que nos permiten realizar videos y fotos a vista de pájaro. Lo que no siempre se tiene en cuenta es la capacidad de estos dispositivos para poder ejecutar algoritmos complejos, como son por el ejemplo aquellos que permiten al dispositivo hacer el seguimiento de un objeto a través de la cámara a tiempo real. Esto es lo que se conoce como *object tracking* y es el punto fundamental de nuestro trabajo.

Recientes algoritmos y estudios alrededor del *object tracking*, entre ellos [Zhao et al. \(2021\)](#) o [Zhou et al. \(2021\)](#), se centran fundamentalmente en el reconocimiento del objeto y en la posición de éste con respecto a las coordenadas X e Y en relación al centro de la imagen a través de la cámara incorporada para guiar las acciones del dispositivo. Es lo que se conoce como *passive tracking*. Este tipo de algoritmos han ganado más atención debido a la simplicidad del problema y los avances en reconocimiento de objetos. Lo cual los hace idóneos para una aplicación industrial.

Otros intentos de realizar *object tracking* con aprendizaje supervisado se realizan utilizando modelos en simuladores tanto del dron como de las personas. Sin embargo, la idea del presente trabajo es la aplicación de diferentes técnicas de inteligencia artificial para lograr un agente que sea capaz de realizar el seguimiento utilizando simplemente imágenes que provengan del propio dispositivo en el cual se desplegará el agente.

Sin embargo, el objetivo en nuestro caso es poder realizar el seguimiento a tiempo real de una persona a través de la cámara del dispositivo utilizando para ello visión por computador y técnicas de aprendizaje por refuerzo para controlar las acciones a tomar en cada momento. Lo que se conoce como *active tracking*.

Para ello, no solo utilizaremos técnicas de aprendizaje por refuerzo, sino también técnicas de aprendizaje supervisado sobre imágenes para la obtención de nuestro conjunto de datos inicial, tal y como explicaremos en las siguientes secciones.

El desarrollo del proyecto se llevará a cabo utilizando lenguaje Python y la librería [PyTorch](#) para el desarrollo de los modelos. El código completo estará disponible para su visualización en un repositorio de [GitHub](#), a excepción de los pesos de los diferentes agentes que se vayan guardando durante el entrenamiento debido al tamaño individual de cada uno de estos archivos.

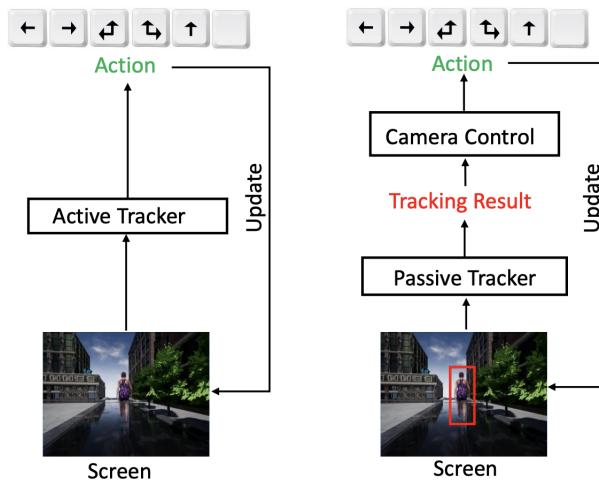


Figura 1.1: A la izquierda, el pipeline utilizado durante active tracking. A la derecha, el pipeline utilizado en passive tracking. Imagen tomada de [Luo et al. \(2019\)](#).

1.1. Acotación del problema

Debido a que nos movemos en un entorno real, en el cual las posibilidades tanto de actuación sobre el entorno como de error son muy amplias, no solo debido a la aleatoriedad del mismo sino también al número de acciones que podemos tomar, debemos definir ciertas restricciones.

En primer lugar, nos centraremos en obtener una solución que sea viable en entornos en los cuales las condiciones externas no sean un impedimento para el buen funcionamiento de nuestro algoritmo. Esto quiere decir que durante el desarrollo y testeо de los algoritmos obviamos factores tales como el viento, las condiciones de humedad o el grado de iluminación en un momento dado, que pudiesen afectar al rendimiento del dispositivo.

Por otro lado, debido a que el espacio de acciones disponibles es muy alto, en una primera fase dispondremos de tan solo 3:

- Girar a la derecha.
- Girar a la izquierda.
- Mantenerse en el mismo lugar.

Tanto el giro a la derecha como el giro a la izquierda se realizará en el dispositivo de mane-
ra controlada, esto quiere decir que nos moveremos siempre utilizando el mismo ángulo de giro.
Si quisieramos añadir además un ángulo de giro variable junto con la acción a tomar podríamos
hacerlo como parte de la salida de nuestro algoritmo, aunque involucraría un entrenamiento más
prolongado en el tiempo y se escaparía a las restricciones de tiempo de este proyecto.

De tal forma que nuestra solución será valorada positiva o negativamente con respecto al eje X exclusivamente. Esto quiere decir que consideraremos que el algoritmo funciona de forma correcta

si es capaz de moverse de tal manera que la persona se encuentre siempre centrada en el eje horizontal y no en el eje vertical.

En cuanto a la detección de la persona, tenemos en cuenta que nuestro dispositivo solo realizará el seguimiento de una sola persona, debido a la complejidad que esto supondría en cuanto al funcionamiento de nuestro algoritmo. Obviamos un escenario real y plausible que es el de encontrarnos en una misma imagen con múltiples personas, en cuyo caso deberíamos también desarrollar un mecanismo de selección de la persona a la cual quisiéramos seguir.

Por lo tanto, quedándonos con una parte simplificada del problema, podemos centrarnos en conseguir un algoritmo que pueda considerarse como el mínimo viable para conseguir nuestro objetivo.

1.2. Dispositivo utilizado

Para el desarrollo del proyecto se utilizará el dispositivo [DJI Tello](#), ya que nos proporciona una interfaz de programación compatible con nuestras necesidades: control del dispositivo y transmisión de datos a través de una API en lenguaje Python.



Figura 1.2: Imagen del dispositivo utilizado durante el proyecto.

El dispositivo cuenta con una cámara integrada con una resolución máxima de 1280x720 píxeles, la cual creemos que es suficiente para el desarrollo del trabajo.

Un punto negativo del uso de drones como dispositivo de captura es el poco rendimiento de las baterías durante el vuelo. Concretamente, el dron utilizado permanece en vuelo unos 13 minutos como máximo.

Además, durante el desarrollo inicial del proyecto se pudo observar una latencia reseñable al intentar comunicar el dispositivo con el ordenador a la hora de transmitir las imágenes y de poder

enviar señales de control debido a la débil conexión WiFi entre ambos puntos de comunicación. El intentar solucionar este problema no solo consumió tiempo de ejecución del proyecto sino que también nos impide poder llevar a cabo pruebas de control más realistas y nos acotará el margen de ejecución de nuestras pruebas finales.

1.3. Marco teórico

En esta sección haremos una breve introducción a los diferentes componentes teóricos que nos iremos encontrando a lo largo del trabajo. Empezaremos repasando los principales conceptos relacionados con el aprendizaje por refuerzo y en qué se diferencia de otros tipos de entrenamiento.

Después haremos una introducción a cada uno de los algoritmos que vamos a utilizar y comentaremos brevemente las diferencias y el por qué los elegimos. Por último comentaremos los *Transformers*([Vaswani et al., 2017](#)) y en especial los *Vision Transformers*([Dosovitskiy et al., 2020](#)), ya que serán utilizados durante uno de los experimentos que explicaremos en la sección [4](#).

1.3.1. Aprendizaje por refuerzo

"La idea de que nosotros aprendemos a través de la interacción con nuestro entorno es probablemente lo primero que se nos ocurre cuando pensamos en la naturaleza del aprendizaje", Sutton y Barto ([Sutton y Barto, 2018](#)).

El aprendizaje por refuerzo es la formalización de esta idea, en concreto, es la rama de la inteligencia artificial cuyo principio se basa en la interacción de un agente con un entorno a través de lo que llamamos acciones. Dichas acciones obtienen una recompensa que el agente recibe y en base a estas recompensas, se refuerzan las acciones positivas y se penalizan las negativas, es decir, el objetivo del agente es maximizar las recompensas obtenidas a lo largo del tiempo.

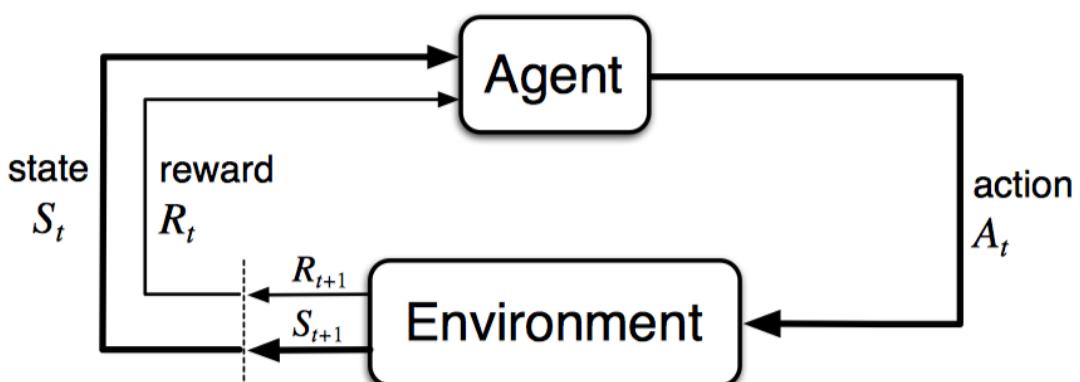


Figura 1.3: Componentes del aprendizaje por refuerzo y la interacción entre ellos.

Durante los últimos años se pudo ver cómo esta rama fue cobrando más importancia por casos exitosos como el de *AlphaGo* ([Silver et al., 2016](#)), creado por la empresa [DeepMind](#), que derrotó

al campeón mundial en el juego Go, un hito que era considerado como extraordinario debido al número de combinaciones por movimiento que el juego contiene.

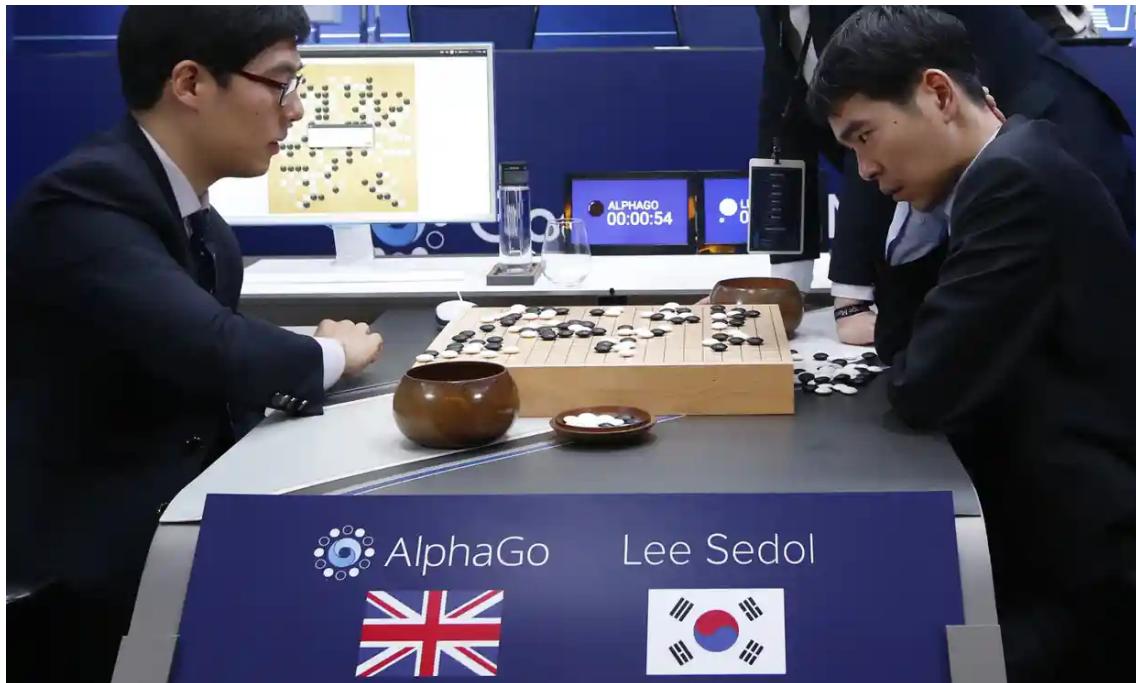


Figura 1.4: Imagen tomada durante una de las partidas entre *AlphaGo* y Lee Sedol. Obtenida de [The Guardian](#).

Y aunque se suele asociar al aprendizaje por refuerzo con aplicaciones como *AlphaGo*, o el aún más reciente caso de *AlphaStar* (Vinyals et al., 2019), gracias a los avances tanto en algoritmos como en el campo del *deep learning*, vemos que con más frecuencia nos encontramos con ejemplos reales en el uso de estas técnicas, lo cual nos anima a realizar nuestro trabajo sobre ellas.

1.3.2. Algoritmos *Policy Gradient*

El subgrupo de algoritmos *Policy Gradient* (Sutton et al., 1999) se basan en la optimización de la política de elección de acciones del agente, es decir, producen políticas de acción estocásticas. Esto contrasta con el subgrupo de algoritmos *Q-Learning* (Mnih et al., 2013) que se basan en la optimización de la función de valor y que producen políticas deterministas.

Una ventaja de estos algoritmos es que por lo general son más estables y confiables que los basados en *Q-Learning*.

1.3.3. Algoritmo Actor-Critic

1.3.4. Transformers y Vision Transformers

Sin entrar en muchos detalles en cuanto a cómo funcionan los *Transformers* y los *Vision Transformers*, haremos un breve repaso de las características únicas de estas redes, ya que uno de

los experimentos nombrados en la sección 4.3.4 será el de un agente con una red de tipo *Vision Transformer* (Dosovitskiy et al., 2020), que se encargará de realizar la transformación de la imagen de entrada.

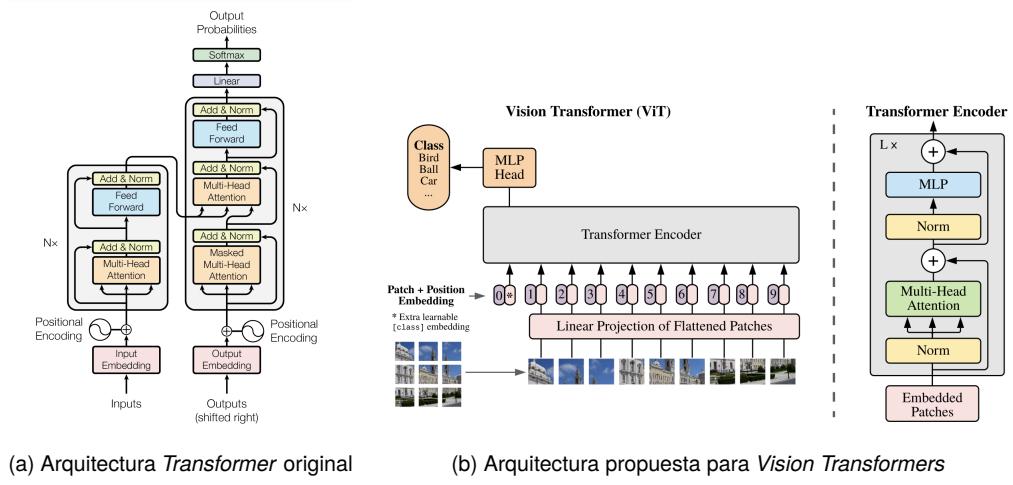


Figura 1.5: Imágenes de las arquitecturas (a) *Transformer* y (b) *Vision Transformer*

Las redes basadas en *Attention*, en particular las redes *Transformers*, desarrolladas inicialmente por Google, se han convertido en la elección por defecto en problemas de lenguaje natural o NLP, ya que gracias a su eficiencia computacional y su escalabilidad, es posible entrenar modelos de más de 100 billones de parámetros. La idea detrás de estas redes es solucionar los problemas que presentan las redes neuronales recurrentes o RNN, tales como las redes LSTM o GRU, las cuales no son capaces de captar dependencias a largo plazo, tal y como podemos encontrarnos en un texto de una novela.

Aunque los *Transformers* son ampliamente utilizados en procesamiento de lenguaje natural, las arquitecturas de redes convolucionales o seguían siendo el tipo de red dominante en el campo de la visión por computador. Es por ello que, inspirados en la arquitectura original de *Transformer* y en el éxito que estas contaban (y aún cuentan), se buscó aplicar los mismos principios en la visión por computador. Esto es lo que dio lugar a los *Vision Transformer*, cuya idea principal es la de partir una imagen en segmentos o *patches* y ser capaces de aplicar capas de *Attention* a cada uno de ellos.

A día de hoy nos encontramos muchas versiones disponibles y preentrenadas. Durante la ejecución de nuestro experimento utilizaremos un modelo disponible en la web [HuggingFace](#). En concreto utilizamos un modelo entrenado por Google, cuya entrada es una imagen de 224x224 píxeles, que será luego dividida en 16 parches de 14x14 píxeles y que fue entrenada en el conjunto de datos *ImageNet21k* (Ridnik et al., 2021). La salida de este modelo, es decir, la entrada a nuestro agente será un vector de tamaño 197x512, donde la primera dimensión representa el tamaño de la secuencia en la cual nuestra imagen ha sido partida tras el paso por el *Vision Transformer*, y cada elemento de esa secuencia consta de un total de 512 características.

Objetivos

2

Al principio del trabajo nos planteamos resolver los siguientes objetivos, algunos de los cuales no pudieron ser resueltos tal y como describiremos en el presente documento:

- 1. Ser capaces de descomponer nuestro problema en subproblemas de menor complejidad.**
- 2. Adaptar esos subproblemas al marco de la inteligencia artificial y en especial al del aprendizaje por refuerzo.**
- 3. Analizar y estudiar soluciones ya implementadas relacionados con nuestro caso.**
- 4. Implementar una solución inicial que sirva como *baseline*.**
- 5. Iterar sobre la solución inicial para conseguir un mejor resultado que esta usando para ello técnicas de aprendizaje por refuerzo más complejas.**
- 6. Conseguir una solución que pueda ser desplegada en nuestro dispositivo.**

Metodología

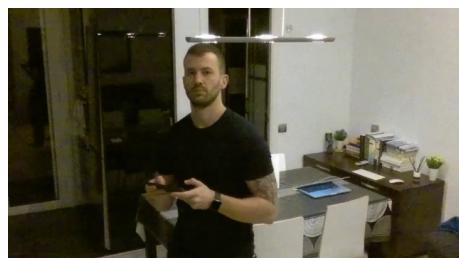
3

En este capítulo entraremos más en profundidad en los detalles técnicos y en las diferentes decisiones tomadas a lo largo del proyecto en cuanto a la descomposición del problema para adecuarlo al paradigma del aprendizaje por refuerzo.

Comenzaremos hablando de los datos, cómo se obtuvieron y cómo los utilizaremos en nuestro proyecto y hablaremos de las diferentes piezas de este y los algoritmos escogidos para llevar a cabo los experimentos.

3.1. Obtención de los datos

La toma de datos se hace mediante el análisis de archivos de vídeo tomados desde el propio dispositivo en un entorno controlado. Es importante destacar que la calidad de las imágenes no es alta, sobre todo en entornos en los cuales la iluminación no es favorable, por ejemplo en entornos con iluminación artificial.



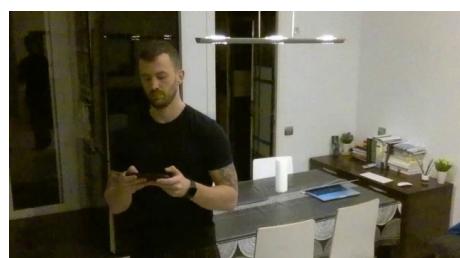
(a)



(b)



(c)



(d)

Figura 3.1: Imágenes de ejemplo del conjunto final de datos.

El dispositivo nos proporciona una aplicación móvil para controlar la grabación y el almacenamiento de los videos que luego serán volcados en el ordenador para su posterior procesamiento, del cual hablaremos posteriormente.

Se ha decidido también utilizar sólo imágenes procedentes del dispositivo y no incluir imágenes de terceras fuentes que podrían haber ayudado a mejorar el tamaño final de nuestro conjunto de datos ya que la idea es acercarse lo máximo posible a un entorno final real.

Debido a las leyes vigentes sobre vuelo de drones, los escenarios en los cuales fueron grabados los videos se redujeron drásticamente a un entorno cerrado, lo cual nos lleva a pensar que más adelante podríamos tener un problema de overfitting en nuestro modelo.

En un escenario ideal contaría con diferentes entornos que nos permitiesen descartar la posibilidad de que el agente aprenda que la posición de determinados objetos es condicionante para tomar una decisión y que se centrara solo en las persona de la imagen.

3.2. Preprocesamiento de los datos

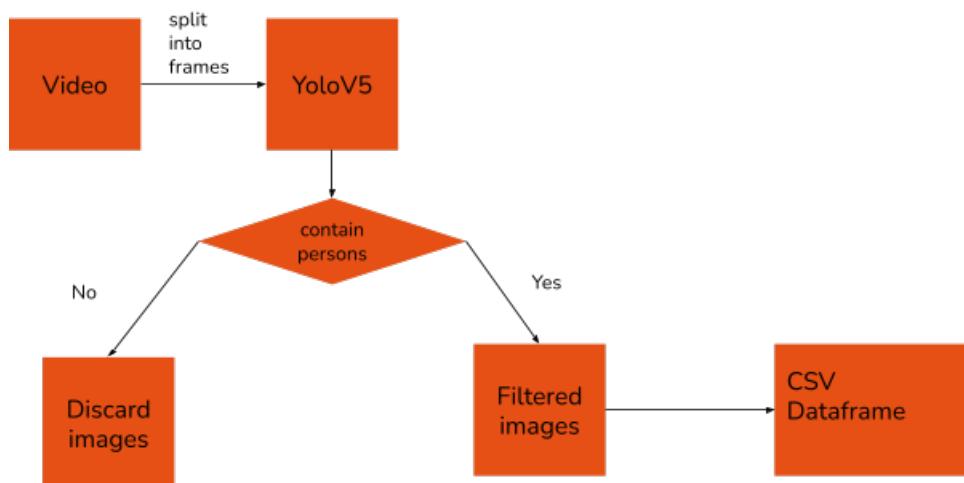


Figura 3.2: Pipeline del procesamiento del conjunto de datos.

Tras la recolección de videos y fotos desde el dispositivo, se realiza un primer proceso de filtrado. Este primer paso consiste en realizar un split de los archivos de video en frames individuales y realizar una operación de reducción del tamaño sobre los frames originales, pasando de 1280 píxeles de ancho y 720 píxeles de alto, a 640 píxeles de ancho y 360 píxeles de alto. Más adelante se tendrá que realizar una segunda reducción de las imágenes para adaptarlas al tamaño de entrada de la red convolucional previa al agente.

Este primer paso de filtrado es posible gracias a los algoritmos de detección de objetos, mediante el uso de redes convolucionales neuronales, que nos proporcionan las coordenadas de las bounding box de aquellas clases de objetos que queremos encontrar en nuestro conjunto de datos.

Durante el transcurso de nuestro proyecto utilizaremos *YOLOv5*([\(Redmon et al., 2016\)](#)), cuya implementación se encuentra disponible de forma gratuita como librería open source. Dicha implementación nos proporciona además múltiples configuraciones de la red. En nuestro caso, esto no fue muy relevante dado que solo lo utilizaríamos para una primera fase de preprocesamiento de las imágenes.

Debido a que la red *YOLOv5* fue entrenada sobre el conjunto de datos *ImageNet* ([\(Deng et al., 2009\)](#)) para detectar 1000 clases de objetos diferentes, debemos modificarla para que sea capaz de detectar tan solo una persona en las imágenes que le pasamos, si es que la hay. La documentación de la librería nos ayuda a conseguir esto, de manera que tras una serie de pruebas tanto de diferentes arquitecturas de *YOLO* como de configuración de los parámetros, en concreto: el intervalo de confianza y el IoU threshold, la red nos devuelve la ubicación de la bounding box en donde se encuentra la persona si es que la hubiese.

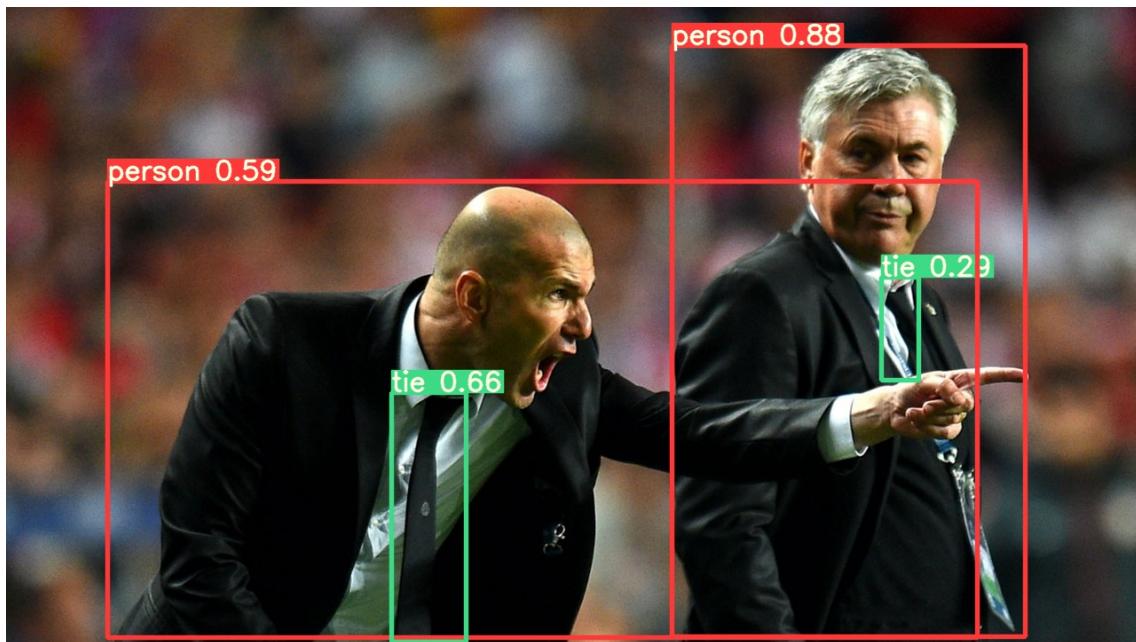


Figura 3.3: Salida de la red *YOLOv5* en donde se pueden observar las *bounding boxes* de los objetos reconocidos en esta. Obtenida de [Yolov5 Tutorial](#).

Sin embargo, para poder adaptar la salida de la red a nuestras necesidades, lo que hicimos fue calcular el punto central de la bounding box en coordenadas X e Y, para que nuestro agente intente acercarse a ese punto durante el entrenamiento en el menor número de pasos posible. El resultado fue por un lado, un dataframe de Pandas que contenía la ruta de la imagen que *YOLO* había filtrado y las coordenadas X e Y del punto central de la bounding box.

Un punto a destacar es que las imágenes que no contenían personas fueron descartadas del conjunto de datos. Esta fue una decisión que se valoró al inicio con el fin de conseguir un agente y una definición más sencilla del problema. De no haber sido así, tendríamos que haber tenido en cuenta el hecho de que la imagen no presenta un punto objetivo y por lo tanto el dron debería permanecer quieto o girar hasta encontrar una persona.

El número total de imágenes de nuestro conjunto de datos inicial tras este primer paso se reduce a tan solo 1690 imágenes. Sin embargo, y debido a que nuestro objetivo es modelar los movimientos de giro del dron tenemos que ser conscientes de cuántas imágenes contamos para que el dron nos detecte a la derecha, a la izquierda o en el centro de la imagen.

En un primer análisis, nuestro conjunto de datos se encontraba completamente desbalanceado. Lo que hicimos para comprobar esto fue tomar las coordenadas centrales de la bounding box sobre el eje X que calculamos con *YOLO* y comprobar en qué parte de la imagen se encontraba. Si esta coordenada se encontraba entre los píxeles 0 y 280, entonces el dron tendría que girar a la izquierda, si se encontraba entre el 280 y el 360, entonces podríamos decir que el dron se mantendría prácticamente quieto, y si la coordenada se encontraba por encima de 360, el dron tendría que girar hacia la derecha.

Los resultados de este análisis fueron los siguientes:

- 340 imágenes se encontraban con la coordenada X en el lado izquierdo.
- 870 imágenes se encontraban con la coordenada X en el lado derecho.
- 480 imágenes se encontraban con la coordenada X en el centro de la imagen.

Debido a este desbalance, se tomó la decisión de igualar la cantidad de imágenes en cada lado, dejando un total de 340 imágenes por cada categoría, lo que hizo un total de 1020 imágenes finales para el entrenamiento y la validación de nuestro agente, lo que en contrapartida podría incrementar aún más el overfitting, pero nos aseguramos de que las decisiones que toma nuestro agente no se verán condicionadas por el número de acciones totales que debería tomar sobre un lado u otro.

Sin embargo, antes de tomar la decisión de descartar las imágenes se estudió la posibilidad también de realizar técnicas de data augmentation tal y como se recomienda en la literatura, utilizando técnicas de crop y volteo horizontal y vertical en cada una de las imágenes, pero la complejidad añadida de tener que recalcular el nuevo punto central hizo que descartemos esa posibilidad.

Debido a que nuestro conjunto de datos no es muy grande, decidimos que el conjunto de entrenamiento sea el 90 %, del cual el 80 % se usará para la fase de entrenamiento de nuestro modelo y el otro 20 % se usará para la fase de test. El 10 % restante de nuestro conjunto, aproximadamente 100 imágenes, lo reservamos para la etapa de validación, de manera que probaremos los resultados de nuestro agente sobre un conjunto de imágenes que no haya visto previamente.

3.3. Análisis de la solución e implementación

En la siguiente sección hablaremos de la implementación de cada uno de los elementos que componen nuestro problema, así como de las decisiones iniciales tomadas.

3.3.1. Elección de los algoritmos utilizados por el agente

Para entrenar nuestro agente utilizaremos métodos que optimicen la *policy* del agente, es decir, que optimicen la toma decisiones para maximizar la *reward* del entorno.

Comenzaremos utilizando el algoritmo *REINFORCE Policy Gradient*, el cual se considera de los más sencillos dentro de la familia de *Policy Gradient* y nos servirá como punto de partida para analizar los posibles problemas y limitaciones de nuestro entorno y nuestras definiciones. Con este algoritmo, el agente colecciona ejemplos del episodio utilizando para ello la *policy* actual.

Algoritmo 1: Algoritmo REINFORCE *Policy Gradient*

$$\begin{aligned} Q^{\pi_\theta}(s_t, a_t) &= v_t \\ \Delta\theta_t &= \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t \end{aligned}$$

```
function REINFORCE
    Initialise  $\theta$  arbitrarily
    for each episode  $\{s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$  do
        for  $t = 1$  to  $T - 1$  do
             $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t$ 
        end for
    end for
    return  $\theta$ 
end function
```

Como siguiente paso, escogeremos nuevamente un algoritmo de la familia de *policy gradient* para el entrenamiento del agente. En este caso, el algoritmo *Actor-Critic* ([Haarnoja et al., 2018](#)). La mayor diferencia entre este algoritmo y el anterior, es que en este caso contamos con 2 componentes separados: el *Actor*, que será el responsable de devolver la distribución de probabilidades de cada una de las acciones del agente y el *Critic*, que será el encargado de estimar el valor del estado en el que se encuentra dicho agente.

3.3.2. Definición del estado

El estado es la representación del problema que el agente debe resolver. En este caso, el estado será una imagen de la cámara del dron, a la que se le aplicó previamente un procesamiento, tal y como comentamos en la sección [3.2](#), junto con el punto en el que se encuentra el agente, expresado en un número entre 0 y 1, representando el 0 estar en el borde izquierdo de la imagen y 1 estar en el borde derecho de esta.

Es decir, el agente recibirá un vector como entrada, de tamaño 4097. Los primeros 4096 elementos representan la imagen, la cual ha sido procesada por una CNN preentrenada, que en nues-

Algoritmo 2: Algoritmo Actor Critic

Algorithm 1 Monte Carlo on policy actor-critic.

Require: Initialize policy π with parameters θ_π and value critic v_π with parameters θ_v

```

1: for each episode do
2:   Get initial state  $s$ 
3:   Initialize storage buffer  $S, A, R, S'$ 
4:   for  $i = 1, 2, 3 \dots N$  steps do
5:     Sample action with policy:  $a \sim \pi_\theta(s)$ 
6:     Run action through environment, obtain reward and post state:  $r, s' \leftarrow ENV(s, a)$ 
7:     Collect and store:  $S, A, R, S' \leftarrow s, a, r, s'$ 
8:      $s \leftarrow s'$ 
9:   end for
10:  Compute discount returns:  $\hat{V} = \sum_{l=0}^{N-1} \gamma^l r_{t+l}$ 
11:  Update  $\theta_v$  to minimize  $\sum_{n=1}^N \|v_\pi(s_n) - \hat{V}_n\|^2$ 
12:  With learning rate  $\alpha$ , update policy:  $\theta_\pi \leftarrow \theta_\pi + \alpha \nabla_\theta \log \pi(A|S)v_\pi(S)$ 
13: end for
```

tro caso ha sido VGG19 ([Simonyan y Zisserman, 2014](#)), la cual recibe como entrada un vector de tamaño 224x224x3, es decir, una imagen en formato RGB de 224x224 píxeles. Si bien la elección de nuestra CNN no fue basándose en ningún criterio en particular, la idea era disminuir el tamaño de la entrada de nuestro agente.

También se tuvieron en cuenta diferentes soluciones para la definición del estado. Entre ellas, la posibilidad de que el estado se formase de varias imágenes al mismo tiempo, debido el componente temporal con el que cuenta nuestro problema. Este estado condicionaría también la arquitectura de la red de nuestro agente, dado que en este caso necesitaríamos tratar con redes recurrentes o en su defecto con redes de tipo Transformers ([Vaswani et al., 2017](#)), tal y como se explica en el trabajo realizado por [Luo et al. \(2019\)](#) y podemos ver en la figura 3.4, donde podemos ver que después del *encoder*, tal y como se cita en el trabajo, se utiliza una red LSTM.

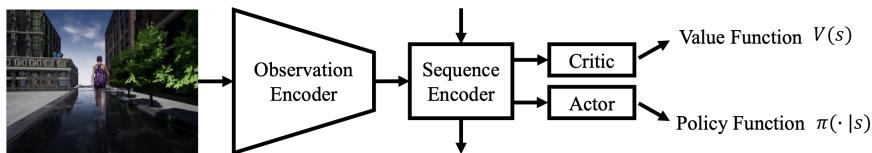


Figura 3.4: Arquitectura de red propuesta por [Luo et al. \(2019\)](#).

Para aumentar la velocidad de nuestro entrenamiento, la imagen será preprocesada por la red convolucional al inicio de cada episodio y se mantendrá constante, dado que lo que iremos cambiando con cada acción durante el entrenamiento es el punto en el que se encuentra el agente.

Por último, el estado contiene en su último elemento, el punto en el que se encuentra el agente en un momento dado, de manera que este debería ser capaz de tomar decisiones basándose no solo en la representación de la imagen sino también del valor que este valor tenga en ese instante.

3.3.3. Definición de la recompensa

El desarrollo de la recompensa fue uno de los puntos críticos en cuanto a la implementación y la definición de nuestro problema. Por un lado, tenemos que intentar que nuestro agente sea capaz de ir aprendiendo los pasos intermedios que lleven a la obtención de la recompensa óptima, en el menor número de pasos posibles y por otro, definir cuánta recompensa en términos absolutos el agente recibirá cuando acaba correctamente el episodio, en contraposición a cuando llega al número máximo de intentos sin haber obtenido la recompensa máxima.

La idea inicial fue crear una zona de recompensa proporcional a la distancia del punto en el que se encontraba el agente y el punto obtenido durante el preprocesamiento de los datos. Esto lo hicimos dividiendo la imagen en secciones longitudinales de tamaño fijo. Dado que nuestra imagen era de 640 píxeles de ancho, lo que hicimos fue dividirla en 32 secciones de 20 píxeles cada una. La recompensa final obtenida sería inversamente proporcional a la distancia, medida en número de secciones, en que el agente se encontraba con respecto al punto objetivo. De manera que si el agente se encontraba a 5 secciones de nuestro punto final, obtendría menos recompensa que si se encontrara a 2 secciones, tal y como se puede observar en el algoritmo 3:

Algoritmo 3: Algoritmo de recompensa inicial

1. Dividimos la imagen en secciones longitudinales de tamaño fijo $Secciones_{total}$.
 2. Establecemos una recompensa máxima del entorno $Recompensa_{max}$.
 3. Calculamos la sección en la que se encuentra nuestro agente:
$$Seccion_{agente} = round(P_{agente} \text{pixels} / Ancho_{imagen}).$$
 4. Calculamos la sección en la que se encuentra nuestro punto final:
$$Seccion_{objetivo} = round(P_{objetivo} \text{pixels} / Ancho_{imagen}).$$
 5. Si $Seccion_{agente} = Seccion_{objetivo}$ devolvemos $Recompensa_{max}$.
Sino devolvemos $1 - (Seccion_{agente} - Seccion_{objetivo}) / Secciones_{total}$
-

Lo que conseguimos con esta recompensa fue obtener un *heatmap* alrededor del punto objetivo, de manera que el agente podría interpretar si se estaba acercando o alejando.

La recompensa máxima fue establecida a 1, de manera que los saltos entre las recompensas parciales y la recompensa final fuese siempre proporcional a la distancia. Más adelante también comprobaremos cómo este valor puede afectar al desarrollo de una solución por parte de nuestro agente.



Figura 3.5: La misma imagen, con 2 puntos definidos por el agente a diferente distancia del punto objetivo. La imagen superior obtendría la puntuación máxima mientras que la imagen inferior solo obtendría una recompensa relativa a la distancia del punto objetivo.

En la figura 3.5 podemos observar cómo funciona nuestro sistema de recompensa inicial. En la imagen situada en la parte superior, el punto amarillo, que representa el punto en el que se encuentra el agente, está dentro de la misma sección que el punto objetivo, y por lo tanto la recompensa será máxima. Mientras que en el caso de la imagen inferior, el punto donde se encuentra el agente está a una distancia mayor a una sección (definida por defecto en 20 píxeles), y por lo tanto la recompensa será proporcional al número de secciones que se encuentra hasta llegar al punto objetivo.

3.3.4. Definición del entorno

Para desarrollar el entorno, lo que se hizo fue implementarlo siguiendo como ejemplo los entornos propuestos por OpenAI Gym ([Brockman et al., 2016](#)). Esto nos facilitó tener una primera estructura y una primera definición de los métodos que deberíamos implementar.

Al inicio de cada episodio, nuestro entorno se encargará de devolver una imagen del conjunto de datos (entrenamiento, test o validación), así como las coordenadas del punto en el que se encuentra la persona, expresados en píxeles, que es lo que compone nuestro estado tal y como describimos en la sección 3.3.2. Estos valores se guardan durante toda la ejecución del episodio y se renuevan una vez que se empieza uno nuevo.

La ejecución de cada acción sobre el entorno nos devolverá un nuevo estado, la recompensa asociada con ese estado y un valor que nos indicará si el estado es final o no, es decir, si el episodio finaliza en ese estado.

El algoritmo 4 detalla el pseudocódigo utilizado tras recibir una acción por parte del agente. Podemos observar que este algoritmo mueve el punto del agente basándose en la acción que este toma, siempre en una cantidad fija de píxeles, que corresponde con el tamaño de cada una de las secciones en las que fue dividida la imagen tal y como se explica en el apartado 3.3.3. Esto se hizo de tal forma que el agente se moviese siempre a una sección diferente y por lo tanto la recompensa obtenida también sea diferente.

Algoritmo 4: Ejecutar acción en el entorno

```
width ← 640
distance ← 20
if accion == 'LEFT' then
    | point ← point - distance
end
if action == 'RIGHT' then
    | point ← point + distance
end
point ← torch.clamp(point, 1, width - 1)
reward ← calculateReward(point)
done = EnvironmentMaxReward == reward
return reward, point, done
```

Podemos observar también que la acción de permanecer quieto no tiene ninguna consecuencia en el estado del agente y la utilizaremos como posible acción para finalizar el episodio por parte del agente.

Durante las fases de entrenamiento, testeo y validación de nuestro agente, cada una de las acciones sobre el entorno involucra el cálculo de la nueva posición del punto en el que se encuentra el agente. Una vez el agente fuese desplegado para controlar el dron, cada ejecución de la acción involucraría la ejecución de esa acción en el dron, ya sea girar a la derecha, a la izquierda, o permanecer quieto y obtener una nueva imagen. En este caso, el punto de vista del dron siempre sería el centro de la imagen.

3.3.5. Definición del agente

El agente es el componente de nuestra solución que se encarga de interactuar con el entorno. En este caso, el agente será el encargado de tomar la decisión de girar la cámara del dron a la izquierda, a la derecha o de permanecer quieto, dado un estado del entorno, el cual está definido como se detalla en la sección 3.3.2.

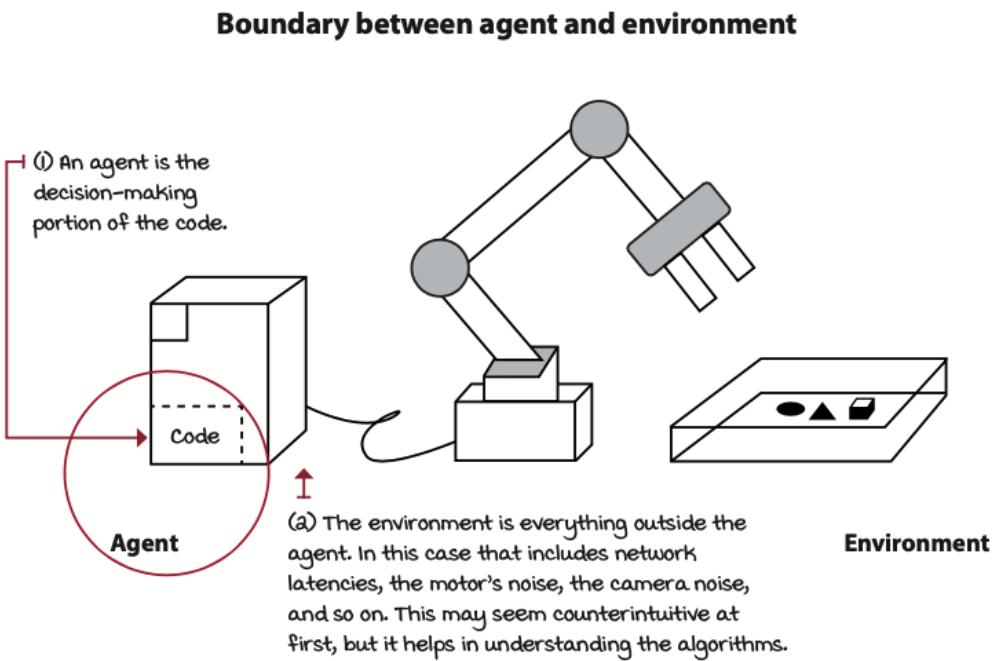


Figura 3.6: Interacción agente-entorno. Imagen tomada del libro *Grokking Deep Reinforcement Learning Morales* (2020).

El objetivo de nuestro agente será encontrar aquellas acciones que maximicen las recompensas obtenidas. Este objetivo será alcanzado dependiendo del algoritmo que utilicemos para ello, pero independientemente de esto, las recompensas obtenidas y el conjunto de acciones será el mismo.

En cuanto a la implementación, el agente será una red neuronal cuya arquitectura iremos variando según los diferentes experimentos, que extenderá de la clase `nn.Module` de `PyTorch` y a la cual le añadiremos métodos propios para facilitar la lectura del código.

3.3.6. Entrenamiento

En cuanto al proceso de entrenamiento, destacamos el hecho de que cada episodio está compuesto por una sola imagen. Es decir, cuando el agente llegue a la recompensa máxima establecida en cada imagen o cuando se llegue al número máximo de acciones que definimos durante el bucle de ejecución, el episodio acabaría.

La inclusión de un número máximo de pasos por imagen fue una decisión tomada para que nuestro agente no entrase en un bucle infinito si decidiese por ejemplo, girar siempre hacia la derecha, llegando al extremo de la imagen donde no se encuentra la persona. Esta decisión también nos lleva a experimentar con un parámetro extra en nuestro entrenamiento, ya que un número muy elevado de acciones puede llevarnos a que el agente no aprenda correctamente y por lo tanto la exploración sobre el entorno no tenga ningún efecto, y por otro lado, un número muy pequeño podría llevarnos a que el agente no tuviese el tiempo suficiente para aprender la policy adecuada

dado el estado. Por lo general, este valor oscilaba entre 50 y 100 acciones por imagen, aunque también se realizaron pruebas con valores menores y mayores a estos.

Debido al problema de la falta de imágenes en diferentes entornos, tal y como comentamos en la sección 3.2. Se tomó la decisión de que el punto inicial se inicialice de manera aleatoria, entre un valor de 0 y 1, que luego será multiplicado por el ancho de la imagen para darnos las coordenadas reales y calcular la recompensa en ese punto. La idea detrás de esta decisión es dotar al entrenamiento de un componente aleatorio y por lo tanto evitar que nuestro agente pueda realizar overfitting sobre el conjunto de datos, dado que para una misma imagen , durante el entrenamiento, el punto de partida sea diferente.

Durante el proceso de entrenamiento se realizaron multitud de pruebas, no solo en lo que respecta a los parámetros de nuestro modelo, sino también a los criterios de parada de cada episodio, el número máximo de acciones a tomar o la recompensa obtenida al seleccionar la acción de permanecer quieto. De estas diferentes decisiones hablaremos en cada uno de los experimentos.

3.3.7. Evaluación

Para evaluar la calidad de los resultados del agente, comprobaremos sobre cada imagen del conjunto de validación, cuándo el agente decide quedarse quieto y la recompensa que obtiene al hacerlo. Esto se asemeja al comportamiento real que tendría al ser desplegado en el dron.

La evaluación comienza con las coordenadas del punto del agente en las coordenadas centrales de la imagen, de manera que simula el punto central de la cámara del dron. Lo que se hace a continuación es ejecutar el agente sobre ese estado y aplicar las acciones de este hasta encontrarnos con la acción de permanecer quieto, lo cual sería el indicador de que el agente se encuentra en la posición en la que se encuentra la persona. En ese momento, dado que contamos con las coordenadas reales obtenidas durante el procesamiento, calculamos la recompensa.

Esta operación se realiza tanto para el conjunto de test durante el entrenamiento, como para el conjunto de validación después del entrenamiento. En ambos casos, nuestro modelo funciona en modo evaluación y por lo tanto no se propagan cambios a los pesos de este.

Además de la recompensa total, nos interesa en este caso obtener un agente que decida moverse de manera uniforme hacia una dirección en el menor número de acciones posible. Pensemos que cada acción que tome el agente será una acción que el dispositivo tendrá que realizar y por lo tanto es tiempo de ejecución que se pierde. Es decir, un agente que consigue la recompensa máxima en 5 acciones será más valioso que uno que la consiga en 10.

Aunque el hecho de conseguir la recompensa máxima del entorno en nuestro caso no es indicativo de la calidad de nuestro agente, por ejemplo podemos pensar que nuestro agente se acerca al punto en el que se encuentra la persona, sin llegar a estar exactamente en donde debería, pero sin embargo lo hace de manera consistente, en un número de acciones reducido por frame, lo que al final nos permite ser más rápidos en la ejecución final y por lo tanto, a efectos prácticos, nos puede incluso llegar a ser más valiosos.

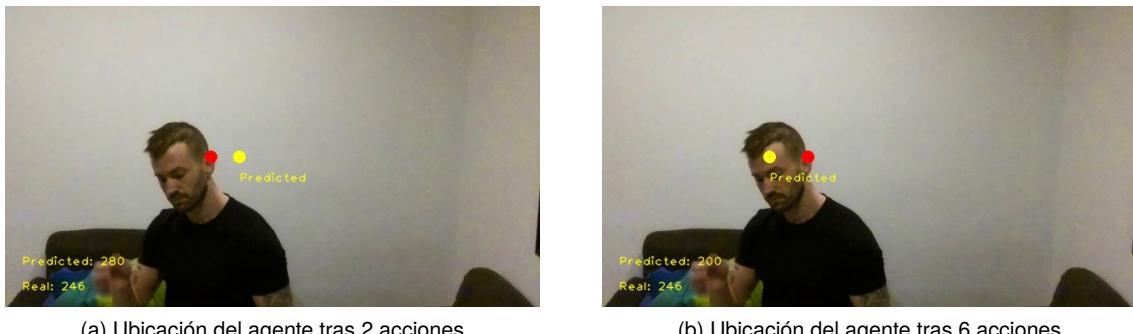


Figura 3.7: Imagen ejecutada en dos experimentos distintos, cuyo punto predicho se encuentra a la misma distancia del punto objetivo, con diferente número de acciones.

Un ejemplo de este comportamiento lo vemos en la figura 3.7, en el que se puede ver que sobre la misma imagen, ejecutada en 2 experimentos diferentes, tenemos 2 puntos de vista por parte del agente a prácticamente la misma distancia. La imagen de la izquierda requirió de solamente 2 acciones, mientras que la imagen de la derecha de 6 acciones. Por lo tanto, para esta imagen nos interesaría el agente que produjo la imagen de la izquierda.

Debido a la alta latencia que se producía al conectar el dron con nuestro ordenador, la ejecución en el dispositivo real no fue posible. Lo que se hizo en su lugar fue analizar frame por frame una serie de videos grabados desde el dispositivo, ejecutando una acción en cada uno de ellos para que simulara la acción a tiempo real que tomaría el agente. Todos estos factores están sujetos a una evaluación subjetiva que iremos comentando más adelante en los diferentes experimentos que realicemos.

Experimentación

4

4.1. Introducción

A continuación haremos un repaso de los resultados obtenidos en el subconjunto de experimentos más representativo, realizados con los diferentes algoritmos propuestos.

Detallaremos brevemente también las conclusiones parciales que obtenemos en cada experimento, así como posibles soluciones a los problemas que se nos plantean en cada uno.

Debido al coste computacional de ejecutar cada experimento sobre el conjunto de datos total, cada experimento se compone de subexperimentos previos, en los cuales intentábamos probar nuestras soluciones con un conjunto de datos menor al original, alrededor de unas 20 imágenes.

La idea detrás de esta subexperimentación era conseguir un modelo de agente que sea capaz de realizar *overfitting* sobre el conjunto de datos de manera relativamente rápida y que obtuviese una convergencia tanto en la recompensa obtenida tanto en las imágenes del conjunto de entrenamiento como en el conjunto de test como en el número de pasos que realizaba sobre cada una de las imágenes. Esto nos permitió también acortar el tiempo entre las diferentes pruebas, ya que el conseguir este *overfitting* nos aseguraba que el agente seguiría en un principio las ideas que queríamos implementar y además provocó que pudiésemos experimentar con diferentes definiciones tanto de nuestra recompensa como de nuestro entorno, así como hacer *fine tuning* de nuestros hiperparámetros.

Ligado a la subexperimentación comentada, en cuanto a los hiperparámetros, vimos como el learning rate jugaba un papel muy importante en todos y cada uno de los algoritmos usados. Un *learning rate* muy alto hacía que nuestros agentes solo tomasen una decisión, llevando a escorarse prácticamente hacia un lado en todas las ocasiones. Por otro lado, un *learning rate* muy bajo no permitía que nuestro agente fuese capaz muchas veces de tomar una decisión clara, y aunque esto pueda entenderse como parte de la exploración del agente muchas veces no permitía al agente obtener la recompensa máxima del entorno en el número máximo de pasos que se había establecido y obtenía por lo tanto solo recompensas parciales.

Durante los experimentos podremos comprobar que en alguno de ellos el número de métricas que mostraremos es más alto que en otros, esto no es solo debido a que queramos destacar una cualidad específica del experimento, sino que también formó parte de la experimentación y el descubrimiento de algunos de los problemas que nos fuimos encontrando, que nos llevó a tener que

realizar el seguimiento de algunos aspectos del entrenamiento que no estaban contemplados en experimentos anteriores. Estos problemas estuvieron relacionados principalmente con la convergencia en la toma de decisiones del agente tal y como se discutirá más adelante.

Por último también mostraremos algunos resultados de las imágenes que fuimos obteniendo en cada caso. Cabe recordar que la ejecución del agente durante el test se detiene cuando decide ejecutar la acción de permanecer quieto, y que tanto esta recompensa como el número de acciones hasta llegar a ella se tienen en cuenta como medidas subjetivas de la eficacia del agente, tal y como comentamos en la sección 3.3.7.

El total de los experimentos está disponible en el repositorio de [GitHub](#) para un análisis más extenso y se adjunta el título de cada uno de ellos en el Apéndice A.

La herramienta utilizada para monitorear el progreso de las diferentes métricas durante los diferentes entrenamientos es [Tensorboard](#), ya que además de su fácil integración con [PyTorch](#) nos permitía comparar dichas métricas entre los diferentes experimentos.

4.2. Policy Gradient

El algoritmo *Policy Gradient* fue nuestra primera opción a la hora de implementar nuestro agente. Se buscaba obtener una primera aproximación a nuestro problema y buscábamos sobre todo tener un modelo que sirviese como *baseline*, así como definir la estructura general de nuestro programa.

4.2.1. Experimento 1

En este primer experimento utilizamos como criterio de finalización del episodio que el agente realizase un máximo de 50 acciones o que obtuviese la máxima recompensa del entorno, en este caso se mantuvo la recompensa proporcional con la distancia al punto objetivo y no se hizo ninguna recompensa extra por llegar al final del episodio correctamente.

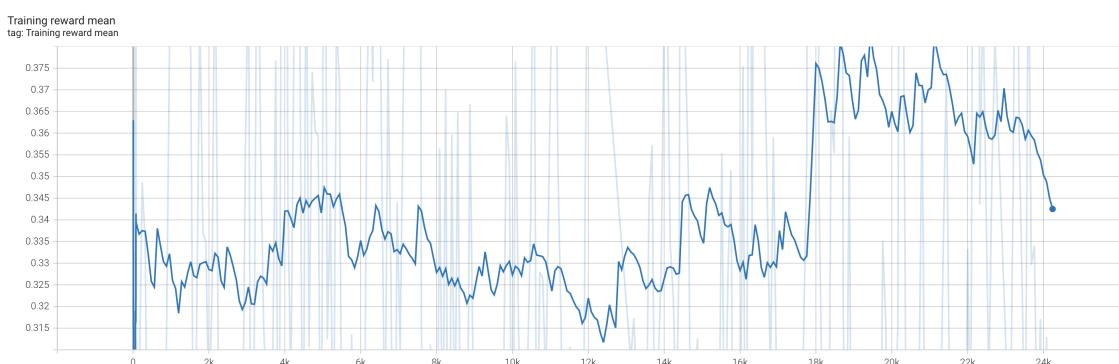


Figura 4.1: Experimento Policy Gradient 1 - Recompensa media en el conjunto de entrenamiento

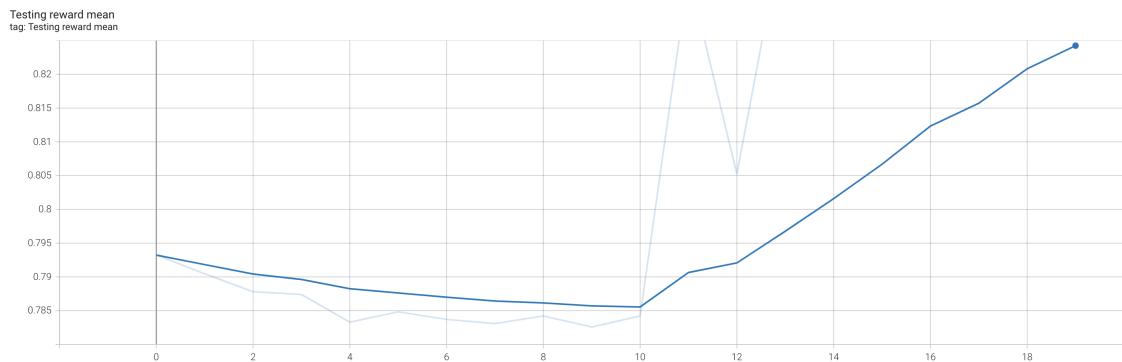


Figura 4.2: Experimento Policy Gradient 1 - Testing reward mean



Figura 4.3: Experimento Policy Gradient 1 - Duración de los episodios

Lo que se puede apreciar es que existe una tendencia a la baja de nuestro algoritmo en cuanto al número de acciones en el conjunto de entrenamiento. Por otra parte, se aprecia que la recompensa en el conjunto de test comienza a subir a partir de la época 10, lo cual marca un punto importante de inflexión en esta.

4.2.2. Experimento 2

Los parámetros utilizados en este experimento son similares al anterior, pero en este caso el entrenamiento se ejecuta durante unas 50 épocas, unas 30 más que el experimento 4.2.1.

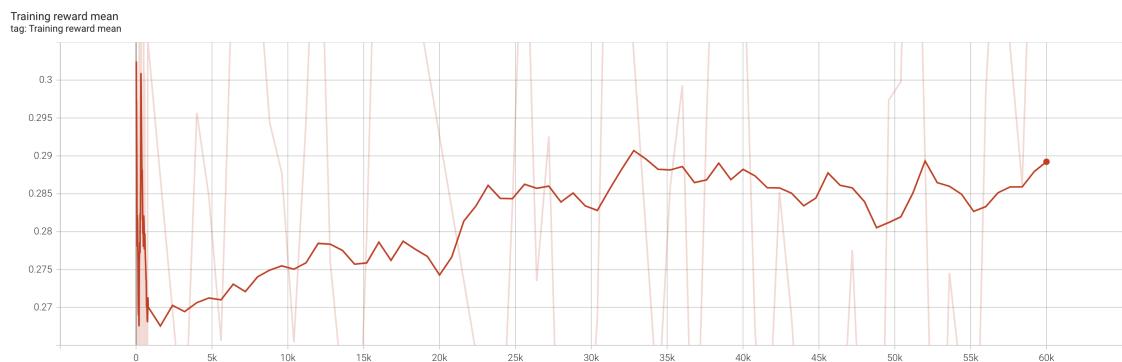


Figura 4.4: Experimento Policy Gradient 2 - Recompensa media en el conjunto de entrenamiento

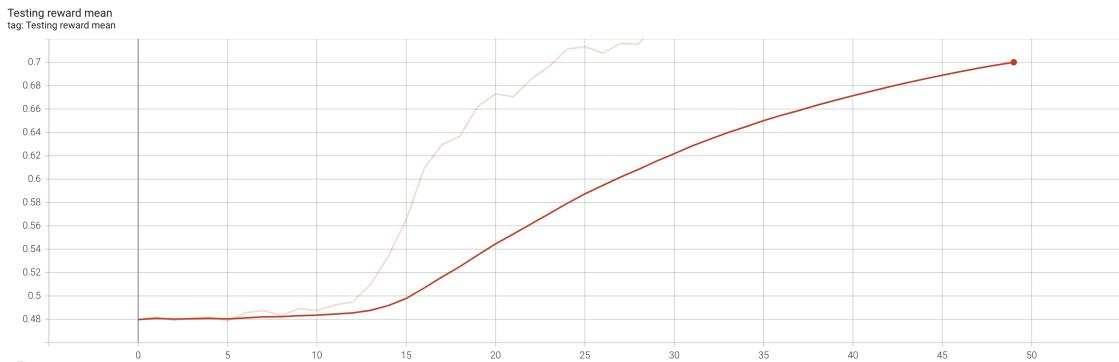


Figura 4.5: Experimento Policy Gradient 2 - Testing reward mean

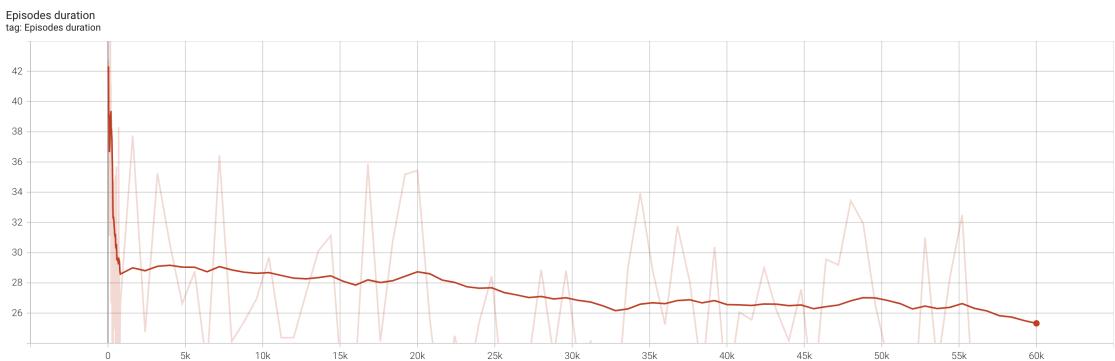


Figura 4.6: Experimento Policy Gradient 2 - Duración de los episodios

Vemos que en este experimento la tendencia que vimos en el experimento anterior se confirma: el número de acciones por imagen en el conjunto de entrenamiento parece seguir una tendencia a la baja y tanto la recompensa en el conjunto de test como en el entrenamiento aumentan según van avanzando las épocas.

4.2.3. Conclusiones

Los resultados obtenidos por este primer algoritmo fueron destacables en cuanto a las expectativas esperadas en un principio. Sin embargo nos encontramos con algunas debilidades en la definición de nuestro problema y sobre todo en la manera en la que recompensábamos cada acción del agente.

En este sentido, un problema difícil de detectar al principio y que se repitió también a lo largo de los demás experimentos, fue el hecho de que el agente decidiese moverse, es decir, a medida que el entrenamiento avanzaba, se podía comprobar no solo que era fácil que el agente realizase overfitting sino que también decidiese que la mejor acción a tomar fuese la de permanecer quieto. Recordemos que la recompensa en estos casos era inversamente proporcional a la distancia en la que se encontraba nuestro punto objetivo comparado con el punto de nuestro agente.

Por concluir, podemos decir que este primer conjunto de experimentos nos permitió entender

la importancia de nuestra política de recompensa y resaltó problemas que no habíamos tenido en cuenta previamente en relación con las decisiones del agente.

4.3. Actor-Critic

Usando los resultados previos con el algoritmo Policy Gradient, lo que buscamos con estos nuevos experimentos es obtener una mejora en cuanto a resultados y en cuanto a la estabilización de nuestro entrenamiento, gracias a la doble salida que obtiene nuestro agente, por un lado la distribución de probabilidad de las acciones a tomar y por otro lado el valor del estado en el que se encuentra.

Para intentar evitar los problemas mencionados en los experimentos anteriores, en esta fase nos centramos ya no tanto en la búsqueda de los hiperparámetros correctos para el entrenamiento del agente, ya que usaremos las mismas técnicas utilizadas previamente, sino en refinar y construir un sistema de recompensa que se adecue más al objetivo que queríamos lograr: llegar al punto objetivo con el menor de acciones posibles.

4.3.1. Experimento 1

(temporal) Tensorboard: Actor-Critic-v2

Para este primer experimento, dejamos el sistema de recompensa previo, pero modificamos el número total de acciones por imagen, aumentándolo de 50 a 100.

El entrenamiento se ejecutó durante 20 epochs, basándonos en la convergencia obtenida en los experimentos anteriores y en el tiempo que conlleva.

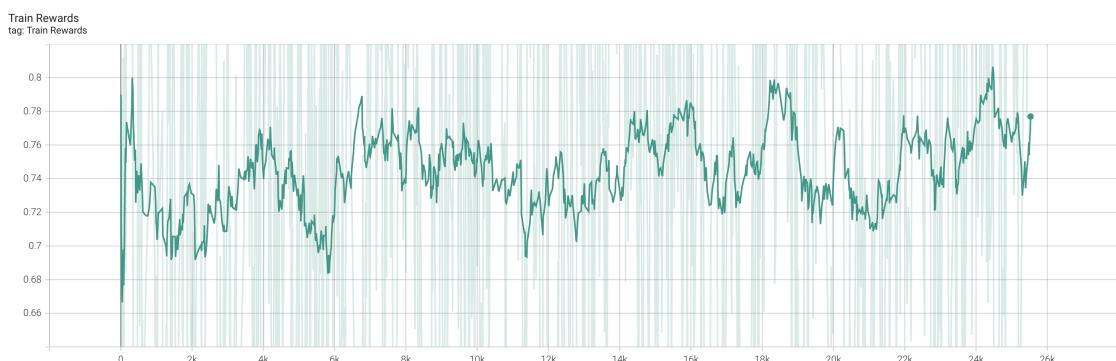


Figura 4.7: Experimento Actor Critic 1 - Recompensa media en el conjunto de entrenamiento

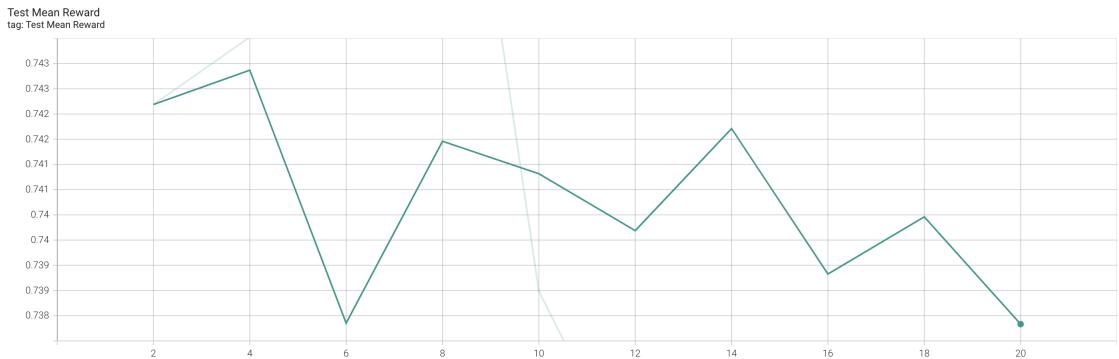


Figura 4.8: Experimento Actor Critic 1 - Testing reward mean

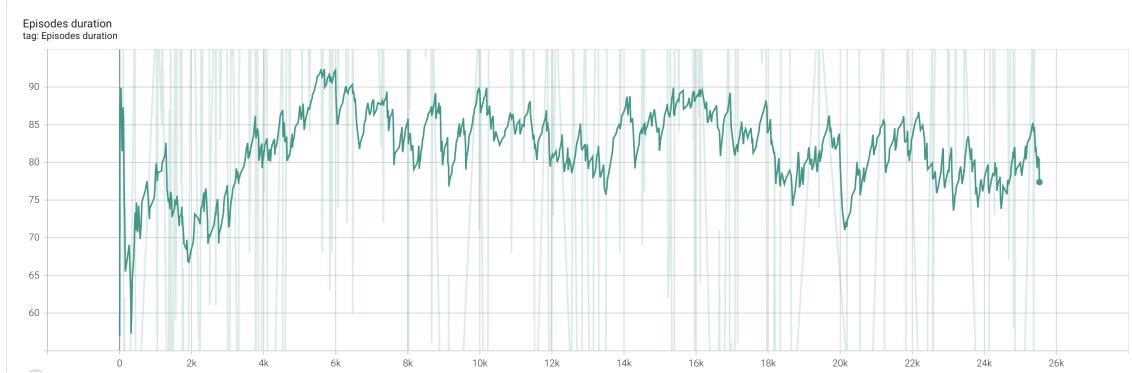


Figura 4.9: Experimento Actor Critic 1 - Duración de los episodios

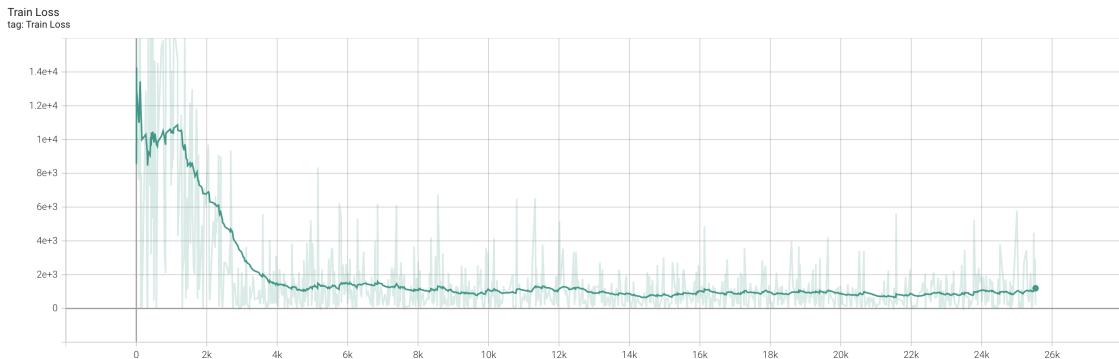


Figura 4.10: Experimento Actor Critic 1 - Train loss

Lo que podemos ver en este primer experimento es que, a pesar de nuestras expectativas iniciales, el entrenamiento fue muy inestable, incluso más que en el caso de *Policy Gradient*. Si bien se aprecia una tendencia a la alza en cuanto a la recompensa en el conjunto de datos de entrenamiento (figura 4.7), vemos que no ocurre lo mismo en el conjunto de datos de test (fig. 4.8).

Sin embargo, observamos que la función de coste en el entrenamiento desciende rápidamente y que se mantiene estable en un nivel bajo comparado con el inicial, lo cual nos hace sospechar

de que el entrenamiento queda estancado a partir de ese punto, siendo esto un posible síntoma de que la red neuronal de nuestro agente no es lo suficientemente potente.

4.3.2. Experimento 2

(temporal) (temporal) Tensorboard: Actor-Critic-ac-no-rewards-till-complete Para este ejemplo decidimos incrementar el número de *epochs*, de las 20 del experimento anterior hasta las 100, esperando que se produjese un salto en cuanto a la calidad del entrenamiento. Además disminuimos levemente el *learning rate*.

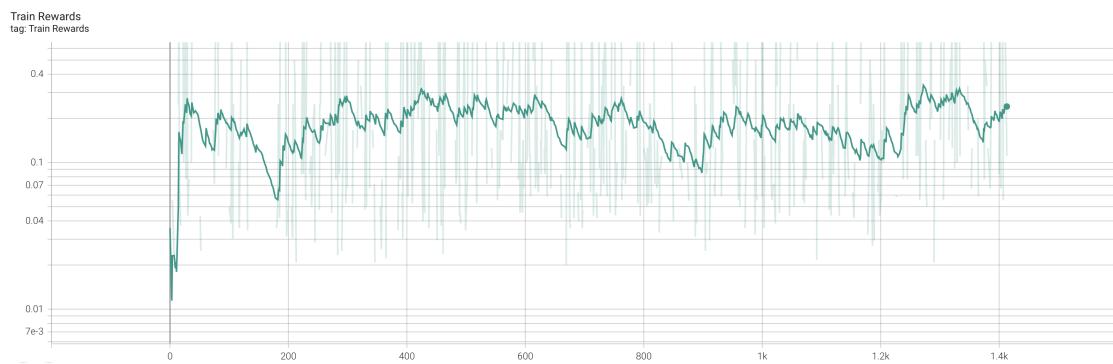


Figura 4.11: Experimento Actor Critic 2 - Recompensa media en el conjunto de entrenamiento

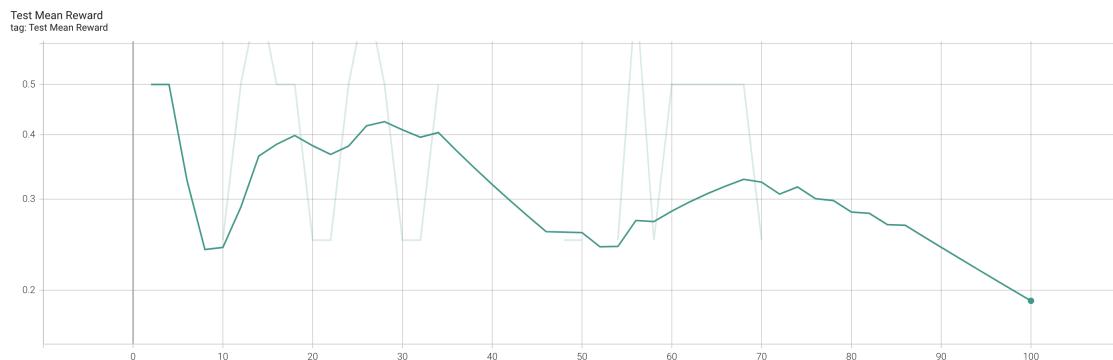


Figura 4.12: Experimento Actor Critic 2 - Testing reward mean

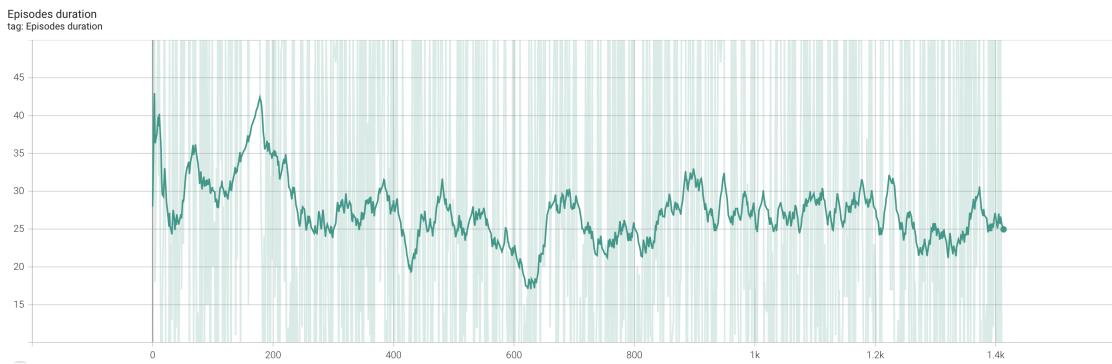


Figura 4.13: Experimento Actor Critic 2 - Duración de los episodios

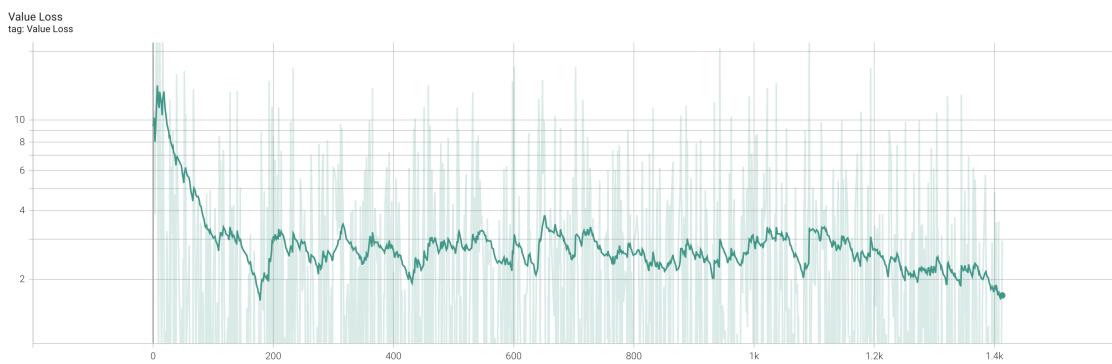


Figura 4.14: Experimento Actor Critic 2 - Value loss

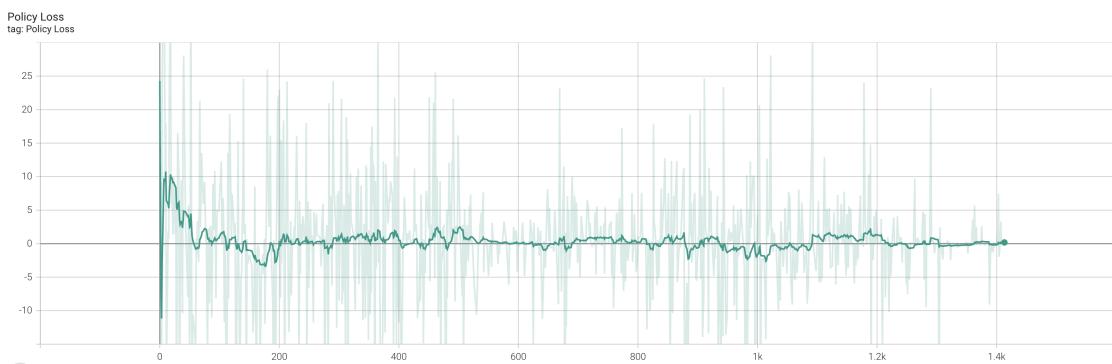


Figura 4.15: Experimento Actor Critic 2 - Policy loss

En este caso, cambiamos el sistema de recompensa al agente. Lo que hicimos fue darle una recompensa de 0 hasta que no llegase al final del episodio si es que no superaba el número de acciones máxima. La idea era reforzar aquellas experiencias en las que el agente tenía un resultado positivo y castigar aquellos intentos en los que no llegase al punto objetivo.

Sin embargo, para disminuir la dificultad del episodio, lo que hicimos fue crear una zona de unos 30 píxeles tanto a la izquierda como a la derecha del punto objetivo, de manera que si el agente se acercaba por alguno de los lados, el episodio acabaría de manera satisfactoria.

Al igual que en el experimento anterior, lo que podemos observar es que el agente aprende rápido durante las primeras iteraciones pero luego no se ve ninguna mejora significativa. En la figura 4.14 vemos como la función de coste del agente si bien se mantiene prácticamente estable, va disminuyendo lentamente durante todo el entrenamiento, mientras que la *policy loss* (fig. 4.15) es prácticamente la misma.

4.3.3. Experimento 3

(temporal) Tensorboard: Actor-Critic-ac-reward-2-rew-by-two-stop-with-none

En este experimento lo que haremos será intentar atacar varias situaciones que nos encontramos en experimentos anteriores. Entre ellas el problema que también comentamos durante el análisis de los experimentos con *Policy Gradient* y es el hecho de que el agente decide permanecer quieto nada más comenzar el episodio cuando se ejecuta el conjunto de test, llevando por un lado a una rápida finalización de los episodios, pero limitando la opción de explorar más allá del punto inicial. Como comentamos en un principio, idealmente nuestro agente debería finalizar, o lo que es lo mismo, decidir no moverse cuando está seguro de que el punto en el que se encuentra es óptimo. Esta situación en particular se discutirá en el apartado 4.3.5.

Para ello, el sistema de recompensa fue modificado de la siguiente manera:

- En el caso de que el agente haya llegado al punto objetivo, se le dará una recompensa de 2, en vez de 1 (se mantiene el mismo formato de recompensas parciales si no se llega al punto objetivo).
- Si el agente escoge la acción de permanecer quieto y no lo hace en el punto de recompensa máxima, la recompensa que obtiene para esa acción se divide entre 2.

El objetivo de estas modificaciones es que el agente tenga una motivación extra por llegar al final del episodio dándole una recompensa extra en caso de que así sea, lo que bajo nuestra hipótesis inicial incitaría a que el agente no decida detenerse nada más comenzar el episodio.

El siguiente objetivo, relacionado directamente con el anterior, es desfavorecer la elección de permanecer quieto a no ser que el agente esté seguro de ello dividiendo la recompensa que obtiene en dicha acción si no se encuentra en el punto esperado.



Figura 4.16: Experimento Actor Critic 3 - Recompensa media en el conjunto de entrenamiento



Figura 4.17: Experimento Actor Critic 3 - Función de perdida

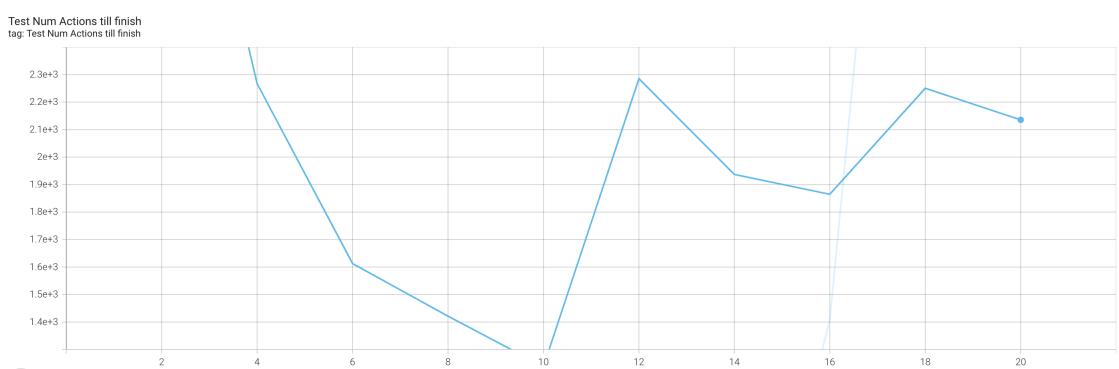


Figura 4.18: Experimento Actor Critic 3 - Número de acciones media por imagen

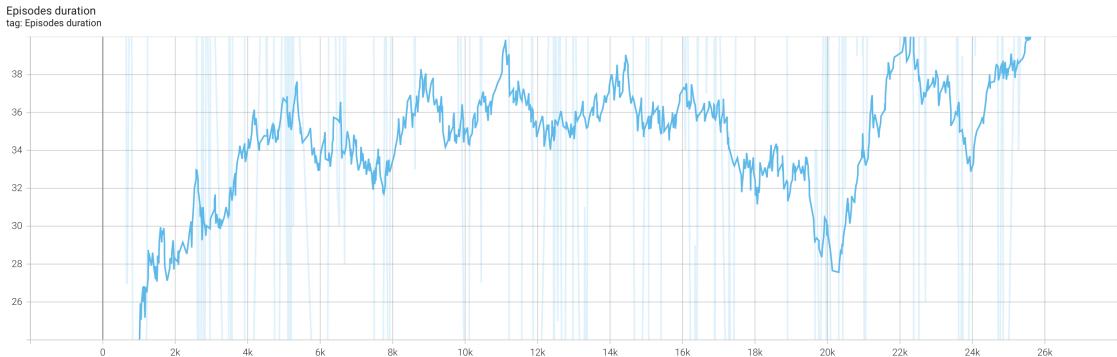


Figura 4.19: Experimento Actor Critic 3 - Duración de los episodios

Como podemos observar, los cambios aplicados sobre nuestro sistema de recompensa tuvo un gran impacto en el entrenamiento del agente. Si observamos la figura 4.16 vemos que la recompensa media no siguió una tendencia, sino que fue variando hasta llegar a un pico casi al final del entrenamiento para finalmente descender. En ese mismo punto vemos que la función de perdida en el conjunto de entrenamiento disminuye (fig. 4.17), lo cual en general nos podría indicar un buen rendimiento de nuestro agente en ese punto.

Durante este experimento, dados los cambios realizados sobre el *reward*, nos pareció interesante realizar un seguimiento al número de acciones que el agente realizaba sobre las imágenes del conjunto de testeo, lo que se puede ver en la figura 4.18. Lo que observamos en la gráfica es que, si bien nuestro objetivo era incrementar la confianza del agente a la hora de tomar la acción de permanecer quieto, lo que conseguimos fue que el número de acciones por episodio en el conjunto de test se disparara sustancialmente, llegando prácticamente a las 2000 acciones de media por imagen, lo que haría prácticamente inviable su despliegue en un entorno real. Este resultado nos lleva a pensar que nuestra hipótesis desarrollo una desconfianza en el agente a la hora de tomar la decisión de no moverse, lo que hace que prefiera moverse hacia la izquierda o hacia a la derecha.

Esta tendencia queda confirmada cuando observamos el archivo de registro del experimento (disponible en el repositorio del proyecto) en el que muestra las acciones escogidas por el agente en cada imagen:

Época / Episodio	Moverse a la izquierda	Moverse a la derecha	Permanecer quieto
Época 0/ Episodio 65	10	8	4
Época 0/ Episodio 285	109	37	5
Época 1/ Episodio 675	77	101	0
Época 2/ Episodio 715	107	94	0

Tabla 4.1: Acciones escogidas por el agente durante el entrenamiento

Al principio del entrenamiento el agente comienza tomando acciones aleatorias, lo cual es esperable debido a que no tiene experiencias previas con el entorno. Según va avanzando su cono-

cimiento, vemos que incluso ya durante la primera época, el agente aprende a relacionar la acción de no moverse con una recompensa negativa y por lo tanto decide cada vez con mayor frecuencia no usarla, lo cual implica disminuir la probabilidad de esa acción en particular. Al cabo de la primera época ya observamos que la acción prácticamente ya no es escogida, lo cual nos lleva a confirmar nuestra teoría de que nuestro nuevo sistema de recompensa no está produciendo los resultados esperados.

La investigación alrededor de este problema fue un punto clave en el desarrollo del proyecto y fue lo que nos llevó a probar diferentes alternativas no solo en cuanto al sistema de recompensa/castigo, probando con diferentes opciones como castigar al agente si el número de acciones era demasiado bajo o penalizar la trayectoria del episodio si no se terminó este con la recompensa máxima, sino también a explorar diferentes criterios de parada en el episodio.

En lo relativo a los criterios de parada se decidió también probar como condición que fuese la propia acción de no moverse en lugar de la obtención de la recompensa máxima, junto con el número máximo de acciones por imagen para evitar un bucle infinito en el entrenamiento.

4.3.4. Experimento 4 - Vision Transformers

Para finalizar con esta sección de experimentos, decidimos incluir nuestro intento usando Vision Transformers como parte del preprocesamiento de la imagen. Esto a su vez involucró una transformación de la salida del Transformer a la capa de entrada de nuestro agente. Esto se realizó utilizando una red neuronal intermedia, ya que nuestro agente inicialmente recibía un vector con 4097 elementos como entrada y la salida del Transformer era una matriz de 197x512 elementos.

Lo que intentábamos buscar con este experimento era el poder descartar la posibilidad de que la representación de nuestra imagen no fuese lo suficientemente característica para que nuestro agente pudiese distinguir qué era lo importante y que no.

Debido a la carga computacional de este modelo y las restricciones en cuanto a tiempo, no se pudieron realizar múltiples experimentos con la misma facilidad que antes. Sin embargo, nos sirvió para explorar e investigar otras soluciones, y aunque esta no fue la única solución que tuvimos en cuenta, nos pareció interesante considerarla en la memoria debido a la popularidad que este tipo de redes cuentan a día de hoy y de la posibilidad de introducirlas en un problema de aprendizaje por refuerzo.

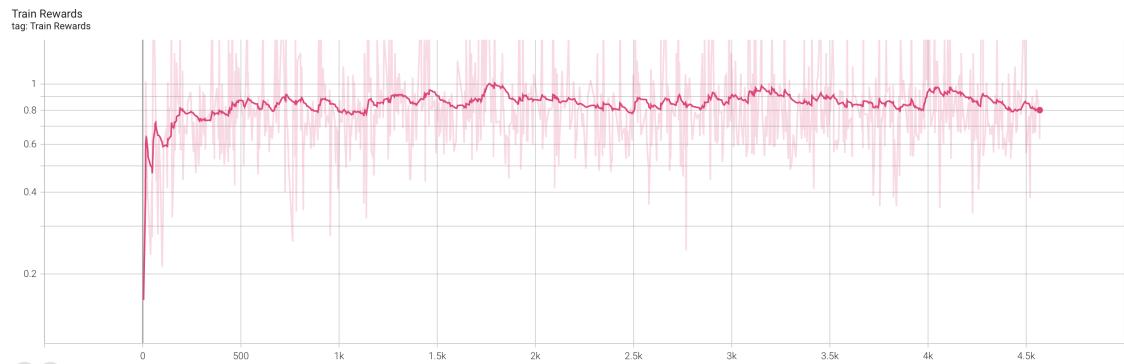


Figura 4.20: Experimento *Vision Transformer* - Recompensa media en el conjunto de entrenamiento

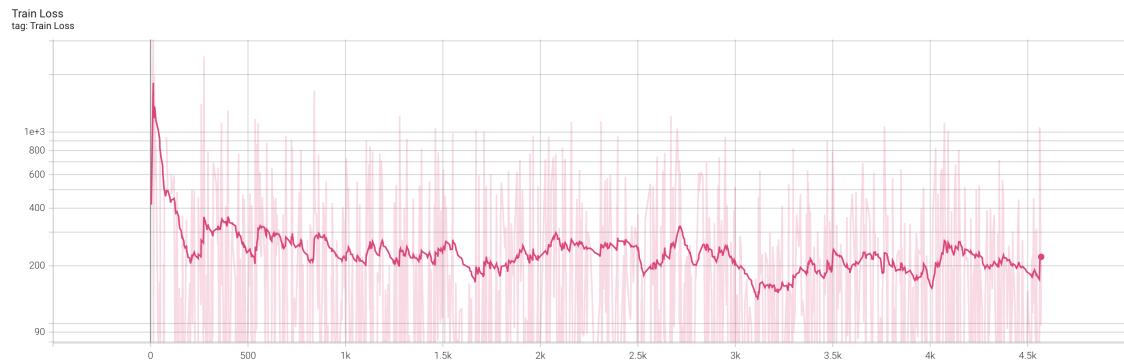


Figura 4.21: Experimento *Vision Transformer* - Función de perdida en el conjunto de entrenamiento

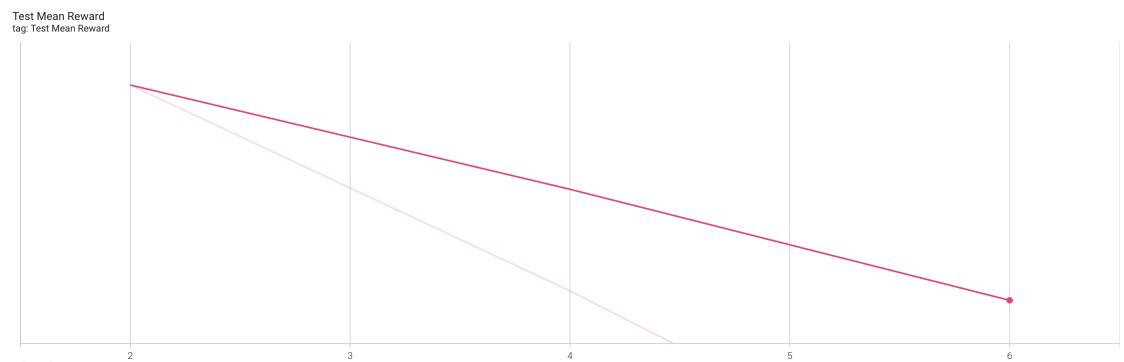


Figura 4.22: Experimento *Vision Transformer* - Testing reward mean



Figura 4.23: Experimento *Vision Transformer* - Duración de los episodios

Lo que se puede apreciar en las gráficas del entrenamiento anteriores es que no obtuvimos ninguna ventaja significativa con el uso de *Vision Transformers*, y aunque esto puede deberse a múltiples factores distintos, descartamos la posibilidad de seguir realizando experimentos con esta arquitectura dado el tiempo de entrenamiento que conllevaba y que no veíamos con claridad que fueseemos a obtener una convergencia o un mejor resultado que usando la red *VGG19* preentrenada.

Al igual que con los experimentos anteriores, vemos en la figura 4.22 que la recompensa sobre el conjunto de test disminuye con el paso del entrenamiento, lo cual nos indicaría que nuestro problema no se encuentra en un principio en la capacidad de abstracción de características de la imagen, sino que podría deberse a una falta de datos en general o incluso a que este tipo de redes necesitan un mayor tiempo de entrenamiento hasta conseguir buenos resultados. Sin embargo, y dado que esto es solamente el paso previo a nuestro agente, podríamos afirmar que la inestabilidad del entrenamiento y el no avance en la recompensa en el conjunto de entrenamiento (figura 4.20), se debe a una falta de capacidad en la red de nuestro agente.

4.3.5. Conclusiones

Tras analizar los diferentes experimentos utilizando el algoritmo *Actor Critic* concluimos que, por un lado, se trata de un algoritmo más potente en cuanto a la capacidad de aprendizaje, lo cual es apreciable en la velocidad con la cual la función de perdida disminuye al principio del entrenamiento.

Por otro lado y para finalizar, vemos que aunque esa convergencia hacia una solución es muy fuerte, solo sucede al principio del entrenamiento, lo cual nos indica que nuestro límite podría no estar en el propio algoritmo sino en las diferentes decisiones que tomamos a la hora de definir nuestro entorno y sobre todo nuestro sistema de recompensa.

5

Resultados

En este capítulo discutiremos los resultados obtenidos a lo largo de los diferentes experimentos realizados, utilizando para ello las imágenes del conjunto de validación.

Comenzaremos evaluando los resultados obtenidos con el algoritmo de *Policy Gradient*. Podemos observar en la figura 5.1 que los resultados fueron en general satisfactorios con lo que buscábamos en un principio y en general el entrenamiento se mantuvo particularmente estable con respecto a los demás experimentos.

Sin embargo, en algunas ocasiones se aprecia que si el sujeto se encuentra en un extremo de la imagen, el agente no es capaz de moverse con soltura hacia ese lado tal y como vemos en la figura 5.1 (b) y (c). Por lo general podemos decir que el agente se comporta de manera correcta, pero el número de acciones y su rendimiento en imágenes que podemos denominar ".extremas" todavía tiene un margen considerable de mejora.



(a) 4 acciones tomadas



(b) 9 acciones tomadas



(c) 5 acciones tomadas



(d) 7 acciones tomadas

Figura 5.1: Imágenes obtenidas en los experimentos usando *Policy Gradient* sobre el conjunto de test.

A continuación analizaremos los resultados obtenidos con el algoritmo *Actor Critic*. En la figu-

ra 5.2 vemos un ejemplo del comportamiento que comentamos durante el análisis en la sección anterior y es que el agente decide no moverse al principio del entrenamiento. En las 4 imágenes presentadas el agente decidió no tomar ninguna acción sino directamente permanecer quieto y con ello terminar el episodio.

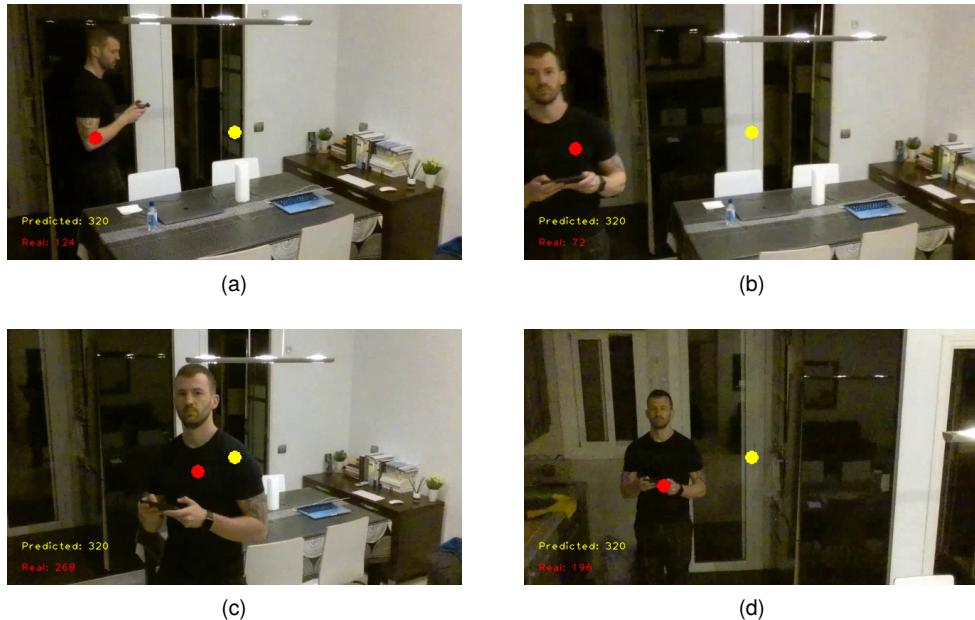


Figura 5.2: Imágenes obtenidas con *Actor Critic* sin incentivar al agente a moverse.

La dificultad de diagnosticar este problema fue que, siguiendo las gráficas del entrenamiento, el agente parecía estar teniendo una muy buena *performance* para el conjunto de entrenamiento y de test y era difícil descubrir que decidía no moverse hasta que hicimos un registro de las acciones tomadas por este en cada etapa del entrenamiento.

Época	Recompensa media	Número de acciones tomadas
2	0.8755	0.366
4	0.873	0.088
6	0.872	0.071
10	0.873	0.059
20	0.874	0.053

Tabla 5.1: Resultados Actor Critic - Media de acciones tomadas por el agente en el conjunto de test

La tabla 5.1 muestra un ejemplo de la evolución de la media de acciones tomadas por el agente sobre el conjunto de test en las diferentes etapas del entrenamiento:

Cabe destacar que este patrón se produjo desde los primeros experimentos con *Actor Critic*, y que como vemos en la anterior tabla, la recompensa media no baja significativamente, mientras que el número de acciones media si lo hace. Lo cual también nos hizo recapacitar en cuanto a las diferentes opciones de recompensa de nuestro agente tal y como comentamos en el capítulo previo.

También pudimos apreciar que el rendimiento del agente por lo general disminuía según avanzaba el entrenamiento y al mismo tiempo la duración de los episodios aumentaba, por lo que se producía un punto de inflexión el cual no pudimos explicar. La figura 5.3 muestra un ejemplo de esta evolución del agente durante el entrenamiento.

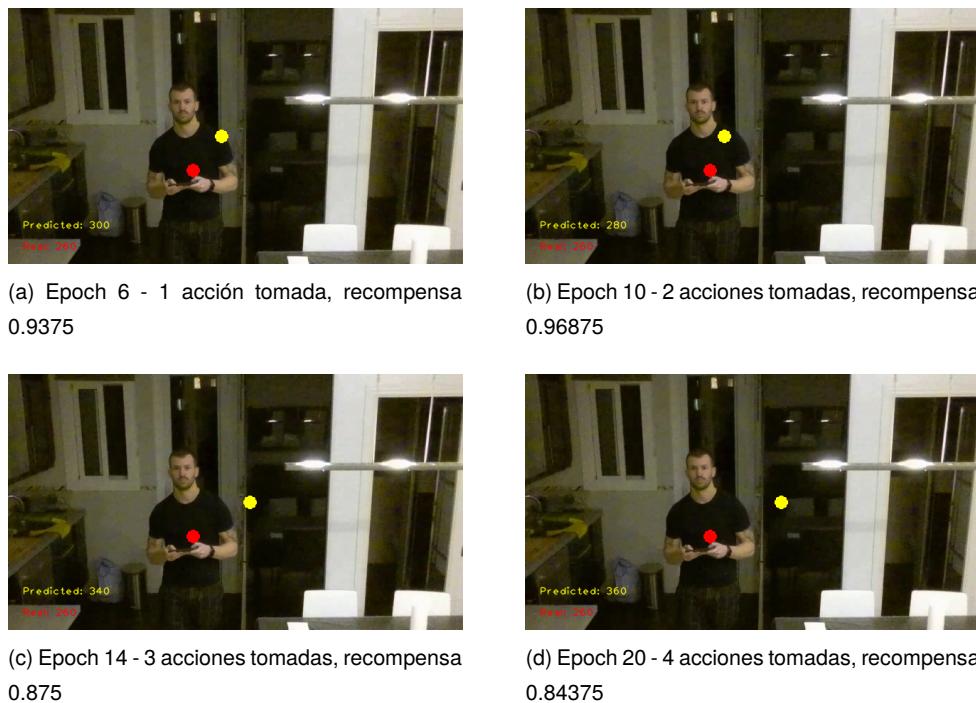


Figura 5.3: Misma imagen en diferentes etapas del entrenamiento usando *Actor Critic*.

Alrededor de las primeras épocas es cuando nuestro algoritmo tuvo un mejor rendimiento y es en ese momento en el que podríamos decir que nuestro agente podría desplegarse en el dispositivo final ya que combina en general pocas decisiones y hacia el lugar adecuado.

Para finalizar el análisis de los diferentes experimentos, podemos comentar que nuestro experimento que incluyó el uso de *Vision Transformers* no obtuvo ninguna ventaja con respecto a los anteriores, aunque esto también puede deberse a la calidad y cantidad de nuestros datos y al hecho de que el tiempo de entrenamiento no fue lo suficiente para permitir aprovechar la potencia de estas redes.

Conclusiones

6

Tras haber analizado los diferentes experimentos, haremos un breve repaso de las conclusiones principales que obtuvimos:

1. La recompensa juega un papel importante.

Como vimos durante los diferentes experimentos, y aunque solo hayamos comentado un subconjunto de los totales realizados, la función de recompensa jugó sin duda un papel clave en todo el trabajo. Las diferentes decisiones sobre esta intentaban apaciguar los diferentes problemas que iban surgiendo tras analizar que los resultados obtenidos no se asemejaban a nuestro idea inicial.

2. Los hiperparámetros no fueron decisivos.

Generalmente asociamos que la convergencia de nuestras redes neuronales está asociada con la elección de buenos hiperparámetros en el entrenamiento, tal y como aprendimos a lo largo del master. Sin embargo, en nuestro caso esto no fue un problema puesto que, si bien el agente podía llegar a un punto de recompensa media antes o después, esto no determinaba el resultado final. Hicimos pruebas con diferentes técnicas de inicialización de pesos, diferentes tasas de aprendizaje e incluimos técnicas conocidas como *Dropout*, sin embargo el resultado no variaba de manera significativa.

3. Definir correctamente el problema y sus componentes.

Esta es la conclusión que finalmente podríamos destacar. Bajo nuestra experiencia a lo largo del desarrollo de este proyecto pudimos observar una mejoría y un empeoramiento del resultado de nuestro agente modificando levemente en algunos casos las definiciones de nuestro entorno, ya sea a través de un cambio en la función de recompensa o en el criterio de parada del entrenamiento. Lo cual nos indica la dificultad que tiene el aplicar algoritmos de aprendizaje por refuerzo a escenarios reales (recordando incluso que el problema que afrontamos contaba con limitaciones con las cuales un producto industrial real tendría que lidiar).

Para finalizar también nos gustaría dejar las conclusiones que obtuvimos a nivel personal tras la realización de este trabajo. Empezaremos diciendo que sin duda alguna el trabajo nos pareció

interesante desde el primer momento, no solo porque queríamos aplicar toda la teoría aprendida durante el master, sino también porque era una rama a la cual no le pudimos dedicar todo el tiempo que nos hubiese gustado en su momento.

A pesar de que bajo nuestro punto de vista el trabajo quedó lejos de ser como nos hubiese gustado, debido a los diferentes problemas que comentamos en secciones anteriores: el entrenamiento de los agentes, la definición del problema y en general el intentar llevar al mundo real un tema que por lo general siempre se relaciona con entornos simulados, con todas las dificultades que ello conlleva, nos quedamos satisfechos de que fue una prueba de las capacidades de este tipo de aprendizajes (y nuestra también) y de que nos motiva aún más a seguir estudiando sobre el tema y que con la experiencia aprendida durante estos meses, poder incluso pensar en realizar nuestros propios proyectos.

Limitaciones y Perspectivas de Futuro

7

En este capítulo hablaremos de las limitaciones con las cuales nos encontramos a la hora de desarrollar nuestro trabajo y haremos un análisis de los principales puntos que podrían mejorarse en futuras exploraciones.

Comenzaremos hablando de los datos. Por regla general podemos asociar que un buen conjunto de datos marca la diferencia en el aprendizaje de una red neuronal profunda, no solo por su cantidad sino también por la variedad y la veracidad de estos comparados con un entorno real. En cuanto al último punto creo que se tomó la decisión correcta, ya que utilizamos imágenes capturadas por el propio dispositivo con el cual esperábamos usar nuestro agente y por lo tanto la calidad de las imágenes sería la misma. Sin embargo, es posible que nos hayamos quedado cortos en cuanto a la variedad de los escenarios, lo cual hizo que ya desde un principio tuviésemos que pensar que el *overfitting* podría llegar a ser un problema. Esto podría ser fácilmente solucionable si procesamos imágenes de diferentes entornos, de manera que nos aseguramos que el agente aprende realmente sobre el problema y no el entorno.

Otro punto que supuso una limitación importante fue la limitada capacidad computacional con la que entrenamos nuestros agentes. En un principio se hicieron pruebas utilizando Google Colab, pero la limitación en cuanto a tiempo y el *workflow* al cual nos obligaba a recurrir, hizo que lo descartásemos. También valoramos el uso de una máquina con procesadores gráficos compatibles con la librería *PyTorch*, pero su costo operativo también hizo que no fuese una opción viable.

También tenemos que hablar de nuestra falta de experiencia en el desarrollo de estos algoritmos y en general en la rama del aprendizaje por refuerzo, lo cual hizo que sumadas a las dificultades técnicas tuviésemos también que repasar los conceptos aprendidos en multitud de ocasiones.

Pensando en trabajos futuros en esta línea de trabajo, dejamos abierta la posibilidad de usar algoritmos que nos ofrezcan una mejor convergencia en los resultados. Un ejemplo de esto podría ser la implementación de los algoritmos A2C y A3C ([Mnih et al., 2016](#)), que nos ofrecen la posibilidad de realizar múltiples entrenamientos al mismo tiempo o incluso podríamos pensar en la combinación de diferentes técnicas de inteligencia artificial como podrían ser el aprendizaje no supervisado o incluso la utilización de algoritmos genéticos para ayudarnos a encontrar una estructura de red óptima de nuestro problema ([Cai et al., 2018](#)).

CAPÍTULO 7. LIMITACIONES Y
PERSPECTIVAS DE FUTURO

Apéndize A

A

El siguiente apéndice muestra el conjunto de experimentos totales realizados durante la ejecución del proyecto. Cada uno de los experimentos se puede encontrar dentro de la carpeta runs/ del código del proyecto en el repositorio de [GitHub](#).

1. Actor-Critic-ac-continuous-action
2. Actor-Critic-ac-discourage
3. Actor-Critic-ac-discourage-reward-1
4. Actor-Critic-ac-discourage-reward-2
5. Actor-Critic-ac-discourage-stop-action
6. Actor-Critic-ac-no-rewards-till-complete
7. Actor-Critic-ac-no-rewards-till-complete-divided
8. Actor-Critic-ac-no-rewards-till-complete-full-training
9. Actor-Critic-ac-reduce-reward-by-four
10. Actor-Critic-ac-reward-2-divide-rewards-lr-e6
11. Actor-Critic-ac-reward-2-min-3-disc-none-only-1000-steps
12. Actor-Critic-ac-reward-2-min-actions
13. Actor-Critic-ac-reward-2-min-actions-3-discount-none-only
14. Actor-Critic-ac-reward-2-normalized-dataframe
15. Actor-Critic-ac-reward-2-normalized-dataframe-lr-e7
16. Actor-Critic-ac-reward-2-rew-by-two-stop-with-none
17. Actor-Critic-ac-reward-2-rew-by-two-stop-with-none-only
18. Actor-Critic-ac-vit-encoder
19. Actor-Critic-no-sigmoid-state-value-adam-l2-reg-30steps-per-image-no-limit

APÉNDICE A. APÉNDIZE A

20. Actor-Critic-no-sigmoid-state-value-adam-l2-reg-30steps-per-image-stop-action
21. Actor-Critic-softmax-state-value-adam-l2-reg-30steps-per-image-stop-action-is
22. Actor-Critic-softmax-state-value-rmsprop-30steps-per-image-stop-action-is-non
23. Actor-Critic-v1
24. Actor-Critic-v2
25. Policy gradient normalized image rewards
26. Policy gradient normalized image rewards v2
27. Policy gradient recompensa 10
28. softmax-state-value

Bibliografía

- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., y Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.
- Cai, H., Chen, T., Zhang, W., Yu, Y., y Wang, J. (2018). Efficient architecture search by network transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., y Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Haarnoja, T., Zhou, A., Abbeel, P., y Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290.
- Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., y Wang, Y. (2019). End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1317–1332.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., y Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., y Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Morales, M. (2020). *Grokking Deep Reinforcement Learning*. Manning.
- Redmon, J., Divvala, S., Girshick, R., y Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Ridnik, T., Ben-Baruch, E., Noy, A., y Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- Simonyan, K. y Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sutton, R. S. y Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Sutton, R. S., McAllester, D., Singh, S., y Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.
- Zhao, W., Meng, Z., Wang, K., Zhang, J., y Lu, S. (2021). Hierarchical active tracking control for uavs via deep reinforcement learning. *Applied Sciences*, 11(22).
- Zhou, D., Sun, G., y Lei, W. (2021). Space non-cooperative object active tracking with deep reinforcement learning. *arXiv preprint arXiv:2112.09854*.