UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escuela Técnica Superior de Ingeniería Informática
Universidad Politécnica de Valencia

Detección de defectos en objetos en movimiento mediante Redes Neuronales Convolucionales con optimizaciones específicas para hardware NVIDIA

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Haro Armero, Abel

Tutor: Flich Cardo, José

Curso 2024-2025

Resum

????

Paraules clau: ????, ?????????, ????, ??????????????

Resumen

????

Palabras clave: ?????, ???, ??????????????

Abstract

????

Key words: ?????, ????? ?????, ?????????????

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Durante los últimos años, la inteligencia artificial ha experimentado un crecimiento en popularidad sin precedentes, transformando nuestra capacidad tecnológica con herramientas revolucionarias. Este avance ha sido impulsado por la disponibilidad de grandes volúmenes de datos y el desarrollo de algoritmos avanzados, que han permitido a las máquinas aprender y adaptarse a situaciones complejas. Algunos campos destacados de aplicación incluyen el procesamiento del lenguaje natural, la visión por computador y la robótica. En particular, la visión por computador ha visto un auge significativo, con aplicaciones en áreas como la seguridad, la medicina y la automoción. Este creciente interés se refleja en la evolución del interés público en inteligencia artificial, como muestra la Figura 1.1, basada en datos de Google Trends [2].

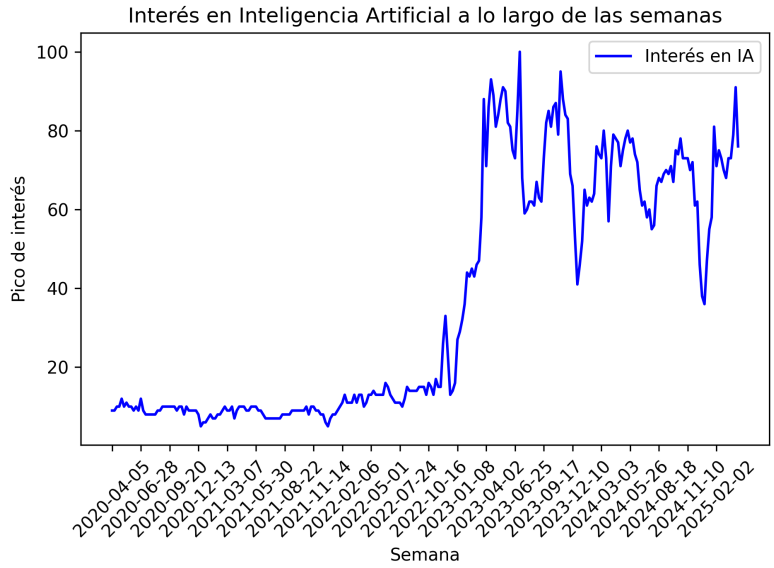


Figura 1.1: Evolución del interés público en inteligencia artificial según datos de Google Trends (2020-2025)

Este progreso ha sido posible gracias a los avances en redes neuronales convolucionales, que han revolucionado la capacidad de los sistemas para detectar y clasificar objetos en imágenes y vídeos con una gran precisión y velocidad.

Estos algoritmos de visión artificial requieren una potencia computacional significativa tanto para su entrenamiento como para su ejecución. Las CPUs (Unidades Centrales de Procesamiento) tradicionales resultan insuficientes para estas tareas, por lo que la industria ha desarrollado arquitecturas específicas como las GPUs (Unidades de Procesa-

miento Gráfico), TPUs (Unidades de Procesamiento Tensorial) y DLAs (Aceleradores de Aprendizaje Profundo). Estos componentes están optimizados para ejecutar operaciones de entrenamiento e inferencia de manera eficiente, permitiendo implementar sistemas de visión artificial capaces de procesar información visual en tiempo real. Sin embargo, estos aceleradores suelen presentar un consumo energético elevado, lo que plantea importantes retos de eficiencia y sostenibilidad.

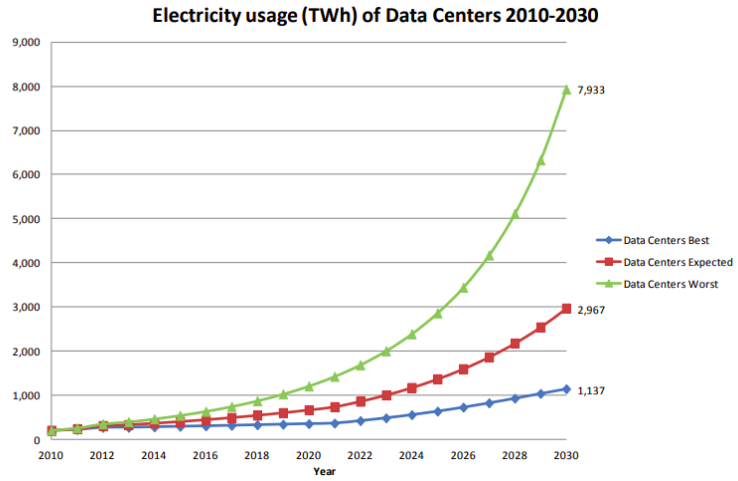


Figura 1.2: Proyección del consumo eléctrico de los centros de datos en el mundo

Como se observa en la Figura 1.2, el consumo eléctrico de los centros de datos en el mundo ha ido aumentando de forma exponencial, lo que plantea un desafío significativo para la sostenibilidad del crecimiento tecnológico [1]. En el peor escenario, esta tendencia podría llevar a un incremento insostenible en la huella de carbono del sector tecnológico, mientras que en el mejor de los casos, la adopción de tecnologías eficientes podría moderar este crecimiento. Este aumento del consumo energético no solo afecta a los centros de datos, sino también a los dispositivos embebidos y móviles, donde la eficiencia energética es crucial para prolongar la vida útil de las baterías y reducir el impacto ambiental.

Para enfrentar estos desafíos, se han desarrollado diversas técnicas de optimización y compresión que reducen el tamaño y la complejidad de los modelos neuronales manteniendo su rendimiento. Paralelamente, han surgido arquitecturas hardware específicamente diseñadas para la inferencia de modelos de aprendizaje profundo en entornos con restricciones energéticas. En este contexto, los dispositivos de la serie Jetson de NVIDIA destacan por ofrecer un equilibrio entre alto rendimiento en tareas de inteligencia artificial y un consumo energético contenido, ideal para aplicaciones embebidas de visión artificial.

La combinación de redes neuronales convolucionales y aceleradores hardware ha permitido la creación de sistemas de visión artificial que pueden detectar y clasificar objetos en movimiento, lo que es esencial en aplicaciones como la vigilancia, la conducción autónoma y la robótica.

1.1 Motivación

Los humamos somos capaces de ver y entender el mundo que nos rodea. Dada una imagen, podemos identificar objetos, reconocer patrones y tomar decisiones basadas en la información visual. Sin embargo, esta capacidad no es innata en las máquinas. La visión por computador es la ciencia que busca dotar a las máquinas de la capacidad de

interpretar y comprender imágenes y vídeos, emulando la forma en que los humanos percibimos el entorno.

Como se mencionó anteriormente, la inteligencia artificial ha revolucionado la forma en que interactuamos con la tecnología. Se ha convertido en una herramienta esencial para aplicar soluciones innovadoras en una amplia gama de campos. En particular, la visión por computador ha demostrado ser un área de gran potencial. También la existencia de dispositivos de bajo consumo, como los de la serie Jetson de NVIDIA, ha permitido llevar la inteligencia artificial a entornos de edge computing (cómputo en el borde), donde se acerca el procesamiento de datos a la fuente de información. Esto reduce la latencia y el consumo energético. Con todo esto, se abre un abanico de posibilidades para la implementación de sistemas de visión artificial en aplicaciones industriales.

Centradánose en el ámbito industrial, la detección y clasificación de objetos en movimiento es crucial para optimizar procesos, mejorar la seguridad y aumentar la eficiencia. En la mayoría de entornos productivos, la detección de defectos se realiza de forma manual, lo que puede ser ineficiente y propenso a errores. La automatización de este proceso mediante sistemas de visión artificial puede reducir costos, aumentar la precisión y mejorar la calidad del producto final.

La motivación de este trabajo radica en la necesidad de desarrollar un sistema de visión artificial capaz de detectar y clasificar objetos en movimiento en un entorno industrial, específicamente en una cinta transportadora.

1.2 Objetivos

El objetivo principal de este trabajo es desarrollar un sistema de visión artificial capaz de detectar y clasificar objetos en movimiento en una cinta transportadora utilizando redes neuronales convolucionales y aceleradores hardware de bajo consumo. Para lograr este objetivo, se plantean los siguientes objetivos específicos:

- Realizar un estudio del estado del arte en redes neuronales convolucionales, aceleradores hardware de bajo consumo y técnicas avanzadas de optimización para visión artificial.
- Desarrollar un conjunto de datos para el entrenamiento y evaluación del sistema, mediante la captura y etiquetado de imágenes de objetos en movimiento.
- Diseñar, entrenar y validar un modelo de red neuronal convolucional optimizado para la detección y clasificación en tiempo real de defectos en objetos en movimiento.
- Implementar un sistema completo de visión artificial que integre el modelo entrenado con los aceleradores hardware NVIDIA, enfocado en maximizar la eficiencia y minimizar la latencia.
- Analizar los cuellos de botella del sistema, y aplicar técnicas específicas de optimización para mejorar el rendimiento y la eficiencia energética.
- Cuantificar de manera exhaustiva el rendimiento del sistema mediante métricas precisas de exactitud (mAP, precisión, recall), latencia (FPS) y consumo energético (W, J/inferencia).
- Realizar un análisis comparativo sistemático entre diferentes configuraciones de hardware, software y parámetros de optimización para identificar la combinación que ofrezca el mejor equilibrio entre precisión, velocidad y eficiencia energética.

1.3 Estructura de la memoria

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 2
Estado del arte

En este capítulo se realizará un estudio del estado del arte en los diferentes componentes que constituyen la base teórica y técnica de este trabajo. Primero, se examinarán las redes neuronales convolucionales, desde sus fundamentos hasta los modelos más recientes en detección de objetos. A continuación, se analizarán los aceleradores hardware de bajo consumo, con especial énfasis en la arquitectura y capacidades de los dispositivos NVIDIA Jetson. Posteriormente, se estudiarán los algoritmos de seguimiento de objetos en tiempo real, fundamentales para aplicaciones con elementos en movimiento. Finalmente, se explorará la técnica de Slicing Aided Hyper Inference (SAHI), una metodología avanzada para mejorar la detección de objetos pequeños o densamente agrupados. Este marco teórico permitirá contextualizar adecuadamente la solución propuesta para la detección de defectos en objetos en movimiento.

2.1 Redes neuronales convolucionales

En esta sección se realizará un estudio de las redes neuronales hasta las redes neuronales convolucionales, desde sus fundamentos hasta los modelos más recientes en detección de objetos. Se explicarán los conceptos básicos de las redes neuronales y la evolución de las arquitecturas.

La ***Inteligencia Artificial*** es un campo de estudio que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de voz, la toma de decisiones y la comprensión del lenguaje natural. Dentro de este campo, existen diversas subdisciplinas, entre las cuales destacan el ***Machine Learning*** y el ***Deep Learning***.

El ***Machine Learning*** o aprendizaje automático es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos y realizar predicciones o tomar decisiones sin ser programadas explícitamente. Este enfoque se basa en la idea de que las máquinas pueden identificar patrones y relaciones en grandes conjuntos de datos, lo que les permite generalizar y adaptarse a nuevas situaciones.

El ***Deep Learning*** o aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales artificiales para modelar y resolver problemas complejos. Estas redes están compuestas por múltiples capas de neuronas interconectadas, que permiten aprender representaciones jerárquicas de los datos. Además, su capacidad para generalizar a partir de ejemplos les permite abordar tareas complejas con un alto grado de precisión.

Las ***Convolutional Neural Networks*** (CNN) o redes neuronales convolucionales son un tipo específico de red neuronal profunda. Estas redes están diseñadas para procesar imágenes y extraer características relevantes de manera eficiente, lo que las hace especialmente adecuadas para tareas de visión por computador.

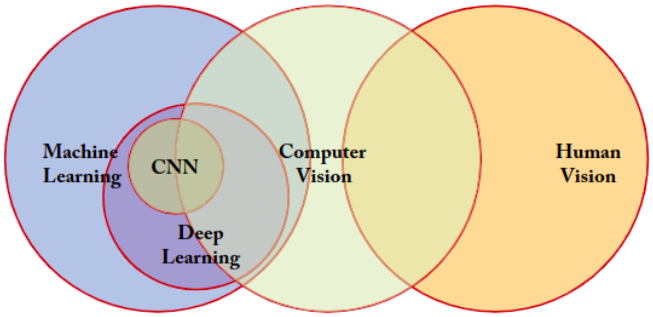


Figura 2.1: Relación entre Machine Learning, Deep Learning, CNN, Computer Vision y Human Vision.

La Figura 2.1 ilustra la relación entre estos conceptos [4]. Las CNN son una subcategoría del Deep Learning, que a su vez es una subcategoría del Machine Learning. Además, las CNN están estrechamente relacionadas con la visión por computador, que busca emular la capacidad de los humanos para interpretar imágenes y vídeos.

Las CNN se inspiran en la forma en que los humanos percibimos el mundo visual. Al igual que nuestro sistema visual, que procesa la información de manera jerárquica, las CNN utilizan capas convolucionales para extraer características de bajo nivel (como bordes y texturas) y capas más profundas para identificar patrones y objetos más complejos. Esta jerarquía de características permite a las CNN aprender representaciones ricas y abstractas de los datos visuales.

La operación de convolución es fundamental en las CNN. Esta operación consiste en aplicar un filtro (o kernel) a una imagen para extraer características locales. El filtro se desliza sobre la imagen, multiplicando sus valores por los valores de la imagen en cada posición y sumando los resultados. Este proceso genera un mapa de activación que resalta las características relevantes de la imagen.

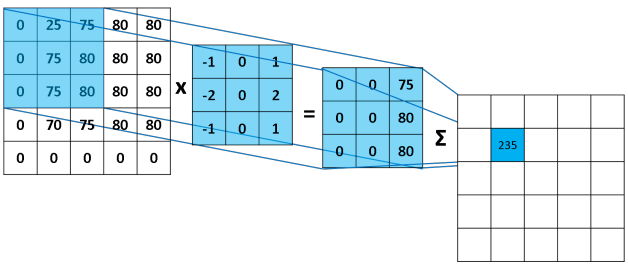
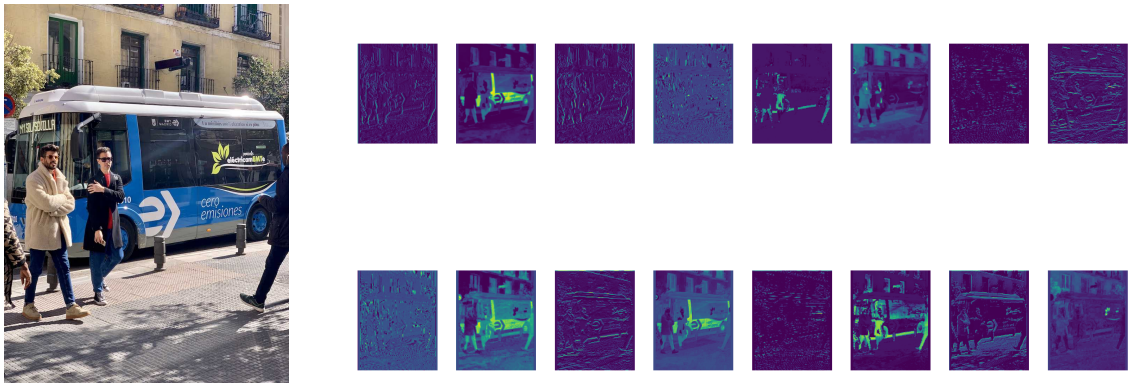


Figura 2.2: Operación de convolución en una imagen.

La Figura 2.2 muestra un ejemplo de la operación de convolución. En este caso, se aplica un filtro de 3x3 a una imagen de entrada, generando un mapa de activación que resalta las características detectadas por el filtro.



(a) Imagen de un autobús. **(b)** Resultado de la operación de convolución.

Figura 2.3: Proceso de convolución aplicado a una imagen de un autobús.

La Figura 2.3 ilustra el proceso de la primera convolución del modelo yolo11n [3]. En la parte izquierda se muestra la imagen original de un autobús, mientras que en la parte derecha se presenta el resultado de aplicar la operación de convolución. En este caso, los 16 filtros de la primera capa convolucional han detectado diferentes características de la imagen, como bordes y texturas. Este proceso se repite en múltiples capas, lo que permite a la red aprender representaciones cada vez más complejas de la imagen.

1. **Fundamentos del Deep Learning**
   - *a*) Concepto de aprendizaje profundo
   - *b*) Redes neuronales artificiales básicas
   - *c*) Función de activación, pesos y capas
2. **Redes neuronales profundas**
   - *a*) Arquitecturas multicapa
   - *b*) Desafío del desvanecimiento del gradiente
   - *c*) Técnicas de inicialización y normalización
3. **Redes neuronales convolucionales (CNN)**
   - *a*) Operación de convolución y su importancia
   - *b*) Capas convolucionales, pooling y fully-connected
   - *c*) Invarianza a la traslación y extracción jerárquica de características
4. **Detección de objetos: primeros enfoques**
   - *a*) Sliding window
   - *b*) Uso de características handcrafted (HOG, SIFT)
   - *c*) Limitaciones de los enfoques tradicionales
5. **Detectores de dos etapas (two-stage)**
   - *a*) R-CNN: region proposals + clasificación
   - *b*) Fast R-CNN: mejoras compartiendo cálculos
   - *c*) Faster R-CNN: Region Proposal Network (RPN)
   - *d*) Mask R-CNN: incorporación de segmentación

6. **Detectores de una etapa (one-stage)**
   - *a*) SSD (Single Shot MultiBox Detector)
   - *b*) RetinaNet y Focal Loss
   - *c*) Ventajas en velocidad vs. precisión
7. **YOLO (You Only Look Once)**
   - *a*) YOLOv1: división en grid y predicción directa
   - *b*) YOLOv2/YOLO9000: mejoras con anchor boxes
   - *c*) YOLOv3: múltiples escalas y características
   - *d*) YOLOv4: mejoras en backbone y técnicas de aumento
8. **Evolución reciente de YOLO**
   - *a*) YOLOv5: optimización y escalabilidad
   - *b*) YOLOv6/v7: avances arquitectónicos
   - *c*) YOLOv8: modularidad y rendimiento
   - *d*) YOLO11: nuevas características y optimizaciones
9. **Métricas de evaluación**
   - *a*) Precisión y recall
   - *b*) IoU (Intersection over Union)
   - *c*) mAP50 y mAP50-95
   - *d*) Velocidad (FPS) y compromiso velocidad-precisión

2.2 Aceleradores de procesamiento gráfico

Evolución de los aceleradores de procesamiento gráfico desde la GPU, hasta los dispositivos de bajo consumo en la serie Jetson de NVIDIA. Comentar también TensorRT y como se utiliza para optimizar los modelos en inferencias para los dispositivos de NVIDIA.

2.3 Seguimiento de objetos en tiempo real

Explicación de como funcionan los algoritmos de multi-object tracking (MOT) en tiempo real, filtro de Kalman hasta BYTETrack.

2.4 Slicing Aided Hyper Inference

Explicación de la técnica de Slicing Aided Hyper Inference, como se utiliza para mejorar la precisión de los modelos de detección de objetos y como se aplica en este trabajo.

CAPÍTULO 3

Análisis del problema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 4
Diseño e implementación de la solución

En este capítulo se explicará la solución propuesta, como se ha diseñado y como se ha implementado.

4.1 Descripción del sistema

Descripción del sistema de visión artificial propuesto, como se ha diseñado y como se ha implementado.

4.2 Diseño de las etapas del sistema

Descripción de las etapas del sistema, como se han diseñado y como se han implementado.

Etapas del sistema:

- **Captura de imágenes:** Descripción de la etapa de captura de imágenes, como se ha diseñado y como se ha implementado.
- **Inferencia:** Descripción de la etapa de inferencia, como se ha diseñado y como se ha implementado.
- **Seguimiento:** Descripción de la etapa de seguimiento, como se ha diseñado y como se ha implementado.
- **Escritura de resultados:** Descripción de la etapa de escritura de resultados, como se ha diseñado y como se ha implementado.

4.3 Segmentación de las etapas del sistema

Tipos de segmentación de las etapas del sistema:

- **No segmentada:** Secuencial
- **Segmentación basada en hilos:** Cada etapa del sistema se ejecuta en un hilo diferente.

- **Segmentación basada en procesos:** Cada etapa del sistema se ejecuta en un proceso diferente.
- **Segmentación basada en hardware:** La etapa de inferencia se ejecuta en GPU, DLA0 y DLA1.
- **Segmentación basada en procesos con memoria compartida:** Cada etapa del sistema se ejecuta en un proceso diferente, pero comparten la memoria.

CAPÍTULO 5
Análisis de la solución

En este capítulo se analizará la solución propuesta variando los parámetros posibles

5.1 Variación de los parámetros

Explicación de los parámetros que se pueden variar en la solución propuesta y su efecto en el rendimiento del sistema.

——PRUEBA———

Model	IoU	CPU_Inference	GPU_Inference	DLA_Inference	CPU_Power	GPU_Power	DLA_Power	CPU_Energy	GPU_Energy	DLA_Energy
YOLOv11-N	0,85	45,2	12,3	15,8	8,2	12,5	6,8	369,64	153,75	107,44
YOLOv11-S	0,87	52,1	14,8	18,2	8,5	13,2	7,1	442,85	195,36	129,22
YOLOv11-M	0,89	68,4	18,2	22,5	9,1	14,8	7,8	622,44	269,36	175,5
YOLOv11-L	0,91	85,6	24,6	28,9	9,8	16,2	8,4	838,88	398,52	242,76

Tabla 5.1: Comparación de modelos en términos de inferencia, consumo de energía y potencia.

——PRUEBA———

5.2 Tipo de segmentación

En esta sección se analizará el rendimiento de la solución propuesta variando el tipo de segmentación de las etapas del sistema con gráficas y tablas.

5.3 Talla del modelo

En esta sección se analizará el rendimiento de la solución propuesta variando la talla del modelo de detección de objetos con gráficas y tablas.

5.4 Precisión del modelo

En esta sección se analizará el rendimiento de la solución propuesta variando la precisión del modelo de detección de objetos con gráficas y tablas.

5.5 Modo de energía y cores de la CPU

En esta sección se analizará el rendimiento de la solución propuesta variando el modo de energía del dispositivo y el número de cores de la CPU con gráficas y tablas.

5.6 Tamaño de la imagen

En esta sección se analizará el rendimiento de la solución propuesta variando el tamaño de la imagen de entrada del modelo con la técnica de Slicing Aided Hyper Inference (SAHI) con gráficas y tablas.

CAPÍTULO 6
Prueba de concepto

Aqui se explicará la implementación de la solución propuesta en el entorno de producción con la cinta transportadora.

6.1 Construcción del entorno

6.2 Instalación del entorno

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 7
Conclusiones

????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

[1] Anders S. G. Andrae and Tomas Edler. On global electricity usage of communication technology: Trends to 2030. *Challenges*, 6(1):117–157, 2015.

[2] Google Trends. Interés en inteligencia artificial (2020–2025). https://trends.google.com/trends/explore?date=2020-04-08%202025-04-08&q=inteligencia%20artificial&hl=es, 2025. Consultado el 08 abril de 2025.

[3] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[4] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. *A Guide to Convolutional Neural Networks for Computer Vision*. Synthesis Lectures on Computer Vision. Springer Cham, 1 edition, 2018.

APÉNDICE A
Configuración del sistema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.1 Fase de inicialització

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.2 Identificación de dispositivos

????? ????????????? ????????????? ????????????? ????????????? ?????????????

APÉNDICE B
??? ????????????? ????

????? ????????????? ????????????? ????????????? ????????????? ?????????????