



## MAESTRÍA EN SISTEMAS EMBEBIDOS

MEMORIA DEL TRABAJO FINAL

### Cámara IoT para detección facial con conectividad Wi-Fi

**Autor:**

**Esp. Ing. Mauricio Barroso Benavides**

Director:

Mg. Ing. Gonzalo Sanchez (FF.AA, FIUBA)

Jurados:

Mg. Ing. Edgardo Torrelli (FIUBA)

Mg. Lic. Leopoldo Zimperz (FIUBA)

Mg. Ing. Sebastián Guarino (FIUBA)

*Este trabajo fue realizado en la ciudad de Tupiza,  
entre junio de 2021 y junio de 2023.*



## *Resumen*

Esta memoria describe el proceso de desarrollo de un dispositivo electrónico compuesto principalmente por un módulo de procesamiento con conectividad Wi-Fi y una cámara, que puede capturar imágenes para procesarlas mediante algoritmos de Inteligencia Artificial y así detectar rostros humanos. Los datos generados por el dispositivo son transmitidos hacia servidores en la nube encargados de procesar, almacenar y facilitar su visualización para los usuarios finales. La principal aplicación de este trabajo es generar información sobre la presencia de personas para, por ejemplo, activar otros dispositivos como bocinas o mecanismos de cierre/apertura de puertas.

En la realización del presente trabajo se utilizaron conocimientos adquiridos a lo largo de la carrera como desarrollo de firmware, visión artificial, diseño de hardware, sistemas distribuidos, gestión de proyectos y gestión de tecnología.



## *Agradecimientos*

A Gonzalo Sanchez, director de este trabajo, por sus valiosos consejos y criterios que se ven reflejados a lo largo de este desarrollo.

A los profesores de la Maestria en Sistemas embebidos, por contribuir en mi formacion academica con sus conocimientos y experiencias.



# Índice general

<b>Resumen</b>	<b>I</b>
<b>1. Introducción general</b>	<b>1</b>
1.1. Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo . . . . .	1
1.1.1. Inteligencia Artificial . . . . .	1
1.1.2. Aprendizaje Automático . . . . .	2
1.1.3. Aprendizaje Profundo . . . . .	2
1.2. Redes neuronales convolucionales . . . . .	3
1.2.1. Capa de convoluciones . . . . .	4
1.2.2. Capa de <i>pooling</i> . . . . .	5
1.2.3. Capa <i>fully-connected</i> . . . . .	5
1.3. Vision artificial . . . . .	6
1.4. Servicios en la nube . . . . .	7
1.4.1. Software como un servicio . . . . .	7
1.4.2. Infraestructura como un servicio . . . . .	7
1.4.3. Plataforma como un servicio . . . . .	7
1.5. Motivación . . . . .	7
1.6. Estado del arte . . . . .	8
1.7. Objetivos y alcance . . . . .	9
1.8. Requerimientos . . . . .	9
<b>2. Introducción específica</b>	<b>11</b>
2.1. Funcionamiento general del sistema . . . . .	11
2.1.1. Tarjeta de desarrollo . . . . .	12
2.1.2. Sensor de movimiento . . . . .	13
Sensor pasivo infrarrojo . . . . .	13
Amplificador operacional . . . . .	14
2.1.3. Cámara . . . . .	15
2.2. Redes Convolucionales en Cascada Multitarea . . . . .	17
2.2.1. P-Net . . . . .	18
2.2.2. R-Net . . . . .	18
2.2.3. O-Net . . . . .	18
2.2.4. Tareas de MTCNN . . . . .	19
2.3. Biblioteca TensorFlow . . . . .	19
2.4. Servicios Web de Amazon . . . . .	20
2.4.1. Servicio IoT Core . . . . .	20
2.4.2. Servicio TimeStream . . . . .	21
2.5. Plataforma Grafana . . . . .	22
<b>3. Diseño e implementación</b>	<b>23</b>
3.1. Detección facial con TensorFlow y TensorFlow Lite . . . . .	23
3.2. Desarrollo del firmware . . . . .	26

3.2.1. Detección facial con TensorFlow Lite para microcontroladores	28
3.2.2. Protocolos de comunicación . . . . .	28
3.2.3. Gestión del consumo energético . . . . .	28
3.3. Procesamiento y visualización en la nube . . . . .	28
3.3.1. Gestión de dispositivos con IoT Core . . . . .	28
3.3.2. Bases de datos de series temporales con TimeStream . . . . .	28
3.3.3. Visualización de datos con Grafana . . . . .	28
<b>4. Ensayos y resultados</b>	<b>29</b>
4.1. Pruebas funcionales del hardware . . . . .	29
<b>5. Conclusiones</b>	<b>31</b>
5.1. Conclusiones generales . . . . .	31
5.2. Próximos pasos . . . . .	31
<b>Bibliografía</b>	<b>33</b>

# Índice de figuras

1.1.	Diferencias entre AI, ML y DL . . . . .	1
1.2.	Arquitectura de una red neuronal artificial. . . . .	3
1.3.	Nodo de una red neuronal artificial. . . . .	3
1.4.	CNN para clasificar dígitos escritos a mano <sup>1</sup> . . . . .	4
1.5.	Convolución de una entrada de 3x3 con un <i>kernel</i> de 2x2 y <i>stride</i> de 1 <sup>2</sup> . . . . .	5
1.6.	Tipos de <i>pooling</i> <sup>3</sup> . . . . .	5
1.7.	Componentes de un sistema de visión artificial y un sistema de visión humana. . . . .	6
1.8.	Imagen procesada por un sistema de detección facial <sup>4</sup> . . . . .	6
1.9.	Diagrama en bloques del sistema propuesto en [13] . . . . .	8
2.1.	Diagrama en bloques del sistema. . . . .	11
2.2.	Componentes del ESP32-S3-DevKitC-1 <sup>5</sup> . . . . .	12
2.3.	Diagrama en bloques del sensor de movimiento PIR. . . . .	13
2.4.	Fotografía del sensor IRA-S230ST01 <sup>6</sup> . . . . .	14
2.5.	Fotografía del TLV8544 en un encapsulado TSSOP-14 <sup>7</sup> . . . . .	15
2.6.	Componentes del módulo ESP-LyraP-CAM <sup>8</sup> . . . . .	16
2.7.	<i>Pipeline</i> de MTCNN [18]. . . . .	17
2.8.	Arquitectura de P-Net [18]. . . . .	18
2.9.	Arquitectura de R-Net [18]. . . . .	18
2.10.	Arquitectura de O-Net [18]. . . . .	19
2.11.	Diagrama de conexión entre dispositivos IoT y AWS <sup>9</sup> . . . . .	21
3.1.	<i>Pipeline</i> detallado de MTCNN. . . . .	23
3.2.	Bloque de postprocesamiento. . . . .	24
3.3.	Bloque de preprocesamiento. . . . .	24
3.4.	Diagrama de flujo de trabajo para la conversión <sup>10</sup> . . . . .	25
3.5.	Diagrama de árbol de decisiones para el proceso de cuantización <sup>11</sup> . . . . .	25
3.6.	Diagrama de capas del firmware. . . . .	27



# Índice de tablas

2.1.	ESP32-S3-DevKitC-1 especificaciones . . . . .	13
2.2.	IRA-S230ST01 especificaciones . . . . .	14
2.3.	TLV8544 especificaciones . . . . .	15
2.4.	OV2640 especificaciones . . . . .	16
3.1.	Modelos comparativa . . . . .	26



*Este trabajo se lo dedico a mi familia, eternas gracias por su apoyo incondicional en cada etapa de mi vida. Ustedes son la luz que guia mi camino*



# Capítulo 1

## Introducción general

En este capítulo se presentan conceptos básicos sobre las tecnologías y técnicas que fueron utilizadas en el desarrollo del trabajo. Se abordan nociones sobre inteligencia artificial, aprendizaje automático, aprendizaje profundo, redes neuronales convolucionales, visión artificial y servicios en la nube. También se citan trabajos anteriores que inspiraron a este, las motivaciones para llevarlo a cabo junto a sus objetivos y alcances.

### 1.1. Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo

AI (*Artificial Intelligence*, Inteligencia Artificial), ML (*Machine Learning*, Aprendizaje Automático) y DL (*Deep Learning*, Aprendizaje Profundo), son términos muy utilizados hoy en día en el mundo del desarrollo tecnológico [1]. Aunque estos términos son muy parecidos, entre ellos existen dependencias que pueden ser visualizadas con ayuda de la figura 1.1.

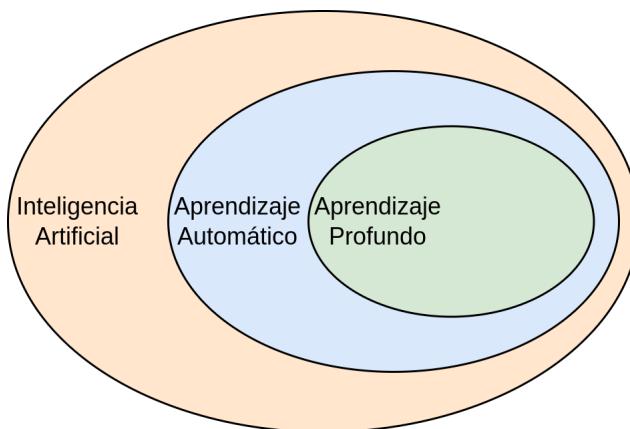


FIGURA 1.1. Diferencias entre AI, ML y DL.

#### 1.1.1. Inteligencia Artificial

AI es un área de la computación que permite a los sistemas computacionales imitar la inteligencia humana para entender su entorno y tomar acciones que maximicen sus posibilidades de lograr sus objetivos [2]. Sus aplicaciones más importante se encuentran en la áreas de comercio, educación, robótica, salud, agricultura, automotriz y finanzas[3]. Todos los sistemas de inteligencia artificial reales e hipotéticos pueden ser clasificados en alguno de los siguientes tipos [4]:

- ANI (*Artificial Narrow Intelligence, Inteligencia Artificial Estrecha*): también conocida como inteligencia artificial débil, su objetivo es llevar a cabo un solo tipo de tarea. Estos sistemas no poseen conciencia y no son manejados por sentimientos como lo haría un humano. Algunos ejemplos de ANI son los *chatbots* o los automóviles autónomos.
- AGI (*Artificial General Intelligence, Inteligencia Artificial General*): también conocida como inteligencia artificial fuerte, es un concepto en el que las máquinas exhiben inteligencia humana. Estos sistemas tendrían la capacidad de aprender, entender y actuar de tal manera que sería indistinguible a un humano. AIG actualmente no existe, pero es utilizado en industrias como el cine
- ASI (*Artificial Super Intelligence, Super Inteligencia Artificial*): ASI también forma parte de la inteligencia artificial fuerte. Se le considera muy poderosa por ser capaz de volverse consciente y autónoma. No sólo replica el comportamiento humano, sino que lo supera. Puede pensar mejor y tener más habilidades. Sin embargo, esta tecnología aún está en desarrollo.

### 1.1.2. Aprendizaje Automático

ML es un subconjunto de AI que utiliza algoritmos de aprendizaje estadísticos para construir sistemas con la habilidad de aprender automáticamente y mejorar a partir de experiencias previas sin ser explícitamente programados para esto [5]. Muchos de los servicios de recomendación utilizados por empresas como Netflix, YouTube o Spotify, utilizan ML para adaptarse a un usuario en particular y ofrecer una mejor experiencia más personalizada [6]. Estos algoritmos pueden ser clasificados de la siguiente manera [7]

- Aprendizaje supervisado: se refiere al aprendizaje modelo a partir de un conjunto de datos, mejor conocidos como *dataset*, cuyas respuestas son conocidas con antelación y están asociadas a una etiqueta o *label*. Por ejemplo, el *dataset* pueden ser muchas fotografías de gatos y el *label* asociado el nombre de este animal. De esta manera el modelo es entrenado para generar predicciones de datos nuevos.
- Aprendizaje no supervisado: es utilizado cuando los datos utilizados para el aprendizaje no tienen *labels*. Su objetivo principal es aprender acerca de los datos e inferir patrones sin ningún tipo de referencia sobre las respuestas esperadas. Es mayormente utilizado como parte del análisis exploratorio de datos [1].
- Aprendizaje reforzado: es el aprendizaje mediante la interacción continua con el entorno con el método de prueba y error, y utiliza continuamente la retroalimentación de sus acciones y experiencias previas. Este tipo de aprendizaje utiliza recompensas si se realizan acciones correctas y penalizaciones si son incorrectas.

### 1.1.3. Aprendizaje Profundo

DL es una técnica de ML que está inspirada en la forma en la que el cerebro humano filtra información [8]. Cómo DL procesa información de manera similar al cerebro humano sus aplicaciones son tareas que un humano generalmente realiza, como distinguir entre un peatón o un poste de luz en el caso de automóviles

autónomos. El componente principal de DL son las redes neuronales artificiales, que son capas de nodos interconectados, donde existe una capa de entrada, una o varias capas ocultas y una capa de salida. En la figura 1.2 se puede observar la arquitectura de una red neuronal artificial.

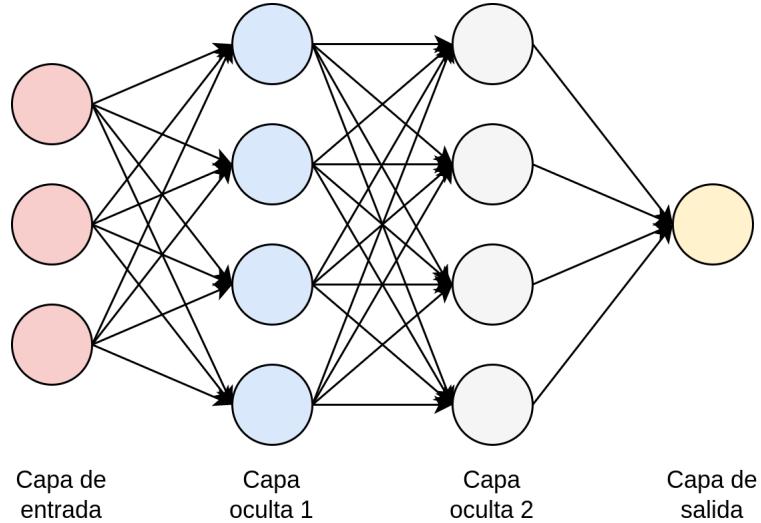


FIGURA 1.2. Arquitectura de una red neuronal artificial.

Cada uno de los nodos de las capas ocultas y de salida, tienen como entrada la salida de los nodos anteriores multiplicadas por unos términos denominados pesos o *weights* y que sumados junto a otro término llamado sesgo o *bias* pasan por una función de activación no lineal para generar su salida. En la figura 1.3 se visualiza un nodo de las capas ocultas o de salida.

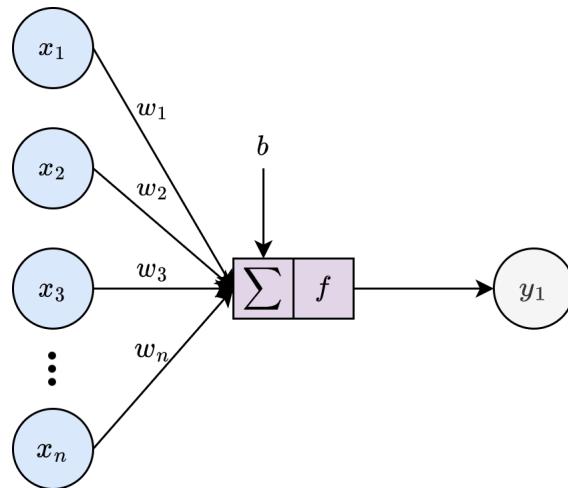


FIGURA 1.3. Nodo de una red neuronal artificial.

## 1.2. Redes neuronales convolucionales

También conocidas como CNN (*Convolutional Neural Networks*, Redes Neuronales Convolucionales) o ConvNet, son un algoritmo de DL que están orientadas a recibir como entrada una, asignarle *weights* y *biases* entrenables a varios aspectos/objetos en la imagen para poder diferenciarlas unas de otras [9]. Su uso reduce

el pre procesamiento de las imágenes de entrada con respecto a otros modelos de clasificación, ya que los filtros necesarios son incorporados en su arquitectura y tienen la habilidad de ser entrenados.

Computacionalmente una imagen puede ser muy difícil de procesar, esto depende del espacio de colores donde se encuentra [ref] y las dimensiones que posee. Por ejemplo una imagen RGB (*Red Green Blue*, Rojo Verde Azul) y de dimensiones 1920x1080 pixeles, tiene un tamaño de 6220800 bytes. El objetivo principal de las CNN es reducir la dimensionalidad de las imágenes de entrada, de tal forma que sean más fáciles de procesar y no pierdan sus características o *features* principales que son críticas para obtener una buena predicción.

La arquitectura de una CNN es independiente del tipo de aplicación, donde las capas que lo componen son elegidas en función de los objetivos que se persiguen. En la figura 1.4 se puede observar la arquitectura de una CNN para clasificar dígitos escritos a mano.

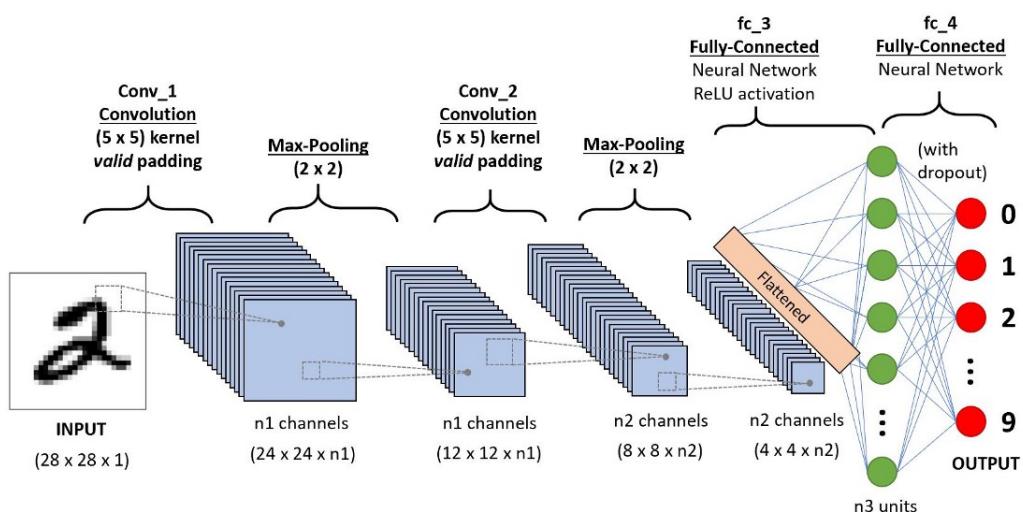


FIGURA 1.4. CNN para clasificar dígitos escritos a mano<sup>1</sup>.

En la arquitectura de la figura 1.4 se pueden observar tres capas principales para construir una CNN: capa de convoluciones, capa de *pooling* y capa *fully-connected*.

### 1.2.1. Capa de convoluciones

Esta capa es la encargada de aplicar la operación de convolución sobre las imágenes de entrada para encontrar patrones que más adelante permitirán clasificarlas. La convolución de una imagen con un *kernel* no es más que la aplicación del operador punto entre ambos. Este tipo de capas se definen por:

- El número de los *kernels* o filtros que se aplican a la imagen, que es el número de matrices por las que se van a convolucionar las imágenes de entrada.
- El tamaño de los *kernels*, donde casi siempre tienen dimensiones cuadradas e impares como 3x3 o 5x5.

<sup>1</sup>Imagen tomada de: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

- El *stride* o paso, se refiere a la forma en como el *kernel* recorre la imagen.

En la figura 1.5 se puede observar como un

Input	Kernel	Output
$\begin{array}{ c c c } \hline 0 & 1 & 2 \\ \hline 3 & 4 & 5 \\ \hline 6 & 7 & 8 \\ \hline \end{array}$	$\begin{array}{ c c } \hline 0 & 1 \\ \hline 2 & 3 \\ \hline \end{array}$	$\begin{array}{ c c } \hline 19 & 25 \\ \hline 37 & 43 \\ \hline \end{array}$

FIGURA 1.5. Convolución de una entrada de 3x3 con un *kernel* de 2x2 y *stride* de 1<sup>2</sup>.

### 1.2.2. Capa de *pooling*

Similar a la capa de convoluciones, tiene el objetivo de reducir la dimensionalidad de los *features* obtenidos mediante las convoluciones aplicadas en la capa anterior, para reducir el poder computacional requerido en un principio. Existen dos tipos de dos tipos: *max pooling* y *Average pooling*. El primero retorna el valor máximo de una porción de la imagen cubierta por el *kernel* y el segundo el valor promedio o *average*. *Max pooling* también funciona como supresor de ruido al mismo tiempo que reduce la dimensionalidad. Mientras que *average pooling* solo sirve para reducir la dimensionalidad. En la figura 1.6 se pueden observar estos tipos de *pooling*.

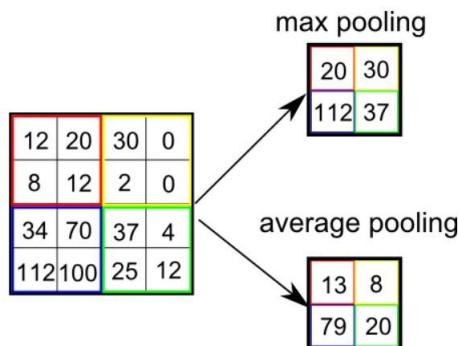


FIGURA 1.6. Tipos de *pooling*<sup>3</sup>.

### 1.2.3. Capa *fully-connected*

También conocida como capa lineal o FC (*Fully Connected*, Totalmente Conectada), es simplemente una red neuronal artificial como la mostrada en la sección anterior y se utiliza después de que las capas de convolución y *pooling* desglosan los *features* más importantes presentes en la imagen de entrada la CNN. La capa FC brinda las probabilidades finales para cada *label* esperado.

<sup>2</sup>Imagen tomada de: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

<sup>3</sup>Imagen tomada de: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

### 1.3. Vision artificial

La visión artificial o *computer vision* es un campo científico interdisciplinario que se encarga de cómo los sistemas computacionales pueden obtener un entendimiento de alto nivel de imágenes y videos digitales, para comprender y automatizar tareas como lo haría un sistema de visión humano. Las tareas que ejecuta un sistema de visión artificial son de adquisición, procesamiento, análisis y entendimiento de imágenes. En la figura 1.7 se pueden apreciar las similitudes de un sistema de visión artificial y un sistema de visión humano.

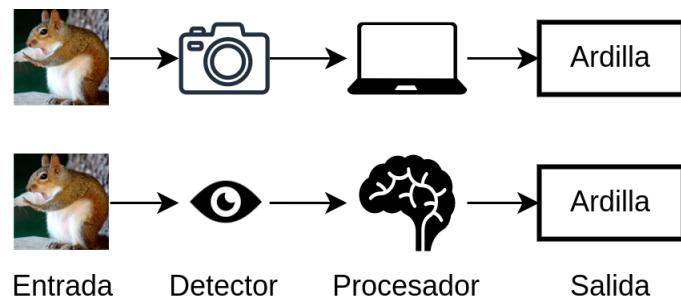


FIGURA 1.7. Componentes de un sistema de visión artificial y un sistema de visión humana.

Uno de los campos de estudio más importantes de la visión artificial es la detección facial. La detección facial puede ser considerada como un caso particular de la detección de objetos y tiene los objetivos de detectar y localizar todos los rostros humanos contenidos en una imagen digital. En la figura 1.8 se puede observar una imagen procesada por un sistema de detección facial.

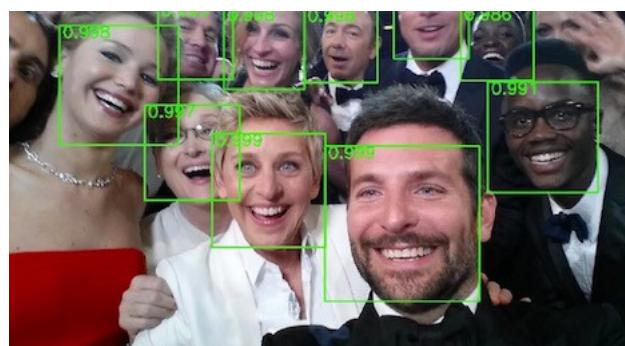


FIGURA 1.8. Imagen procesada por un sistema de detección facial<sup>4</sup>.

Hoy en día, muchos dispositivos comerciales y profesionales como smartphones, tablets y robots, utilizan la detección facial como primer paso para otro tipo de aplicaciones más complejas, entre las que destacan: reconocimiento facial, computación afectiva y grabación de vídeo inteligente [10].

<sup>4</sup>Imagen tomada de: <https://pbarasia.medium.com/use-face-recognition-on-whatsapp-group-pictures-part-1-832c6a74b5e5>

## 1.4. Servicios en la nube

El término servicios en la nube hace referencia a un amplio rango de servicios ofrecidos bajo demanda a compañías y usuarios a través de internet. Estos servicios están diseñados para proveer de una manera fácil y asequible acceso a aplicaciones y recursos, sin la necesidad de una infraestructura o hardware propios [11].

Los servicios en la nube son administrados totalmente por proveedores de computación en la nube o *cloud computing* [11]. Estos se encuentran disponibles para los usuarios desde los servidores de los proveedores, por lo que no es necesario que una empresa aloje aplicaciones en sus propios servidores. De manera general, existen tres tipos básicos de servicios en la nube: SaaS (*Software as a Service*, Software como un Servicio), IaaS (*Infrastructure as a Service*, Infraestructura como un servicio) y PaaS (*Platform as a Service*, Plataforma como un Servicio).

### 1.4.1. Software como un servicio

En este servicio el proveedor sólo proporciona el software o aplicaciones en la nube mediante internet. Los clientes tienen acceso a través de APIs (*Application Programming Interface*, Interfaz de Programación de Aplicaciones) o a través de la web, que les permite interactuar de manera sencilla, sin la necesidad de gestionar, instalar ni actualizar el software.

### 1.4.2. Infraestructura como un servicio

Este servicio implica la contratación de una infraestructura de hardware a un tercero, donde varios cliente comparten los recursos de una máquina física. El proveedor proporciona a sus clientes el acceso a los recursos computacionales necesarios para almacenar o ejecutar tareas que pueden incluir servidores, redes, *backup*, *firewalls*, entre otros.

### 1.4.3. Plataforma como un servicio

Es un servicio que se encuentra conceptualmente entre SaaS e IaaS al eliminar la parte física de la infraestructura y ofrece una plataforma donde los cliente pueden crear, desarrollar, gestionar y distribuir sus aplicaciones. El proveedor es el encargado de la gestión y mantenimiento de la plataforma y permite que los clientes se dediquen exclusivamente al desarrollo.

## 1.5. Motivación

Gracias a la amplia gama de platarformas de hardware y la disponibilidad de bibliotecas de código abierto para implementar AI, ML y DL, ademas de la difusión de información en foros y sitios web especializados, es posible desarrollar sistemas de visión artificial personalizados para distintos tipos de arquitecturas [12].

Estas bibliotecas de código para implementar visión artificial no suelen ser aptas para cualquier dispositivo, ya que son, en la mayoría de los casos, muy grandes en tamaño y requieren de una capacidad de procesamiento lo suficientemente

grande para ejecutarse en tiempo real. Pero los constantes esfuerzos de las empresas para ofrecer *frameworks* optimizados que reducen el tamaño y poder de procesamiento necesarios para ejecutar estos algoritmos, la integración de aceleradores de hardware para redes neuronales y DSP (*Digital Signal Processor, Procesador Digital de Señales*) en SoCs actuales (*System on a Chip, Sistema en un Chip*) y el estudio de nuevas y mejoradas arquitecturas para visión artificial basadas en CNN, permiten el desarrollo de dispositivos embebidos con la capacidad de ejecutar modelos de visión artificial. Algunas de sus aplicaciones más importantes son: industria 4.0, seguridad, vehículos autónomos y robótica.

La motivación principal de este trabajo fue desarrollar un sistema embebido de bajo costo económico, bajo consumo energético y de código abierto, que integre algoritmos de DL para visión artificial enfocado a la tarea de detección facial.

Una motivación adicional fue integrar otra tecnología actual como es IoT (*Internet of Things, Internet de las Cosas*), para trabajar en conjunto con los algoritmos de visión artificial. Así las aplicaciones que se pueden obtener son más versátiles a la hora de su implementación en entornos urbanos.

## 1.6. Estado del arte

Como antecedente existe el trabajo de Ilhan Aydin y Nashwan Adnan Othman, denominado “A new IoT combined face detection of people by using computer vision for security application” [13]. El *paper* donde se describe su trabajo presenta el desarrollo de un dispositivo electrónico que tiene como componentes principales una Raspberry Pi 3, un sensor PIR (*Passive Infra Red, Pasivo Infrarrojo*) y una cámara. Su objetivo principal es detectar personas con ayuda del sensor de movimiento PIR, fotografiarlas y aplicar el algoritmo de detección facial Haar Cascade [ref], para posteriormente guardar una imagen del rostro detectado y visualizarla en un teléfono móvil con ayuda de la aplicación Telegram. En la figura 1.9 se puede observar el diagrama en bloques del sistema descrito anteriormente.

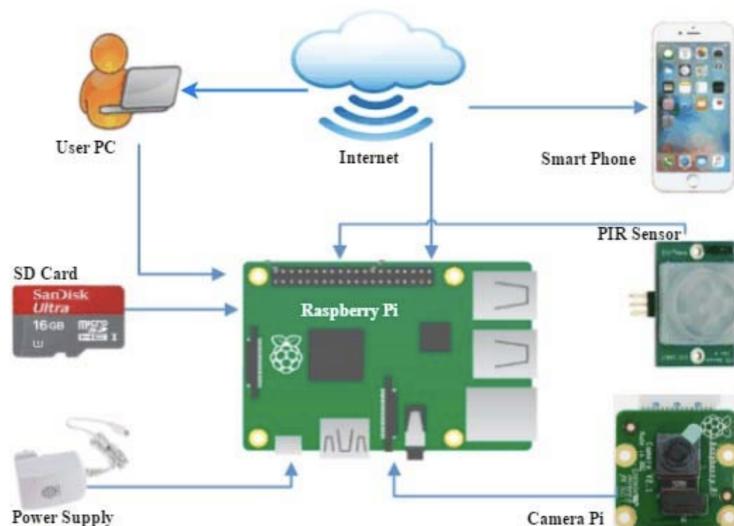


FIGURA 1.9. Diagrama en bloques del sistema propuesto en [13]

## 1.7. Objetivos y alcance

El objetivo principal de este trabajo fue desarrollar un sistema embebido con la capacidad de ejecutar modelos de AI para detectar y localizar rostros humanos de imágenes digitales capturadas por su cámara.

El alcance de este trabajo incluyó:

- Construcción de un prototipo de pruebas
- Desarrollo e implementación de los modelos de AI necesarios
- Implementación de los servicios en la nube necesarios

## 1.8. Requerimientos

Los requerimientos planteados para este trabajo fueron:

### 1. Requerimientos funcionales

- a) El sistema debe detectar y contar todos los rostros existentes de las imágenes obtenidas por su cámara con ayuda de las técnicas de procesamiento de imágenes pyramid image y sliding window, y modelos de DL que alcancen una precisión de al menos 80 %.
- b) El sistema debe conectarse a una red Wi-Fi existente a través de algún mecanismo de aprovisionamiento de credenciales de red.
- c) El sistema debe establecer comunicación con los servidores de AWS.
- d) El sistema debe ser alimentado mediante dos baterías AA de litio.
- e) El sistema debe poseer mecanismos de seguridad implementados tanto en hardware como en firmware para evitar su manipulación incorrecta.
- f) El sistema debe funcionar solamente si se detecta movimiento en el sector donde se encuentra instalada.

### 2. Requerimientos no funcionales

- a) El sistema debe tener un costo de desarrollo igual o menor a US\$200.
- b) El sistema debe tener documentación adecuada sobre su uso y desarrollo.



## Capítulo 2

# Introducción específica

Este capítulo expone una descripción detallada del sistema, del hardware utilizado y las herramientas de software necesarias en el desarrollo del trabajo. Se abarcan la descripción del sistema y sus componentes, los *frameworks* y modelos utilizados para detección facial y las herramientas utilizadas en la web.

### 2.1. Funcionamiento general del sistema

El sistema desarrollado en este trabajo consta de varios componentes de hardware que interconectados entre sí son capaces de cubrir todos los requerimientos funcionales planteados en el capítulo 1. En la figura 2.1 se muestra el diagrama en bloques del sistema.

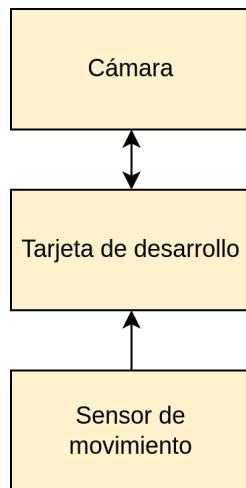


FIGURA 2.1. Diagrama en bloques del sistema.

En el sistema de la figura 2.1, cuando el sensor de movimiento detecta el movimiento de una persona genera una señal que notifica a la tarjeta de desarrollo sobre este evento. Entonces la tarjeta de desarrollo activa la cámara y obtiene una fotografía para procesarla. La imagen digital obtenida es procesada y utilizada como entrada para los modelos de DL. Cuando se obtienen las inferencias deseadas de los modelos, los resultados son procesados para transmitirlos hacia los servidores en la nube encargados de procesarlos y mostrarlos a los usuarios finales.

### 2.1.1. Tarjeta de desarrollo

El componente central del sistema es la tarjeta de desarrollo ESP32-S3-DevKitC-1-N8R8 de la empresa Espressif. Tiene como componente central el módulo ESP32-S3-WROOM-1-N8R8 y varios otros componentes que simplifican el proceso de desarrollo de aplicaciones para IoT. En la figura 2.2 se observa una fotografía de la tarjeta con el detalle de sus componentes.

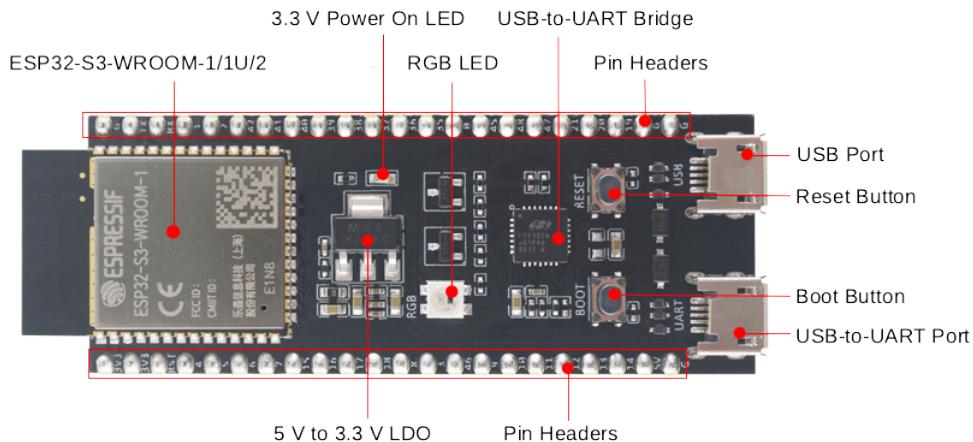


FIGURA 2.2. Componentes del ESP32-S3-DevKitC-1<sup>1</sup>.

El módulo ESP32-S3-WROOM-1-N8R8 es un potente módulo MCU (*Microcontroller Unit*, Unidad de Microcontrolador) de doble núcleo que incorpora Wi-Fi y BLE (*Bluetooth Low Energy*, Bluetooth de Baja Energía) y tiene un amplio conjunto de periféricos. Sus especificaciones técnicas más relevantes se detallan en la tabla 2.1.

En el mercado existen muchos fabricantes que ofrecen tarjetas de desarrollo de características técnicas que podrían haber sido utilizadas para el desarrollo de este trabajo. Sin ir muy lejos, Espressif, fabricante de la ESP32-S3-DevKitC-1-N8R8, tiene toda una familia de módulos y tarjetas muy similares entre sí. La elección de esta tarjeta en particular responde a los siguientes criterios:

- Costo: Espressif ofrece en todos sus SoCs, módulos y tarjetas, un costo muy contenido por la gran cantidad de características ofrecidas.
- Redes neuronales: la serie de SoCs ESP32-S3 ofrece soporte para instrucciones vectoriales, que acelera las tareas de computación de redes neuronales. Esta fue la característica más importante al momento de la elección de esta tarjeta.
- Memoria: como el trabajo implicaba el uso de una cámara y por tanto el manejo de *buffers* de memoria de gran tamaño para manipular las imágenes obtenidas, la cantidad de memoria externa que ofrece esta tarjeta la hizo óptima para la aplicación.

<sup>1</sup>Imagen tomada de: <https://docs.espressif.com/projects/esp-idf/en/latest/esp32s3/hw-reference/esp32s3/user-guide-devkitc-1.html>

TABLA 2.1. Tabla de especificaciones del ESP32-S3-DevKitC-1 [14]

Característica	Descripción
SoC embebido	ESP32-S3R8
Procesador	Xtensa LX7 doble núcleo de 32 bits
Frecuencia	Hasta 240 MHz
ROM	384 KB
SRAM	512 KB
Pines	41
Flash	8 MB
PSRAM	8 MB
Tipo de antena	PCB
Wi-Fi	802.11 b/g/n hasta 150 Mbps
Bluetooth	Bluetooth 5 y Bluetooth <i>mesh</i>
Periféricos	GPIO, I2C, SPI, interfaz LCD, interfaz de cámara, UART, I2S, USB, PWM, ADC, sensor táctil, sensor de temperatura, timer y <i>watchdogs</i>
Rango de temperatura	-40 °C a 65 °C

### 2.1.2. Sensor de movimiento

Un sensor de movimiento PIR basa su funcionamiento al detectar diferencias en la energía IR (*Infrared, Infrarrojo*) en el campo de visión del sensor. Debido a que la señal de salida generada por el sensor es muy pequeña, es necesario aplicar etapas de amplificación y filtrado para elevar el nivel de tensión de la señal de salida y al mismo tiempo filtrar el ruido que puede generar eventos falsos positivos. Esta salida analógica luego se debe convertir en una señal digital mediante la operación de comparación de ventanas y se puede utilizar, por ejemplo, como una interrupción en un MCU. En la figura 2.3 se muestra el diagrama en bloques del sensor de movimiento PIR.

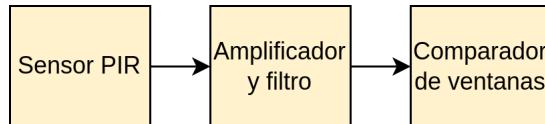


FIGURA 2.3. Diagrama en bloques del sensor de movimiento PIR.

#### Sensor pasivo infrarojo

El IRA-S230ST01 es un sensor PIR fabricado por la empresa Murata. Posee una alta sensibilidad y un rendimiento confiable gracias a la tecnología cerámica y la técnica de IC (*Integrated Circuit, Circuito Integrado*) híbrida de Murata. Tiene además una sensibilidad mejorada a la interferencia de RF (*Radio Frequency, Radiofrecuencia*). Sus aplicaciones más comunes incluyen sistemas de seguridad, aparatos de iluminación, electrodomésticos, entre otros [15]. En la figura 2.4 se puede observar una fotografía del IRA-S230ST01.



FIGURA 2.4. Fotografía del sensor IRA-S230ST01<sup>2</sup>.

En la tabla 2.2 se detallan sus características técnicas más importantes

TABLA 2.2. Tabla de especificaciones del IRA-S230ST01 [15]

Característica	Descripción
Rango de temperatura	-40 °C a 70 °C
SNR	40 dB
Campo de vision	theta1=theta2=45 grados
Electrodo	(2.0x1.0mm)x2
Responsividad	4.6 mV
Filtro óptico	5 µm paso alto
Fuente de alimentacion	2 V a 15 V

La elección del IRA-S230ST01 como sensor PIR responde a los siguientes criterios:

- Marca: Murata es una marca muy reconocida en el mundo de los semiconductores y ofrece productos de muy alta calidad.
- Documentación: el IRA-S230ST01 cuenta con documentación muy clara sobre sus características técnicas.

### Amplificador operacional

El TLV8544 es un amplificador operacional cuadruple de ultra bajo consumo energético de la empresa Texas Instruments, de costo optimizado para aplicaciones de detección en equipos inalámbricos y cableados de bajo consumo. El diseño del TLV8544 minimiza el consumo energético en dispositivos como sensores de movimiento para sistemas de seguridad, donde el tiempo de vida de la batería que los alimenta es crítico. Su uso más común es en configuraciones de amplificadores de transimpedancia con resistencias de *feedback* en el orden de los Mega ohms. Adicionalmente, tiene protección contra EMI (*Electromagnetic Interference*, Interferencia Electromagnética) que reduce la sensibilidad a las señales de RF no deseadas de fuentes como teléfonos móviles, Wi-Fi y transmisores de radio [16]. En la figura 2.5 se puede observar una fotografía del TLV8544 en un encapsulado TSSOP-14.

---

<sup>2</sup>Imagen tomada de: <https://www.murata.com/en-sg/products/productdetail?partno=IRA-S230ST01>

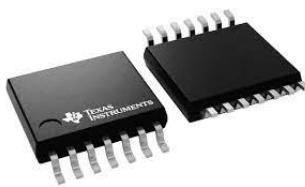


FIGURA 2.5. Fotografía del TLV8544 en un encapsulado TSSOP-14<sup>3</sup>.

Las características técnicas más importantes del TLV8544 se presentan en la tabla 2.3.

TABLA 2.3. Tabla de especificaciones del TLV8544 [16]

Característica	Descripción
Número de canales	4
Fuente de alimentación	1.7 V a 3.6 V
Corriente de salida por canal	30 mA
Corriente de operación	500 nA
CMMR ( <i>Common Mode Rejection Ratio</i> , Relación de Rechazo del Modo Común)	90 dB
Rango de temperatura	-40 °C a 125 °C
Corriente de polarización	100 fA
Ancho de banda de ganancia	8 kHz

Desde hace muchos años los amplificadores operacionales son dispositivos muy utilizados por su gran cantidad de aplicaciones, en el mercado existen una gran variedad de modelos y son fabricados por muchas empresas de semiconductores. Estos fueron los criterios de elección del TLV8544 para el presente trabajo:

- Aplicación: por sus características técnicas, el TLV8544 está diseñado para ser parte de las etapas de amplificación y filtrado en el diseño de un sensor de movimiento PIR.
- Documentación: Texas Instruments, además del correspondiente *datasheet* del TLV8544, ofrece varios documentos técnicos con ejemplos de diseño para el TLV8544.
- Costo: es un dispositivo de precio muy razonable por todas las características que ofrece.
- Consumo energético: con sus 500 nA de corriente de funcionamiento por canal, el TLV8544 es una opción ideal para aplicaciones que requieran el uso de baterías.

### 2.1.3. Cámara

Otro de los componentes principales del sistema es la cámara, que permite obtener imágenes en un formato digital que posteriormente deben ser procesadas por los algoritmos de DL. Para este trabajo se utilizó el módulo ESP-LyraP-CAM.

<sup>3</sup>Imagen tomada de: <https://www.ti.com/product/TLV8544>

Este módulo integra un CCM (*Compact Camera Module*, Módulo de Cámara Compacto) con un sensor OV2640 en conjunto con dos reguladores de tensión para su correcto funcionamiento. En la figura 2.6 se puede observar unas fotografías del módulo ESP-LyraP-CAM y sus componentes.

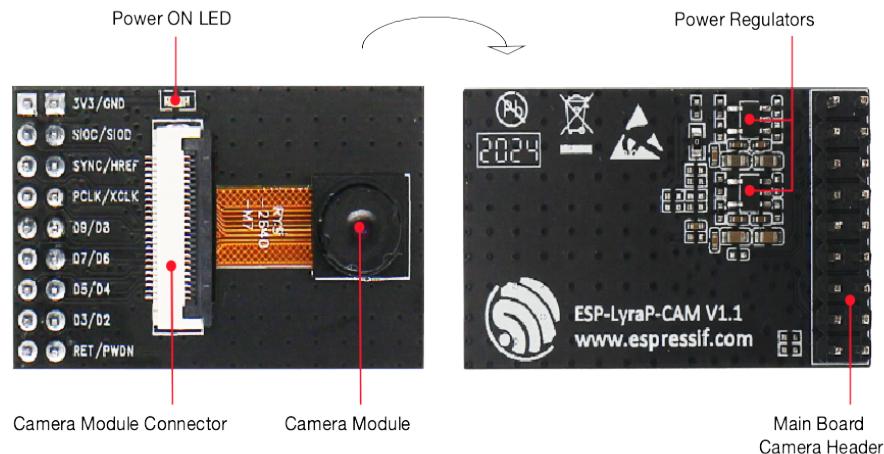


FIGURA 2.6. Componentes del módulo ESP-LyraP-CAM<sup>4</sup>.

El OV2640 de la empresa OmniVision es un sensor CMOS (*Complementary Metal-Oxide-Semiconductor*, Semiconductor de Óxido de Metal Complementario) de 2 MP, cuenta con una interfaz de comunicación compatible con DVP (*Digital Video Port*, Puerto de Video Digital), soporta codificación JPEG (*Joint Photographic Experts Group*, Grupo Unido de Expertos en Fotografía) y es de bajo consumo energético. En la tabla 2.4 se muestran las características técnicas más importantes del OV2640.

TABLA 2.4. Tabla de especificaciones del OV2640 [17]

Característica	Descripción
Tamaño de matriz	1600x1200 (UXGA) Core: 1.3 V ± 5 %
Fuente de alimentación	Analog 2.5 3.0 V I/O: 1.7 V - 3.3 V
Consumo energético	Free running: 125 mW Standby: 600 µA
Formato de imagen del sensor	1/4"
Tasa de transferencia máxima	1600×1200 a 15 fps SVGA a 30 fps CIF a 60 fps
Sensibilidad	0.6 / Lux-sec
SNR	40 dB
Rango dinámico	50 dB
Tamaño de pixel	x2.2x2.2 µm
Formato de salida	YUV/RGB/MJPEG

<sup>4</sup>Imagen tomada de: <https://docs.espressif.com/projects/esp-idf/en/latest/esp32s2/hw-reference/esp32s2/user-guide-esp-lyrap-cam-v1.1.html>

Los criterios para utilizar este módulo como cámara del sistema son los siguientes:

- Costo: los módulos con el sensor OV2640 tienen un costo muy reducido en comparación con otros disponibles en el mercado.
- Bajo consumo energético: como se mostró en la tabla 2.4 el consumo energético del módulo en modo *standby* es lo suficientemente bajo como para funcionar alimentado por baterías.
- Disponibilidad de código: al ser un módulo que ya lleva mucho tiempo en el mercado existen muchas bibliotecas de código para utilizarlo, lo que simplifica en gran medida el tiempo de desarrollo de *firmware*.

## 2.2. Redes Convolucionales en Cascada Multitarea

MTCNN (*Multi-task Cascaded Convolutional Networks*, Redes Convolucionales en Cascada Multitarea) es un *framework* basado en el *paper* “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks” de Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li y Yu Qiao [18], está desarrollado para integrar las tareas de detección facial y alineamiento facial con ayuda de CNNs en cascada mediante aprendizaje multitarea. El proceso consta de tres etapas de CNNs que puede detectar rostros humanos, sus posiciones y las posiciones de sus rasgos faciales (nariz, ojos y boca). En la figura 2.7 se muestra el *pipeline* utilizado en MTCNN.

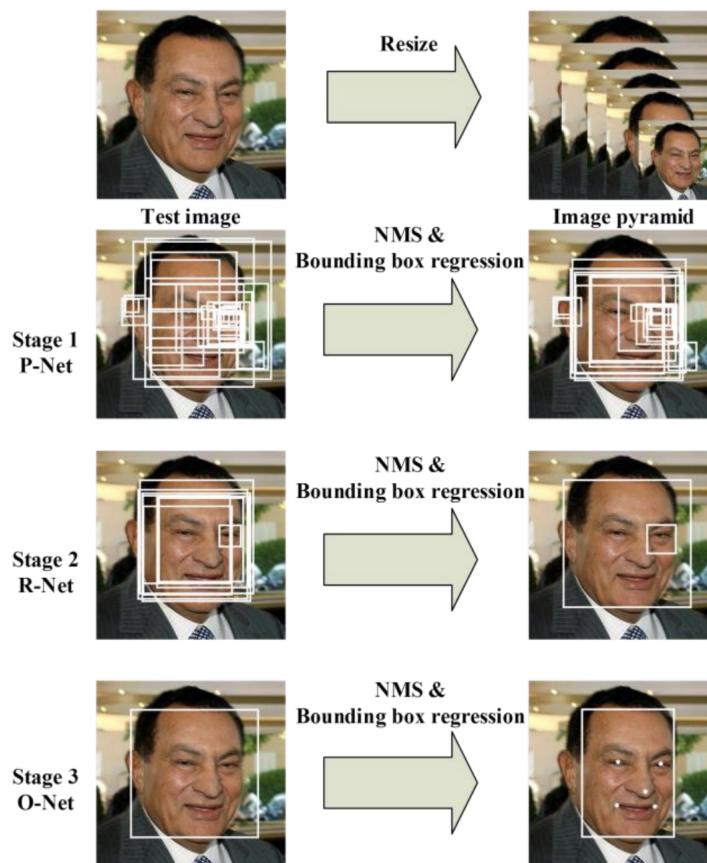


FIGURA 2.7. *Pipeline* de MTCNN [18].

### 2.2.1. Red de propuestas

También conocida como P-Net (*Proposal Network*, Red Propositiva), esta etapa está compuesta de una FCN (*Fully Convolutional Network*, Red Totalmente Convolucional). La diferencia entre una FCN y una CNN es que la FCN no utiliza una capa FC como parte de su arquitectura. Tiene la función de obtener ventanas candidatas y sus vectores de regresión de *bounding box* a partir de varias escalas de la imagen original. La regresión de *bounding box* es una técnica para predecir la localización de un cuadro delimitador en el que se encuentra el objeto que quiere ser detectado, en este caso rostros humanos. Una vez que se obtienen estos vectores, se realiza una operación conocida como NMS (*Non Max Suppression*, Supresión no Máxima) para combinar las regiones superpuestas entre sí. Finalmente las ventanas candidatas resultantes pasan a la siguiente etapa. En la figura 2.8 se muestra la arquitectura de P-Net.

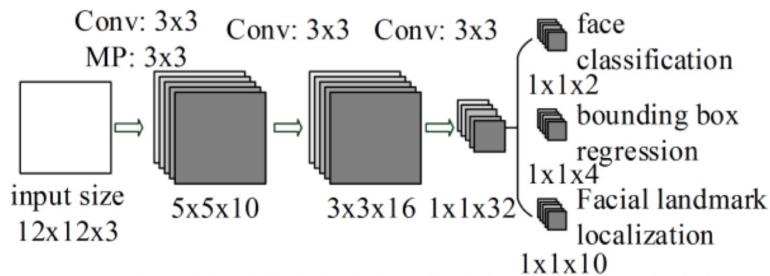


FIGURA 2.8. Arquitectura de P-Net [18].

### 2.2.2. Red de refinamiento

Esta es una CNN denominada R-Net (*Refine Network*, Red de Refinamiento). Los candidatos provenientes de P-Net son la entrada de esta red. La arquitectura de R-Net reduce aún más el número de candidatos, realiza la calibración con regresión de *bounding box* y emplea NMS para fusionar candidatos superpuestos. Para cada candidato de entrada, R-Net obtiene la probabilidad de si es un rostro o no, un vector de 4 elementos que es el *bounding box* y un vector de 10 elementos que representan la localización de rasgos faciales. En la figura 2.9 se muestra la arquitectura de R-Net.

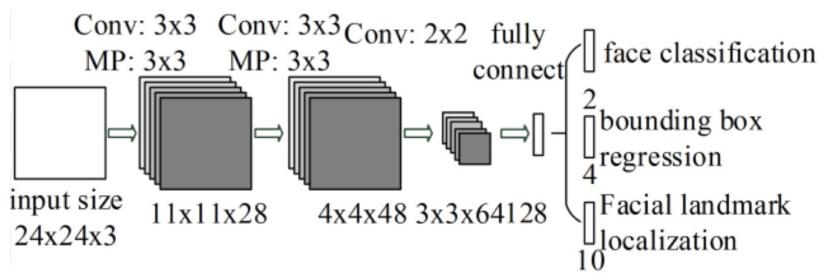


FIGURA 2.9. Arquitectura de R-Net [18].

### 2.2.3. Red de salida

Conocida como O-Net (*Output Network*, Red de Salida), es muy similar a R-Net, pero está enfocada a describir el rostro con más detalle y generar las cinco localizaciones para ojos, boca y nariz. Su arquitectura se muestra en la figura 2.10.

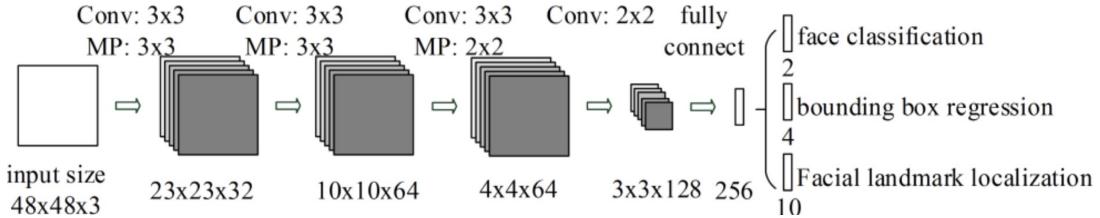


FIGURA 2.10. Arquitectura de O-Net [18].

#### 2.2.4. Tareas

Como se explicó en los puntos anteriores, MTCNN se compone de tres etapas que filtran ventanas candidatas y con la ayuda de NMS y calibración con los vectores de regresión de *bounding box*, se detectan rostros y sus rasgos. Entonces, el propósito de MTCNN es cumplir con las siguientes tareas:

- Clasificación rostro/no rostro: este es un problema de clasificación binaria que utiliza una función de pérdida de entropía cruzada.
- Regresión de *bounding box*: el objetivo de aprendizaje es un problema de regresión. Para cada candidato, se calcula el *offset* entre el candidato y el *ground truth* [ref] más cercano. La función de pérdida Euclidiana es utilizada para esta tarea.
- Localización de rasgos faciales: es formulada como un problema de regresión en el que la función de perdida es la distancia Euclidiana.

### 2.3. Biblioteca TensorFlow

TensorFlow es una plataforma de código abierto para ML. Tiene un ecosistema completo y flexible de herramientas, bibliotecas y recursos comunitarios que permite a los desarrolladores crear e implementar fácilmente aplicaciones basadas en ML. Fue originalmente desarrollado por investigadores e ingenieros que trabajaban en el equipo Google Brain dentro de la organización de investigación de inteligencia artificial de Google y la versión inicial fue lanzada en 2015 bajo la licencia Apache License 2.0 [19]. TensorFlow proporciona APIs estables y oficiales para Python y C++, aunque también existen APIs para otros lenguajes de programación que no están garantizadas de manera oficial.

Sus características principales son:

- Autodiferenciación: es el proceso de cálculo automático del vector gradiente de un modelo respecto a cada uno de sus parámetros. Ejecución ansiosa: significa que las operaciones se evalúan de manera inmediata en lugar de agregarse a un gráfico computacional que se ejecuta más tarde.
- Distribuido: TensorFlow proporciona una API para distribuir el cómputo en múltiples dispositivos tanto para ejecución ansiosa como para gráficos computacionales.
- Funciones de pérdida: TensorFlow proporciona un conjunto de funciones de pérdida, también conocidas como funciones de costo.

- Métricas: TensorFlow brinda acceso a un API de métricas de uso común que se utilizan para evaluar el rendimiento de los modelos de ML. Optimizadores: TensorFlow ofrece un conjunto de optimizadores para entrenar redes neuronales, algunos son ADAM, ADAGRAD y SGD (*Stochastic Gradient Descent*, Descenso de Gradiente Estocástico).

Para el desarrollo de aplicaciones de ML existen varias otras bibliotecas, algunas de las más populares son: PyTorch, Caffe Computer Vision Library, Deeplearning, Neuroph, OpenNN, Theano, Torch y MXNet. Los criterios de elección de TensorFlow en este trabajo sobre las anteriores bibliotecas citadas fueron:

- Experiencia: este fue el criterio más fuerte en elección de TensorFlow como *framework* para el desarrollo de modelos de ML. El autor de este trabajo ya poseía experiencia trabajando con TensorFlow.
- Documentación: TensorFlow tiene mucha documentación oficial sobre su API y una gran variedad de tutoriales de uso.
- Herramientas para cuantización: TensorFlow cuenta con herramientas de cuantización de datos para optimizar el tamaño y tiempos de ejecución de modelos de ML.

## 2.4. Servicios Web de Amazon

Más conocido como AWS (*Amazon Web Services*, Servicio Web de Amazon) por sus siglas en inglés, es una plataforma de *cloud computing* provista por Amazon que incluye una combinación de IaaS, PaaS y SaaS. Los servicios de AWS pueden ofrecer herramientas de poder cómputo, almacenamiento de datos y servicios de entrega de contenido [20].

AWS está dividido en distintos tipos de servicios que pueden ser configurados según las necesidades de cada usuario. Estos servicios pueden dividirse en las siguientes categorías: computación, almacenamiento, bases de datos, administración de datos, migración, redes, herramientas de desarrollo, monitoreo, administración de *big data*, analíticas, AI, desarrollo móvil, mensajería y notificaciones.

De toda la extensa cantidad de servicios que ofrece AWS, para este trabajo se necesitaron sólo los siguientes: IoT Core y Amazon Timestream.

### 2.4.1. Servicio IoT Core

Proporciona los servicios en la nube necesarios para conectar dispositivos IoT entre sí y a los otros servicios de AWS [21]. AWS IoT proporciona software que puede ayudar a integrar dispositivos IoT en soluciones basadas en las herramientas de AWS. En la figura 2.11 se puede observar un diagrama de interconexión de dispositivos IoT y los servicios de AWS mediante IoT Core.

IoT Core permite seleccionar las tecnologías más adecuadas y actualizadas para interconectar dispositivos IoT. Los protocolos de comunicación soportados son: MQTT (*Message Queuing and Telemetry Transport*, Cola de Mensajes y Transporte de Telemetría), MQTT sobre WSS (*Websocket Secure*, Websocket Seguro), HTTPS

---

<sup>5</sup>Imagen tomada de: <https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html>

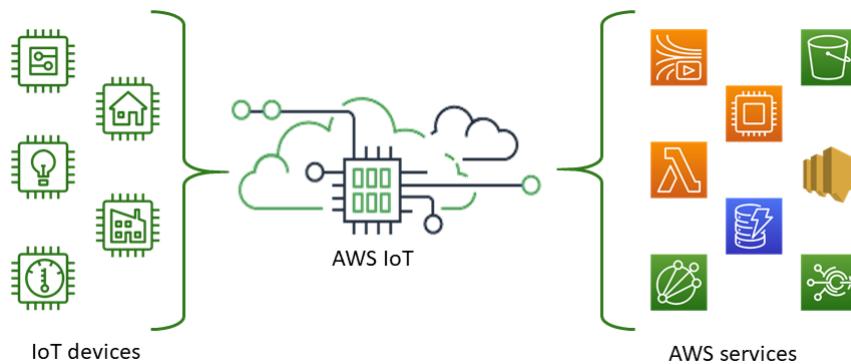


FIGURA 2.11. Diagrama de conexión entre dispositivos IoT y AWS<sup>5</sup>.

(*Hypertext Transfer Protocol Secure*, Protocolo de Transferencia de Hipertexto Seguro) y LoRaWAN.

El *broker* de IoT Core admite dispositivos y clientes que utilizan MQTT y MQTT sobre WSS para publicar y suscribirse a algún tópico. También es compatible con dispositivos y clientes que utilizan HTTPS para publicar mensajes.

#### 2.4.2. Servicio TimeStream

TimeStream es una base de datos de series temporales rápida, escalable y totalmente administrada, que facilita el almacenamiento y el análisis de billones de datos de series temporales al día. Timestream ahorra tiempos y costos con su capacidad de administrar los ciclos de vida de los datos de series temporales, donde mantiene los datos recientes en la memoria y mueve los datos históricos a un nivel de almacenamiento optimizado según las políticas definidas previamente por el usuario. El motor de consultas de Timestream permite acceder y analizar datos recientes e históricos al mismo tiempo. No necesita servidor y su tamaño se acomoda automáticamente para ajustar la capacidad y el rendimiento requeridos [22].

Los beneficios más notables que ofrece Amazon Timestream son:

- Sin servidor con escalado automático: a medida que cambian las necesidades de la aplicación, Timestream escala automáticamente para ajustar la capacidad.
- Administración de los ciclos de vida de los datos: ofrece niveles de almacenamiento, con un almacenamiento de memoria para datos recientes y un almacenamiento magnético para datos históricos. Timestream automatiza el proceso de transferencia entre ambos almacenamientos.
- Acceso simplificado a los datos: el motor de consultas de Timestream permite acceder a los datos de forma transparente, sin la necesidad de especificar el nivel de almacenamiento.
- Diseñado para series temporales: puede analizar datos de series de tiempo con SQL, con funciones integradas de series de tiempo para suavizar, aproximar e interpolar.

- Siempre cifrado: garantiza que los datos de series de tiempo siempre están cifrados. Timestream permite especificar una clave administrada para encriptar datos en el almacenamiento magnético.

## 2.5. Plataforma Grafana

Es una aplicación web multiplataforma de análisis y visualización interactiva. Proporciona tablas, gráficos y alertas a través de la web cuando se conecta a alguna fuente de datos compatible. Los usuarios pueden crear *dashboards* de monitoreo de datos complejos con ayuda de generadores de consultas interactivos [23].

Como herramienta de visualización, Grafana es muy popular gracias a las siguientes características:

- Se conecta a muchas fuentes de datos populares como Graphite, Prometheus, Influxdb, ElasticSearch, MySQL, PostgreSQL, entre otros.
- Es de código abierto y distribuida bajo la licencia AGPL-3.0, que permite desarrollar complementos desde cero para integrar con otras fuentes de datos.
- Ayuda a estudiar, analizar y monitorear datos durante un periodo de tiempo configurable por el usuario.
- Puede ser implementado localmente por organizaciones que quieran mantener sus datos confidenciales sin acceso a internet.
- Se pueden configurar alertas que se envían por otros medios de comunicación bajo ciertas condiciones pre establecidas.

## Capítulo 3

# Diseño e implementación

### 3.1. Detección facial con TensorFlow y TensorFlow Lite

Como se explicó en el capítulo 1, el objetivo principal de este trabajo es detectar rostros humanos con ayuda de algoritmos de AI. Para esto se deben obtener imágenes digitales con ayuda de una cámara, procesarlas y utilizarlas como entrada de una red de modelos de DL capaces de realizar la tarea de detección facial. Esta red de modelos de DL fue descrita en el capítulo 2 y se denomina MTCNN.

Para implementar MTCNN adecuadamente no basta con alimentar P-Net con las imágenes obtenidas por la cámara, R-Net con las ventanas candidatas de P-Net y O-Net con las ventanas candidatas de R-Net. Los datos de entrada de cada uno de los modelos de MTCNN deben ser procesados para conseguir el mejor resultado posible, como se muestra en el diagrama de la figura 3.1.

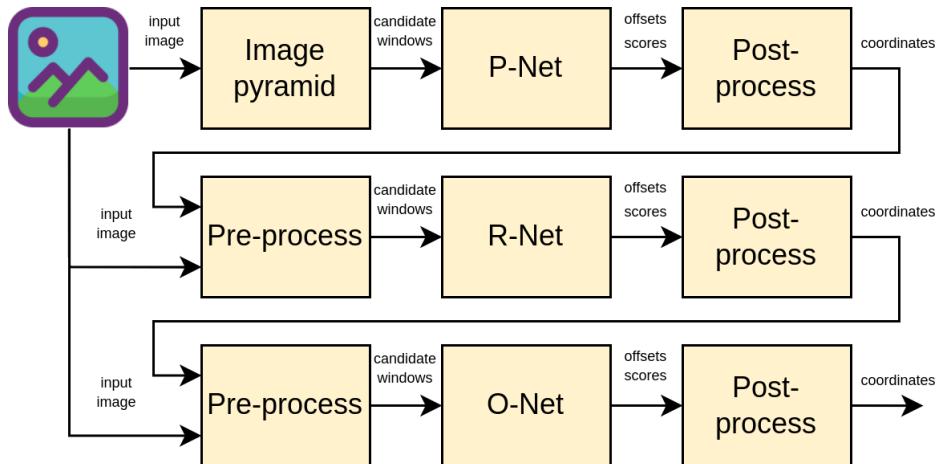


FIGURA 3.1. *Pipeline* detallado de MTCNN.

El diagrama de la figura 3.1 muestra el *pipeline* detallado de la red MTCNN, donde se pueden observar varios bloques de procesamiento, estos son:

- *Image pyramid*: genera a partir de la imagen de entrada otras imágenes de escalas inferiores, lo que permite detectar objetos de distintos tamaños. Cada nivel de escala se obtiene mediante la reducción de la escala anterior, por lo que las imágenes en niveles superiores tienen una escala más baja que las imágenes en niveles inferiores. Después de generadas las imágenes escaladas requeridas de la imagen de entrada estas sirven para alimentar P-Net y así detectar rostros de distintos tamaños.

- *Post-process*: en este bloque se procesan los datos de salida generados por P-Net, R-Net y O-Net. El primer subbloque realiza la operación de NMS para reducir la cantidad de ventanas candidatas que tienen solapamiento entre ellas. El segundo subbloque aplica un proceso de calibración que utiliza los *offsets* generados por los modelos para determinar de manera mas precisa las coordenadas de las ventanas candidatas. Finalmente el último subbloque corrige las coordenadas de las ventanas candidatas para que posean dimensiones cuadradas y estén dentro de los límites de la imagen original.

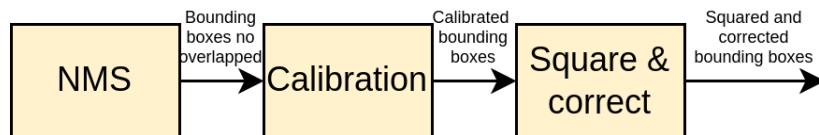


FIGURA 3.2. Bloque de postprocesamiento.

- *Pre-process*: tiene la función de procesar los datos de entrada para las redes R-Net y O-Net. El primer subbloque genera recortes de la imagen original en función de las coordenadas obtenidas del bloque *post-process*. En el segundo subbloque las imágenes recortadas de entrada son redimensionadas con dimensiones de 24x24 px y 48x48 px, para alimentar R-Net y O-Net respectivamente.<sup>1</sup>

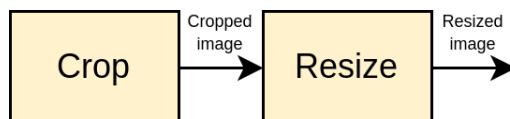


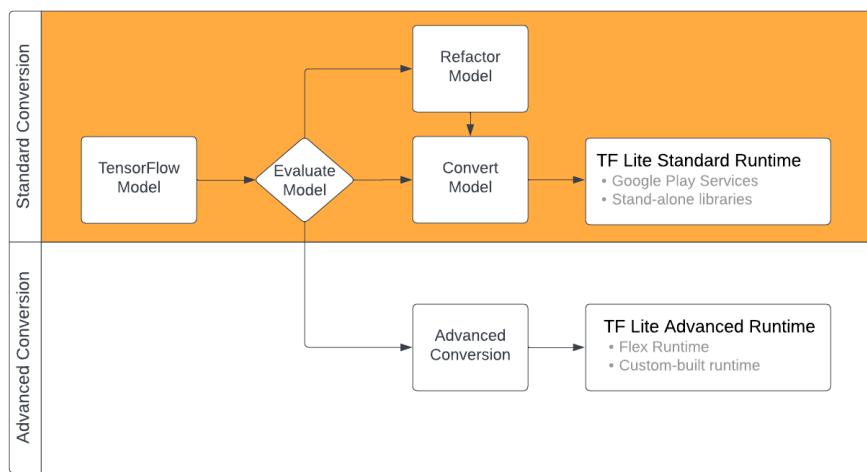
FIGURA 3.3. Bloque de preprocesamiento.

P-Net, R-Net y O-Net fueron creados con ayuda de la biblioteca para redes neuronales Keras, que es parte del *core* de TensorFlow, de acuerdo con lo expuesto en [18]. Para P-Net se crearon tantos modelos como escalas utilizadas, en este caso 3, suficientes para detectar rostros a corta distancia. Con las arquitecturas definidas de los modelos el siguiente paso natural en el desarrollo debería haber sido su entrenamiento con uno o varios *datasets*, pero al ser MTCNN tan popular en el ámbito de detección facial se pudieron encontrar archivos de tipo HDF (*Hierarchical Data Format*, Formato de Datos Jerárquicos) que contenían los *weights* resultantes de un proceso de entrenamiento anterior. En el siguiente fragmento de código se puede observar el código utilizado para crear O-Net.

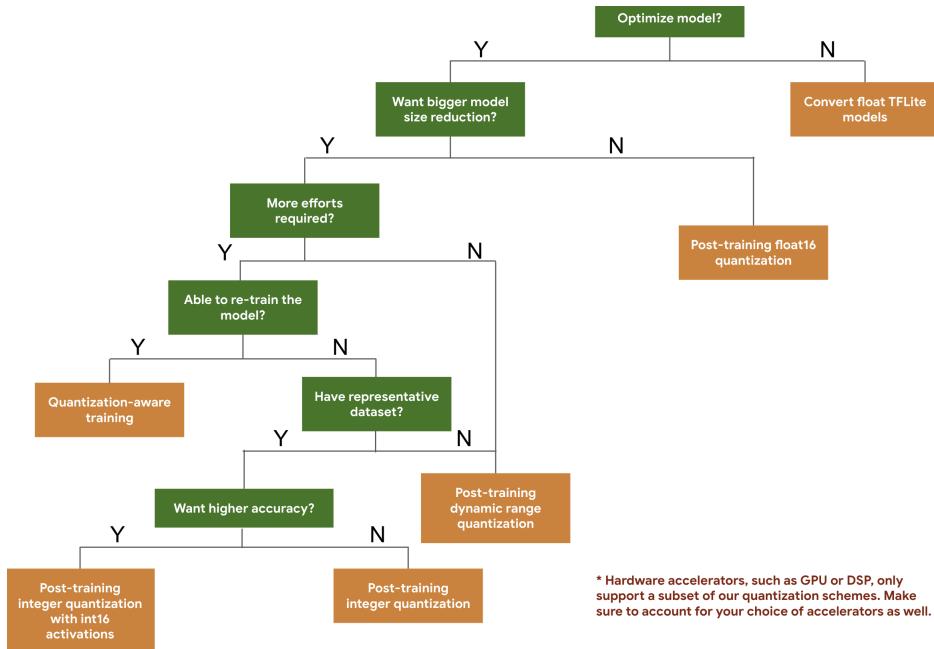
Para que los modelos obtenidos pudieran ser ejecutados en el hardware objetivo de este trabajo tuvieron que ser convertidos a un formato más liviano y eficiente llamado TensorFlow Lite. El conversor de TensorFlow Lite toma un modelo de TensorFlow y genera un modelo de TensorFlow Lite cuya extensión de archivo es .tflite. La conversión puede seguir 2 caminos según como sean evaluados los modelos de TensorFlow, en la figura 3.4 se observa el flujo de trabajo del conversor.

Gracias a que todos los operadores utilizados en los modelos de TensorFlow eran compatibles con los operadores de TensorFlow Lite se realizó una conversión estandar, lo que posteriormente facilitó su implementación en el hardware destino.

<sup>1</sup>Imagen tomada de: <https://www.tensorflow.org/lite/models/convert/>

FIGURA 3.4. Diagrama de flujo de trabajo para la conversión<sup>1</sup>.

Durante el proceso de conversión se aplicaron optimizaciones que responden a una necesidad de reducir aún más el tamaño y la latencia de los modelos. Se realizó una optimización por cuantización, que se refiere a la reducción de la precisión de los números usados para representar los parámetros de los modelos, los cuales por defecto son flotantes de 32 bits. Las opciones de cuantización para los modelos se tomaron del diagrama de la figura 3.5.

FIGURA 3.5. Diagrama de árbol de decisiones para el proceso de cuantización<sup>2</sup>.

La cuantización utilizada para los modelos fue *full integer quantization*, que reduce los picos de memoria utilizados y asegura la compatibilidad con dispositivos de hardware que no pueden utilizar punto flotante. Para este tipo de cuantización se

<sup>2</sup>Imagen tomada de: [https://www.tensorflow.org/lite/performance/model\\_optimization](https://www.tensorflow.org/lite/performance/model_optimization)

necesito crear un *dataset* representativo, compuesto por un pequeño subconjunto (entre 100 a 500 muestras) del *dataset* de entrenamiento. En ... se expone el código utilizado para la conversión de los modelos al formato TensorFlow Lite con cuantización a 8 bits.

La tabla 3.1 muestra las diferencias entre los tamaños y latencias obtenidas para el modelo O-Net de TensorFlow, TensorFlow Lite sin quantización y TensorFlow Lite con cuantización a 8 bits.

TABLA 3.1. Tabla comparativa de O-Net

TensorFlow	TensorFlow Lite	TensorFlow Lite int8
Tamaño (bytes)	2 V a 15 V	x
Latencia (ms)	2 V a 15 V	x

Todo el código para la obtención de los modelos hasta aquí expuesto, las funciones del *pipeline*, las pruebas realizadas a los modelos y el despliegue de estos en el SoC ESP32-S3, se encuentra disponible en el repositorio de acceso público [24].

## 3.2. Desarrollo del firmware

El primer paso para el desarrollo del firmware del dispositivo fue la elección de un conjunto de herramientas de software o SDK (*Software Development Kit*, Kit de Desarrollo de software) por sus siglas en inglés. Estas herramientas permitieron implementar código para utilizar de manera eficiente todos los periféricos disponibles en el ESP32-S3. Para este proyecto el SDK utilizado fue ESP-IDF [25], las razones de su elección fueron:

- Experiencia:
- Compatibilidad:
- Herramientas:
- Soporte:
- Documentación:

Con el conjunto de herramientas definido, otro aspecto de importancia fue la elección de un entorno de desarrollo para optimizar la escritura y depuración de código. El IDE (*Integrated Development Environment*, Entorno de Desarrollo Integrado) escogido fue Eclipse IDE C/C++, los aspectos más importantes para su elección fueron:

- Experiencia:
- Herramientas:
- Complementos:

Otra herramienta importante para el proceso de desarrollo del firmware fue la utilización de software para control de versiones, que permite realizar un seguimiento de los cambios realizados en el código a lo largo del tiempo. Git fue elegido como software de control de versiones, mientras que GitHub como plataforma para alojar el repositorio de Git. Las razones para la elección de ambos son:

- Experiencia:
- Reutilización de código:
- Soporte:
- Documentación:

Con todas las herramientas de software correctamente seleccionadas, el siguiente paso fue el diseño de la arquitectura del firmware. El firmware desarrollado siguió una arquitectura en capas, donde las capas de niveles más bajos tienen una mayor interacción con el hardware, mientras que las de niveles más altos con la aplicación del usuario. En la figura 3.6 se presenta el diagrama en capas del firmware.

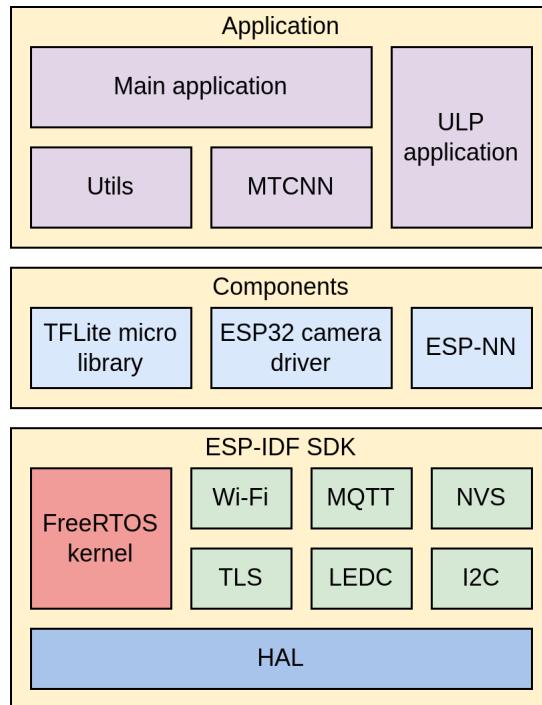


FIGURA 3.6. Diagrama de capas del firmware.

Las capas expuestas en el diagrama de la figura 3.6 son:

- ESP-IDF SDK:
- Components:
  - TFLite micro library:
  - ESP32 camera driver:
  - ESP-NN:
- Application:
  - Main application:
  - ULP application:
  - MTCNN:
  - Utils:

El firmware

**3.2.1. Detección facial con TensorFlow Lite para microcontroladores**

**3.2.2. Protocolos de comunicación**

**3.2.3. Gestión del consumo energético**

**3.3. Procesamiento y visualización en la nube**

**3.3.1. Gestión de dispositivos con IoT Core**

**3.3.2. Bases de datos de series temporales con TimeStream**

**3.3.3. Visualización de datos con Grafana**

## Capítulo 4

# Ensayos y resultados

### 4.1. Pruebas funcionales del hardware

La idea de esta sección es explicar cómo se hicieron los ensayos, qué resultados se obtuvieron y analizarlos.



## Capítulo 5

# Conclusiones

### 5.1. Conclusiones generales

La idea de esta sección es resaltar cuáles son los principales aportes del trabajo realizado y cómo se podría continuar. Debe ser especialmente breve y concisa. Es buena idea usar un listado para enumerar los logros obtenidos.

Algunas preguntas que pueden servir para completar este capítulo:

- ¿Cuál es el grado de cumplimiento de los requerimientos?
- ¿Cuán fielmente se pudo seguir la planificación original (cronograma incluido)?
- ¿Se manifestó algunos de los riesgos identificados en la planificación? ¿Fue efectivo el plan de mitigación? ¿Se debió aplicar alguna otra acción no contemplada previamente?
- Si se debieron hacer modificaciones a lo planificado ¿Cuáles fueron las causas y los efectos?
- ¿Qué técnicas resultaron útiles para el desarrollo del proyecto y cuáles no tanto?

### 5.2. Próximos pasos

Acá se indica cómo se podría continuar el trabajo más adelante.



# Bibliografía

- [1] Towards Data Science. *Difference between AI, ML and DL* | Towards Data Science. <https://towardsdatascience.com/understanding-the-difference-between-ai-ml-and-dl-cceb63252a6c>. Abr. de 2020. (Visitado 29-10-2022).
- [2] Oxford English Dictionary. *artificial intelligence*, n. : Oxford English Dictionary. <https://www.oed.com/viewdictionaryentry/Entry/271625>. Dic. de 2021. (Visitado 28-10-2022).
- [3] Simplilearn. *Top 14 Artificial Intelligence (AI) Applications in 2023* | Simplilearn. <https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/artificial-intelligence-applications>. Nov. de 2022. (Visitado 08-11-2022).
- [4] Medium. *Three Types of AI. ANI, AGI, and ASI* | by Jordan Franics | Predict | Medium. <https://medium.com/predict/three-types-of-ai-92ea7c8e57a8>. Abr. de 2022. (Visitado 28-10-2022).
- [5] AWS. *¿Qué es el machine learning? - Guía empresarial de machine learning para principiantes* - AWS. <https://aws.amazon.com/es/what-is/machine-learning/>. Ago. de 2022. (Visitado 29-10-2022).
- [6] Forbes. *The Amazing Ways YouTube Uses Artificial Intelligence And Machine Learning*. <https://www.forbes.com/sites/bernardmarr/2019/08/23/the-amazing-ways-youtube-uses-artificial-intelligence-and-machine-learning/?sh=53584be45852>. Ago. de 2019. (Visitado 29-10-2022).
- [7] Towards Data Science. *What is Machine Learning: Supervised, Unsupervised, Semi-Supervised and Reinforcement learning methods* | by Serafeim Loukas | Towards Data Science. <https://towardsdatascience.com/what-is-machine-learning-a-short-note-on-supervised-unsupervised-semi-supervised-and-aed1573ae9bb>. Jul. de 2020. (Visitado 30-10-2022).
- [8] IBM. *What is Deep Learning* | IBM. <https://www.ibm.com/cloud/learn/deep-learning>. Mayo de 2020. (Visitado 30-10-2022).
- [9] Towards Data Science. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way* | by Sumit Saha | Towards Data Science. <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>. Dic. de 2018. (Visitado 30-10-2022).
- [10] TechTarget. *What is Face Detection and How Does It Work?* <https://www.techtarget.com/searchenterpriseai/definition/face-detection>. Feb. de 2020. (Visitado 30-10-2022).
- [11] Citrix. *What is a Cloud Service? – Cloud Services Solutions* - Citrix. <https://www.citrix.com/solutions/digital-workspace/what-is-a-cloud-service.html>. Oct. de 2019. (Visitado 31-10-2022).
- [12] NWES. *Computer Vision in Embedded Systems and AI Platforms*. <https://www.nwengineeringllc.com/article/computer-vision-in->

- [embedded-systems-and-ai-platforms.php](#). Ene. de 2020. (Visitado 31-10-2022).
- [13] Ilhan Aydin y Nashwan Adnan Othman. «A new IoT combined face detection of people by using computer vision for security application». En: *Paper* (2017).
- [14] Espressif Systems. *ESP32-S3-DevKitC-1 v1.1 - Espressif Systems*. <https://docs.espressif.com/projects/esp-idf/en/latest/esp32s3/hw-reference/esp32s3/user-guide-devkitc-1.html>. Dic. de 2021. (Visitado 02-11-2022).
- [15] Murata. *IRA-S230ST01 | Pyroelectric Infrared Sensors - Murata*. <https://www.murata.com/en-sg/products/productdetail?partno=IRA-S230ST01>. Nov. de 2017. (Visitado 02-11-2022).
- [16] Texas Instruments. *TLV8544 data sheet, product information and support | TI.com*. <https://www.ti.com/product/TLV8544>. Mar. de 2017. (Visitado 02-11-2022).
- [17] Espressif Systems. *ESP-LyraP-CAM v1.1 - ESP32-S2 - Espressif Systems*. <https://docs.espressif.com/projects/esp-idf/en/latest/esp32s2/hw-reference/esp32s2/user-guide-esp-lyrap-cam-v1.1.html>. Mayo de 2020. (Visitado 03-11-2022).
- [18] Zhifeng Li y Yu Qiao Kaipeng Zhang Zhanpeng Zhang. «Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks». En: *Paper* (2016).
- [19] Github. *tensorflow/tensorflow: An Open Source Machine Learning Framework for Everyone - Github*. <https://github.com/tensorflow/tensorflow>. Nov. de 2015. (Visitado 03-11-2022).
- [20] Amazon AWS. *What is AWS - Amazon AWS*. <https://aws.amazon.com/what-is-aws/>. Ago. de 2009. (Visitado 05-11-2022).
- [21] Amazon AWS. *What is AWS IoT?* <https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html>. Oct. de 2015. (Visitado 05-11-2022).
- [22] Amazon AWS. *What is Amazon Timestream? - Amazon Timestream*. <https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html>. Sep. de 2020. (Visitado 05-11-2022).
- [23] Github. *grafana/grafana - GitHub*. <https://github.com/grafana/grafana>. Oct. de 2022. (Visitado 07-11-2022).
- [24] Mauricio Barroso Benavides. *mauriciobarroso/mtcnnesp32s3: Face detection with MTCNN, TensorFlow Lite Micro and ESP32-S3*. [https://github.com/mauriciobarroso/mtcnn\\_esp32s3](https://github.com/mauriciobarroso/mtcnn_esp32s3). Mayo de 2023. (Visitado 05-05-2023).
- [25] Espressif. *espressif/esp-idf: Espressif IoT Development Framework. Official development framework for Espressif SoCs*. <https://github.com/espressif/esp-idf>. Mayo de 2023. (Visitado 05-05-2023).